RECONHECIMENTO DE PADRÕES
UMA ABORDAGEM ESTATÍSTICA

ANDRÉ FABIO KOHN, Ph.D.

DEPARTAMENTO DE ENGENHARIA ELETRÔNICA
ESCOLA POLITÉCNICA
UNIVERSIDADE DE SÃO PAULO

1998

PREFÁCIO

O presente texto tem a finalidade de servir como roteiro para uma disciplina de pós-graduação sobre Reconhecimento Estatístico de Padrões, tendo sido utilizado tanto na Escola Politécnica da Universidade de São Paulo quanto na Universidade de Aveiro, em Portugal. Apresentamos fundamentos sobre o assunto, bem como algumas das técnicas clássicas de projeto de classificadores com e sem supervisão. Há um pré-requisito de álgebra linear e de teoria de probabilidade.

A concepção da disciplina foi baseada no acoplamento da teoria com a prática, isto significando que o aluno complementa seus estudos teóricos com experiências realizadas em computador. Cada aluno recebe um conjunto de rotinas (não incluídas neste texto, mas disponíveis para os interessados), desenvolvidas em Matlab, que são utilizadas tanto para gerar amostras com propriedades desejadas quanto para, a seguir, projetar classificadores ou estudar seu desempenho. Desta forma, o aluno pode se concentrar quase que unicamente na análise e na comparação dos resultados com a teoria vista em aula, sem se preocupar em desenvolver e validar programas para estas finalidades. Essas rotinas foram desenvolvidas por dois antigos alunos da EPUSP, Emílio del Moral Hernandez e Ricardo Tokio Higuti.

Quando boa parte deste texto já foi coberta em aula, costumamos fazer leituras e discussões em classe de alguns artigos da literatura especializada.

Finalizando, gostaria de expressar meus agradecimentos a Elisabete A. A. Fernandes, da EPUSP, que digitou meus manuscritos, e a Sandro A. Miqueleti e Rogério R. L. Cisi, do LEB/EPUSP, que auxiliaram na edição das figuras e do texto.

ÍNDICE

GLOSSÁRIO

$\underline{a}^T$: transposta do vetor $\underline{a}$

$*$: o asterisco irá denotar uma solução otimizante

B : matriz de espalhamento entre-classes ou entre-agrupamentos

c : número de classes

$C_{ij} = C(\omega_i | \omega_j)$:custo, perda ou penalidade de escolher a classe ω_i quando a classe verdadeira é ω_j

$C_\omega(\underline{x})$: custo médio associado à decisão genérica $\omega(\underline{x})$ para o dado $\underline{x}$

$C_i(\underline{x})$: custo médio (risco condicional) associado à decisão $\omega(\underline{x}) = \omega_i$ para o dado $\underline{x}$

$d(\underline{x},\underline{z})$: distância entre os vetores $\underline{x}$ e $\underline{z}$

d_{ij} : distância entre a i-ésima e a j-ésima classes ou agrupamentos

D : matriz de distâncias

$d(\underline{x})$: função de decisão mapeando $\underline{x} \in \mathbb{R}^d$ em $\mathbb{R}$

$\det(A) = |A|$ = determinante da matriz A (d x d)

$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$

$e_i(\underline{x})$: probabilidade de erro ao classificar $\underline{x}$ em ω_i

E : taxa de erro ou probabilidade de erro global de um classificador

η : dimensão original de cada vetor padrão.

G_i : i-ésimo agrupamento

$G_{N,K}$: partição de U_N em K agrupamentos

J(.) : função critério

L : limiar de decisão

$\underline{\mu}_x$: vetor médio ou vetor esperado de um vetor aleatório $\underline{x}$

N : número total de amostras ou padrões disponíveis

n_i : número de amostras ou padrões da i-ésima classe ou agrupamento ($N = \sum_{i=1}^{c} n_i$)

P_i : probabilidade a priori da classe ω_i

$P(\omega_i|\underline{x})$: probabilidade a posteriori da classe ω_i dado que ocorreu um particular $\underline{x}$

$p(\underline{x}|\omega_i)$: função densidade de probabilidade em $\underline{x}$ quando se sabe que ele pertence à classe ω_i

$p(\underline{x})$: função densidade de probabilidade em $\underline{x}$

Q_N : conjunto de padrões com classificação conhecida; conjunto de treinamento

R : risco médio ou risco de Bayes

S_x : estimador não viciado da matriz Σ_x

Σ_x : matriz de covariância de um vetor aleatório $\underline{x}$, com elementos σ_{ij}

T : matriz de espalhamento total

$tr(A)$: traço da matriz $A(d \times d) = \sum_{i=1}^{d} a_{ii}$

U_N : conjunto de padrões com classificação a determinar

$\underline{v}_o$: vetor peso, utilizado em funções de decisão lineares

v_{d+1} : limiar, utilizado em funções de decisão lineares

$\underline{v}$: vetor peso aumentado, $\underline{v} = \left[\underline{v}_o^T \quad v_{d+1}\right]^T$; utilizado em funções de decisão lineares

W : matriz de espalhamento intra-classes ou intra-agrupamentos

ω_i : i-ésima classe, $i = 1, 2, \ldots c$

$\omega(\underline{x})$: regra de decisão; atribuição que um dado classificador dá ao vetor $\underline{x}$, isto é, $\omega(\underline{x}):\mathbb{R}^d \longrightarrow \{\omega_0, \omega_1, \ldots, \omega_c\}$

Ω : espaço de atributos com dimensão d

Ω_i : região do espaço de atributos em que um vetor $\underline{x}$ é classificado em ω_i.

$\underline{x}$: vetor de atributos, medidas ou variáveis, indica um padrão

$\underline{x}_a$: vetor de atributos aumentado $\underline{x}_a = [\underline{x}^T \ 1]^T$; utilizado em funções de decisão lineares

$\overline{\underline{x}}$: média de amostras de $\underline{x}$, estimando $\underline{\mu}_x$

$\underline{x}_{ik}$: k-ésimo vetor representativo da i-ésima classe ou agrupamento, com $k = 1,\ldots, n_i$, onde n_i é o número de vetores pertencentes à i-ésima classe ou agrupamento.

$\overline{\underline{x}}_i$: média dos vetores $\underline{x}$ pertencentes à classe ou agrupamento ω_i

$||\underline{x}||$: norma do vetor $\underline{x}$

⊙⊙⊙⊙⊙ : indica o início ou final de um exemplo.

TEORIA DE DECISÃO BAYESIANA

INTRODUÇÃO

Este capítulo apresenta um formalismo matemático que permite obter uma regra de decisão ótima para um classificador, baseado no conhecimento estatístico completo da fonte que gera os vetores de padrões ou de atributos. Este classificador ótimo, também conhecido como classificador de Bayes, tem também a utilidade de fornecer uma medida de desempenho que pode servir como referência de comparação para classificadores projetados por outros métodos (sub-ótimos), mas, geralmente, de maior relevância ou exeqüibilidade prática. Além do mais, se, em uma dada aplicação, a estimativa da probabilidade de erro usando um classificador de Bayes é maior do que o erro desejado, sabe-se que é necessário retornar ao problema inicial e se obter medidas melhores e/ou incluir medidas de atributos adicionais.

Modelos probabilísticos são úteis para se descrever ou representar as variabilidades encontradas nos padrões gerados por uma fonte. Estas variabilidades podem tanto ser intrínsecas quanto extrínsecas:

- Variabilidades
  - (*) intrínsecas
    - devidas aos mecanismos de geração dos padrões
  - (*) extrínsecas
    - ruídos de medição (p.ex., gerados por instrumentação)
    - forma não padronizada ou não padronizável de aquisição (p.ex., posição de eletrodos na captação de ECG, EMG e EEG)

A teoria de decisão estatística provê uma base tanto para modelar o

mecanismo de geração de padrões quanto para formalizar o processo de decisão.

ELEMENTOS BÁSICOS DO PROBLEMA DE CLASSIFICAÇÃO

Apresentaremos inicialmente uma formulação matemática para descrever ou modelar uma dada fonte que gera padrões. Em seguida é feita uma formalização de processos de decisão estatística, incluindo-se uma discussão de índices de mérito que são fundamentais para uma avaliação do desempenho de classificadores (sistemas de decisão).

Representa-se cada padrão por um vetor $\underline{x}$ de atributos ou medidas, sendo que na abordagem estatística $\underline{x}$ é um vetor aleatório de dimensão d. Os vetores de atributos pertencem a um espaço Ω de atributos d-dimensional (normalmente $\Omega = \mathbb{R}^d$).

Um dado vetor $\underline{x}$ pode provir de, ou ser associado a, uma de **c** classes $\omega_1,\omega_2,..,\omega_c$, com uma probabilidade P_i. Esta é a chamada probabilidade a priori de cada classe. É óbvio que $\sum_{i=1}^{c} P_i = 1$.

Define-se $p(\underline{x}|\omega_i)$ como a função densidade de probabilidade multivariada de $\underline{x}$ quando se sabe que ele pertence à classe ω_i (i=1,...c). A função densidade de probabilidade global de $\underline{x}$ é

$$p(\underline{x}) = \sum_{i=1}^{c} P_i \, p(\underline{x}|\omega_i)$$

Os elementos apresentados são suficientes para descrever ou modelar uma grande gama de mecanismos de geração de padrões.

Além das c possíveis classificações para $\underline{x}$, isto é, as classes ω_1 ou ω_2 ou ... ω_c, em certos casos de grande ambigüidade ou em que artefatos e ruídos podem ser confundidos com padrões, pode ser interessante rejeitar o

padrão, classificando-o como pertencente a uma classe ω_o de rejeição, podendo-se eventualmente efetuar um tratamento a posteriori especial para os padrões rejeitados. Nestes casos, teremos c+1 decisões possíveis muito embora haja apenas c classes de fato.

Para modelar os processos de decisão temos que definir o conjunto de possíveis regras de decisão $\omega(\underline{x}) : \Omega \rightarrow \{\omega_o,\omega_1,\omega_2,\ldots\omega_c\}$, bem como o criterio de desempenho de classificação. Para esta última tarefa necessitamos introduzir um quantificador que mede o custo de se cometer um erro de decisão. Indica-se o custo ou perda ou penalidade de se decidir por ω_i, quando a classe verdadeira é ω_j, como

$$C(\omega_i|\omega_j) = C_{ij} \qquad i = 0, 1,\ldots,c \; ; \; j = 1,\ldots,c$$

Estes custos se referem a cada par ω_i/ω_j, onde o primeiro elemento (ω_i na notação acima) é dado pelo classificador e o segundo elemento (ω_j) é definido pela fonte geradora de padrões. Portanto, devemos atribuir um custo médio para uma dada decisão ω_i independentemente de se especificar qual a classe verdadeira. Por sua vez, este custo médio por decisão pode ser utilizado como um critério para escolha das regras de decisão $\omega(.)$.

REGRA DE DECISÃO DE BAYES PARA MÍNIMO CUSTO TOTAL

É dado um padrão $\underline{x}$ com classificação desconhecida. A probabilidade de $\underline{x}$ ser da classe ω_j é $P(\omega_j|\underline{x})$, que é a probabilidade a posteriori da classe ω_j (a probabilidade a priori é P_j). Pela regra de Bayes:

$$P(\omega_j|\underline{x}) = \frac{p(\underline{x}|\omega_j)P_j}{p(\underline{x})} \qquad (1)$$

com $$p(\underline{x}) = \sum_{j=1}^{c} p(\underline{x}|\omega_j)P_j \qquad (2)$$

Como já vimos, o custo associado com a decisão $\omega(\underline{x}) = \omega_i$ i=0, 1,...c, quando a classe correta (ou que ocorreu) é ω_j, é C_{ij}, sendo que este custo ocorrerá com probabilidade $P(\omega_j|\underline{x})$ para j=1,...c .

O custo esperado condicional, ou custo médio condicional, ou risco condicional associado à decisão $\omega(\underline{x}) = \omega_i$, para o dado $\underline{x}$ pode assumir os valores:

$$C_i(\underline{x}) = \sum_{j=1}^{c} C_{ij}\, P(\omega_j|\underline{x}) \qquad i = 0, 1,....c \qquad (3)$$

Tomando uma decisão arbitrária $\omega(\underline{x})$ temos o custo médio condicional ou risco condicional

$$C_\omega(\underline{x}) = \sum_{j=1}^{c} C(\omega(\underline{x})|\omega_j)P(\omega_j|\underline{x}) \qquad (4)$$

ressaltando-se que este $C_\omega(\underline{x})$ poderia ser escrito como $C(\omega(\underline{x}))$ para evidenciar a sua dependência na função de decisão que por sua vez é função de $\underline{x}$ (vide Fig.1). Como $C_\omega(\underline{x})$ é uma variável aleatória pois depende do vetor aleatório $\underline{x}$, podendo tomar os valores $C_0(\underline{x})$, $C_1(\underline{x}),....,C_c(\underline{x})$, utiliza-se o seu valor esperado como uma medida global de desempenho. Define-se, então, como custo total (ou custo médio), ou risco médio, ao valor médio de $C_\omega(\underline{x})$

$$R_\omega = \int_\Omega C_\omega(\underline{x})p(\underline{x})d\underline{x} \qquad (5)$$

Utiliza-se como critério de otimalidade para escolha das regras de decisão $\omega(\underline{x})$ o mínimo de R . Como $p(\underline{x})$ é não negativo (vide (5)), e não depende da

decisão $\omega(\underline{x})$ escolhida, este critério é equivalente a se ter o mínimo de $C_{\omega}(\underline{x})$ para cada $\underline{x}$. Como $\omega(\underline{x})$ em (4) pode assumir as classes ω_o , ω_1 ,...ω_c, chega-se, olhando em (3), à regra de decisão de Bayes $\omega^*(\underline{x})$:

$$\omega^*(\underline{x}) = \omega_i \qquad \text{se} \qquad C_i(\underline{x}) \leq C_k(\underline{x}) \quad , \text{ para } \forall\ k, \quad i,k = 0, 1,...,c \tag{6}$$

Uma regra de decisão basicamente fornece uma partição do espaço de atributos Ω de tal forma que as regiões Ω_i definem o conjunto dos $\underline{x}$ que são classificados em ω_i, para i=0,1,..,c. Tem-se $\Omega = \Omega_o \cup \Omega_1 \cup .....\cup \Omega_c$. Caso o custo de rejeição seja relativamente grande, o classificador não utilizará a opção de rejeição, e portanto $\Omega_o = \varnothing$, e $\Omega = \Omega_1 \cup \Omega_2 \cup .... \cup \Omega_c$. No caso geral, em que há região de rejeição Ω_o, pode-se definir uma região de aceitação global $\Omega_a = \bigcup_{i=1}^{c} \Omega_i$ bem como uma taxa ou probabilidade de aceitação:

$$TA = \sum_{i=1}^{c} \int_{\Omega_i} p(\underline{x})d\underline{x}$$

A taxa de rejeição TR é definida como $TR = \int_{\Omega o} p(\underline{x})d\underline{x}$, com TR+TA=1.

A Regra de Decisão de Bayes fornece o mínimo custo médio condicional

$$C^*(\underline{x}) = \min_{i=0\ ,\ 1,\ ..c} C_i(\underline{x}) = \min_{i=0\ ,\ 1,\ ..c} \sum_{j=1}^{c} C_{ij}\ P(\omega_j|\underline{x}) \tag{7}$$

bem como o mínimo custo total (chamado de risco de Bayes)

$$R^* = \int_{\Omega} C^*(\underline{x})p(\underline{x})d\underline{x} \tag{8}$$

Como o que se conhece a priori é P_i e $p(\underline{x}|\omega_i)$ (e por conseguinte $p(\underline{x})$) pode-se escrever

$$C_{\omega}(\underline{x}) = \frac{1}{p(\underline{x})} \sum_{j=1}^{c} C\,(\omega(\underline{x})|\omega_j)p(\underline{x}|\omega_j)P_j \qquad (9)$$

$$C_{\omega}(\underline{x}) = \frac{1}{p(\underline{x})} \sum_{j=1}^{c} C_{ij}p(\underline{x}|\omega_j)P_j \quad , \; i=0,1,..c \qquad (10)$$

e como $p(\underline{x})$ não depende das classes então

$$C^{*}(\underline{x}) = \min_{i=0,1,...c} C_i(\underline{x}) = \frac{1}{p(\underline{x})} \min_{i=0,1,..c} \sum_{j=1}^{c} C_{ij}p(\underline{x}|\omega_j)P_j \qquad (11)$$

Para ilustrar o que foi visto, tomemos um exemplo simples em que $P_1<P_2$.Suponhamos que $C_{11}= C_{22}= 0$ e que $C_{12} < C_{21}$, isto é, o custo de se decidir pela classe ω_2 quando o correto é a classe ω_1 é maior do que o caso contrário. A Fig. 1a mostra as funções densidade ponderadas pelas respectivas probabilidades de classe. Nas Figs. 1b e 1c vemos $p(\underline{x})C_1(\underline{x})$ e $p(\underline{x})C_2(\underline{x})$, notando-se que o pico de $p(\underline{x})C_2(\underline{x})$ é maior que o de $p(\underline{x})C_1(\underline{x})$. Neste exemplo simples, existem apenas duas regiões de decisão no espaço Ω unidimensional, com o limiar em $\underline{x}$ se encontrando no ponto em que $C_1(\underline{x}) = C_2(\underline{x})$ (ou $p(\underline{x})C_1(\underline{x})=p(\underline{x})C_2(\underline{x})$). A Fig. 1d mostra o mínimo custo médio condicional (multiplicado por $p(\underline{x})$) para cada ponto do espaço.

Apresenta-se na Fig. 2 um diagrama de blocos do classificador geral de Bayes utilizando-se diretamente o que é conhecido sobre o gerador de padrões, ou seja, $p(\underline{x}|\omega_i)$ e P_i, bem como os custos C_{ij} estipulados pelo usuário ou projetista do classificador.

RAZÃO DE VEROSSIMILHANÇA

Na teoria de testes de hipótese a razão de verossimilhança tem um papel bastante importante. O problema que está sendo abordado é de múltiplas hipóteses (c-hipóteses), sob o enfoque Bayesiano, e será investigado a

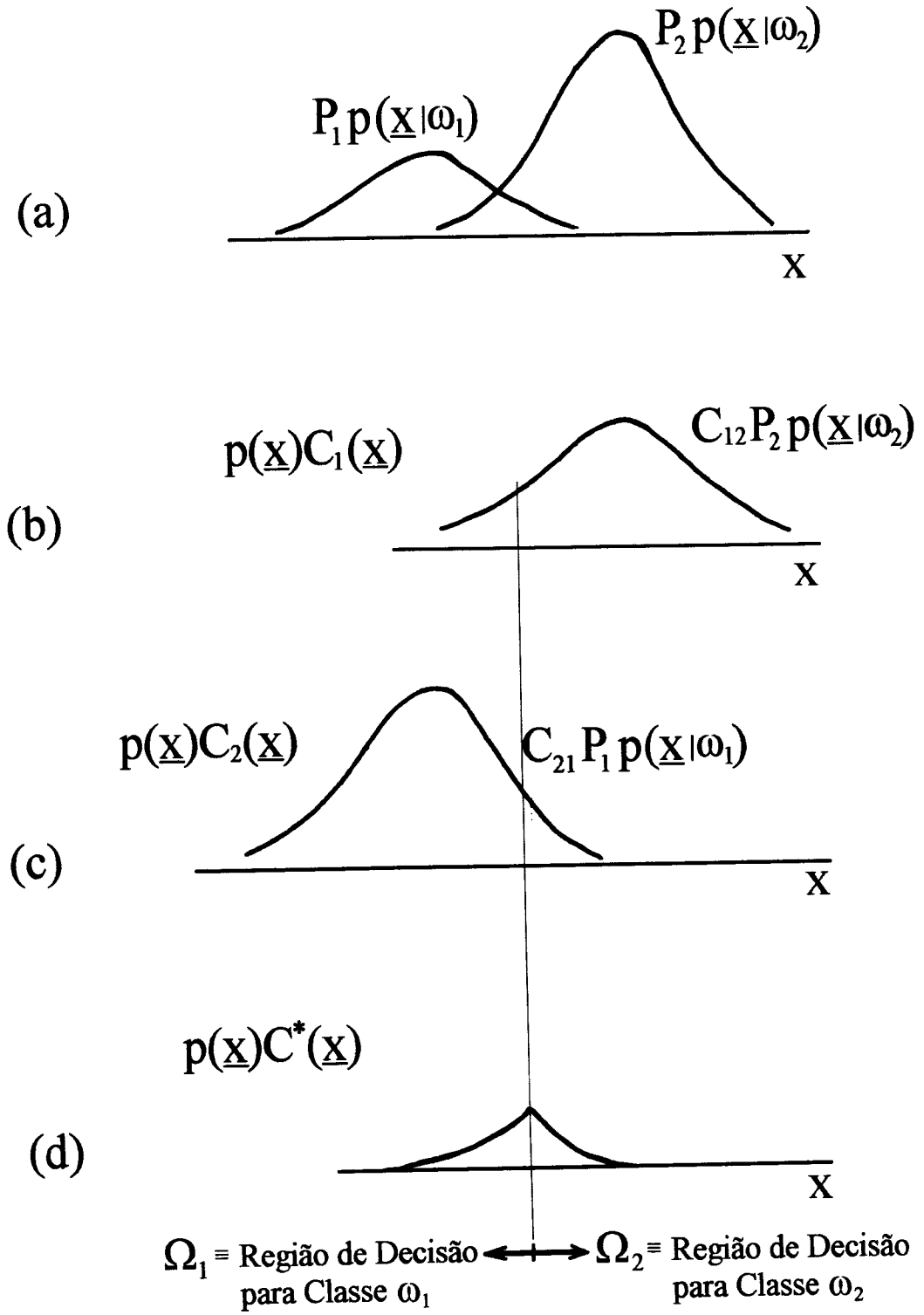


Fig. 1 - Ilustração para o caso unidimensional, com duas classes, do problema de decisão de Bayes.

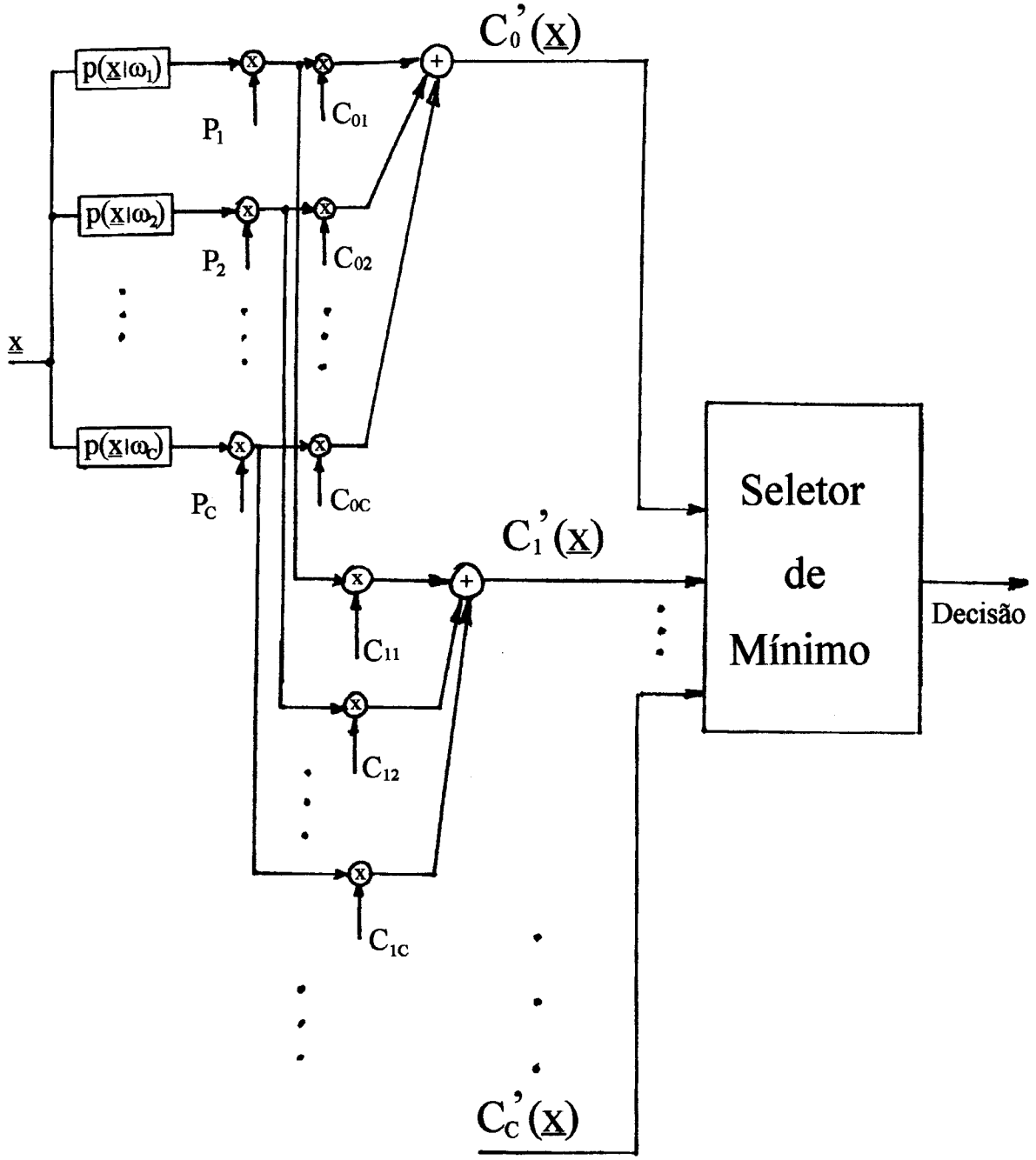


Fig. 2 – Classificador de Bayes, onde $C_i'(x) \triangleq p(x).C_i(x)$

seguir se é possível expressar a regra de decisão através de razão de verossimilhança. Será analisado o caso em que não há classe de rejeição para evitar-se resultados e interpretações artificiosas. Utilizando a expressão (10) na regra de decisão de Bayes dada por (6) tem-se

$\omega^*(\underline{x}) = \omega_i$ se

$$\sum_{j=1}^{c} C_{ij}\ p(\underline{x}|\omega_j)P_j \leq \sum_{m=1}^{c} C_{km}\ p(\underline{x}|\omega_m)P_m \quad , \forall\ k \quad k,i = 1,..,c \tag{12}$$

que expressa a condição a ser satisfeita para $k = 1,\ldots,c$ de modo a se optar pela classe ω_i , com $i = 1,\ldots,c$. Equivalentemente:

$\omega^*(\underline{x}) = \omega_i$ se

$$\sum_{\substack{j=1 \\ j\neq i}}^{c} \left[C_{ij} - C_{kj}\right] p(\underline{x}|\omega_j)P_j \leq \left[C_{ki} - C_{ii}\right] p(\underline{x}|\omega_i)P_i \quad , \forall\ k \quad k=1,\ldots,c \tag{13}$$

Dividindo (13) por $p(\underline{x}|\omega_k)P_k$ (supostos diferentes de zero), obtém-se :

$$\left[C_{ik} - C_{kk}\right] + \sum_{\substack{j=1 \\ j\neq i \\ j\neq k}}^{c} \left[C_{ij} - C_{kj}\right] \frac{p(\underline{x}|\omega_j)P_j}{p(\underline{x}|\omega_k)P_k} \leq \left[C_{ki} - C_{ii}\right] \frac{p(\underline{x}|\omega_i)P_i}{p(\underline{x}|\omega_k)P_k} \tag{14}$$

Finalmente, dividindo (14) por $\left[C_{ki} - C_{ii}\right]P_i/P_k$ obtém-se no lado esquerdo a razão de verossimilhança para o par de classes ω_i e ω_k com $k = 1,..c$ $(k\neq i)$, ou seja, a regra de decisão de Bayes é

$\omega^*(\underline{x}) = \omega_i$ se

$$\frac{p(\underline{x}|\omega_i)}{p(\underline{x}|\omega_k)} \geq \left[\frac{C_{ik}-C_{kk}}{C_{ki}-C_{ii}}\right] \cdot \frac{P_k}{P_i} + \tag{15}$$

$$+ \sum_{\substack{j=1 \\ j\neq i \\ j\neq k}}^{c} \left[\frac{C_{ij}-C_{kj}}{C_{ki}-C_{ii}}\right] \frac{p(\underline{x}|\omega_j)P_j}{p(\underline{x}|\omega_k)P_i} \qquad \forall k,\ k=1,..,c$$

Da expressão (15) nota-se que somente se obtém uma expressão no lado direito independente de outras razões de verossimilhança se

$$C_{ij} = C \quad \text{para} \quad j,i = 1, 2,...c \ , \ \text{com} \ j\neq i.$$

Somente neste caso é que a regra de decisão de Bayes pode ser escrita na forma tradicional de razão de verossimilhança:

$$\omega^*(\underline{x}) = \omega_i \quad \text{se} \quad \frac{p(\underline{x}|\omega_i)}{p(\underline{x}|\omega_k)} \geq \left[\frac{C-C_{kk}}{C-C_{ii}}\right] \frac{P_k}{P_i} \quad \text{para } \forall\ k,\ k\neq i \quad k,i = 1,..,c \tag{16}$$

A desigualdade em (16) ainda pode ser simplificada se $C_{kk} = C' \neq C$, para todo k , o que é em geral adequado em aplicações práticas:

$$\frac{p(\underline{x}|\omega_i)}{p(\underline{x}|\omega_k)} \geq \frac{P_k}{P_i} \qquad \text{para } k=1,..., c \ ; \ k\neq i \tag{17}$$

FUNÇÃO CUSTO SIMÉTRICA E A REGRA DE DECISÃO DE BAYES

Uma função custo que tem grande importância teórica e prática é a assim chamada função custo simétrica:

$$C_{ii} = 0 \qquad i = 1,\ 2,\ldots,\ c$$

$$C_{ij} = 1 \qquad i,\ j = 1,\ldots,\ c\ ;\ \ i \neq j$$

$$C_{oj} = C_r \geq 0 \qquad j = 1,\ldots,c$$

Com esta função custo, temos:

$$\begin{aligned} C_o(\underline{x}) &= \sum_{j=1}^{c} C_{oj}\, P(\omega_j|\underline{x}) = C_r \sum_{j=1}^{c} P(\omega_j|\underline{x}) = C_r \\ C_i(\underline{x}) &= \sum_{\substack{j=1 \\ j\neq i}}^{c} P(\omega_j|\underline{x}) = 1 - P(\omega_i|\underline{x}) \qquad i=1,\ldots,c \end{aligned} \tag{18}$$

Caso não haja classe de rejeição $\underline{w}_o$, dado um $\underline{x}$, ele será classificado em ω_i se

$$P(\omega_i|\underline{x}) \geq P(\omega_j|\underline{x}) \qquad i,j = 1,\ldots,\ c \tag{19}$$

ou seja, a regra de decisão consiste na maximização da probabilidade a posteriori das classes. Na literatura internacional esta regra de decisão é conhecida como a regra MAP ("maximum a posteriori probability").

Se houver classe de rejeição $\underline{w}_o$, a regra de decisão de Bayes, escrita na forma MAP é :

$$\omega^*(\underline{x}) = \begin{cases} \omega_i \ \text{ se } \ P(\omega_i|\underline{x}) = \max\limits_{j=1,..c} P(\omega_j|\underline{x}) \geq 1-C_r \ , i\neq 0 \\ \omega_o \ \text{ se } \ 1-C_r > \max\limits_{j=1,..c} P(\omega_j|\underline{x}) \leftarrow \text{opção de rejeição está ativa} \end{cases} \tag{20}$$

Exemplificamos a derivação para o caso de decisão ω_i, $i\neq 0$. Para ser tomada a decisão ω_i devemos escolher $\min\limits_i C_i(\underline{x}) < C_r$, o que equivale a $\max\limits_i (-C_i) > -C_r$, ou seja, $\max [(P(\omega_i|\underline{x})-1] > -C_r$, de onde segue o resultado em (20).

Verificaremos a seguir qual uma condição básica para se ter a opção de rejeição ativada. Dado um vetor $\underline{x}$ arbitrário, se as c classes forem equiprováveis a posteriori, teremos $\max_j P(\omega_j|\underline{x}) = \frac{1}{c}$ e se as classes não forem

equiprováveis então teremos $\max_j P(\omega_j|\underline{x}) > \frac{1}{c}$ e portanto

$$0 \leq 1 - \max_j P(\omega_j|\underline{x}) < \frac{c-1}{c} \tag{21}$$

Para que a opção de rejeição ω_o possa eventualmente estar ativa é necessário que (vide (20) e (21)):

$$0 \leq C_r < \frac{c-1}{c} \tag{22}$$

Suporemos a seguir que (22) é válido. A regra de decisão (20) particiona o espaço de atributos Ω em c regiões de decisão com aceitação, Ω_i (i=1,...c) , e uma região de decisão com rejeição, Ω_o .

$$\Omega_i = \left\{ \underline{x} \mid P(\omega_i|\underline{x}) = \max_{j=1,\ldots c} P(\omega_j|\underline{x}) \geq 1-C_r \right\} \quad ,i\neq 0 \tag{23}$$

$$\Omega_o = \left\{ \underline{x} \mid 1-C_r > \max_{j=1,\ldots c} P(\omega_j|x) \right\} \tag{24}$$

Temos ainda

$$\Omega_a \cup \Omega_o = \Omega \qquad \text{com} \qquad \Omega_a = \bigcup_{i=1,\ldots c} \Omega_i$$

onde Ω_a é a região de aceitação global.

Deve-se ressaltar que mesmo que (22) seja válida, isto não significa que a opção de rejeição será de fato utilizada, pois pode ser que para todo $\underline{x}$ tenhamos $\max_{j=1\ldots c} P(\omega_j|\underline{x}) \geq 1 - C_r$.

Pode-se também exprimir a regra de decisão em (20) em termos da função de verossimilhança (na classe ω_i) $p(\underline{x}|\omega_i)$ e de P_i conhecidos.

$$\omega^*(\underline{x}) = \begin{cases} \omega_i \quad \text{se} \quad p(\underline{x}|\omega_i)P_i \geq \underset{j=1,\ldots,c}{p(\underline{x}|\omega_j)P_j} \geq (1-C_r)p(\underline{x}) \qquad (i=1,\ldots,c) \\ \\ \omega_o \quad \text{se} \quad (1-C_r)p(\underline{x}) > \underset{j=1,\ldots c}{p(\underline{x}|\omega_j)P_j} \end{cases} \tag{25}$$

A Fig. 3 mostra em forma de diagrama de blocos o classificador de Bayes para função custo simétrica. Deve-se notar que o classificador fica muito mais simples do que quando a função custo é arbitrária (Fig. 2).

A regra de decisão expressa em (25) pode ser expressa em termos da razão de verossimilhança:

$$\omega^*(\underline{x}) = \begin{cases} \omega_i \quad \text{se} \quad \dfrac{p(\underline{x}|\omega_i)}{p(\underline{x}|\omega_j)} \geq \dfrac{P_j}{P_i} \quad \text{e} \quad \dfrac{(1-C_r)p(\underline{x})}{p(\underline{x}|\omega_j)P_j} \leq 1 \ ,\forall j \ (j=1,..c) \\ \\ \omega_o \quad \text{se} \quad \dfrac{(1-C_r)p(\underline{x})}{p(\underline{x}|\omega_j)P_j} > 1 \quad p/ \quad j=1,\ldots c \end{cases}$$

que é a mesma regra de decisão que (17). Isto é razoável pois a função custo simétrica é um caso particular da função custo $C_{ij}=C$ para $i \neq j$ e $C_{ii} = C'$ para todo i .

A regra de decisão Bayesiana, utilizando função custo simétrica, pode ser escrita ainda utilizando-se funções de decisão, sendo mais prático definí-las pelo logaritmo natural da função de verossimilhança e da probabilidade de classe:

$$d_i(\underline{x}) = \ell n \ p(\underline{x}|\omega_i) + \ell n \ P_i \qquad i=1....c$$

$$d_o(\underline{x}) = \ell n \ (1-C_r) + \ell n \ p(\underline{x})$$

Note que as desigualdades em (25) não se alteram se tomarmos o logaritmo em ambos os lados pois a função ℓn é monotônica. Resulta então

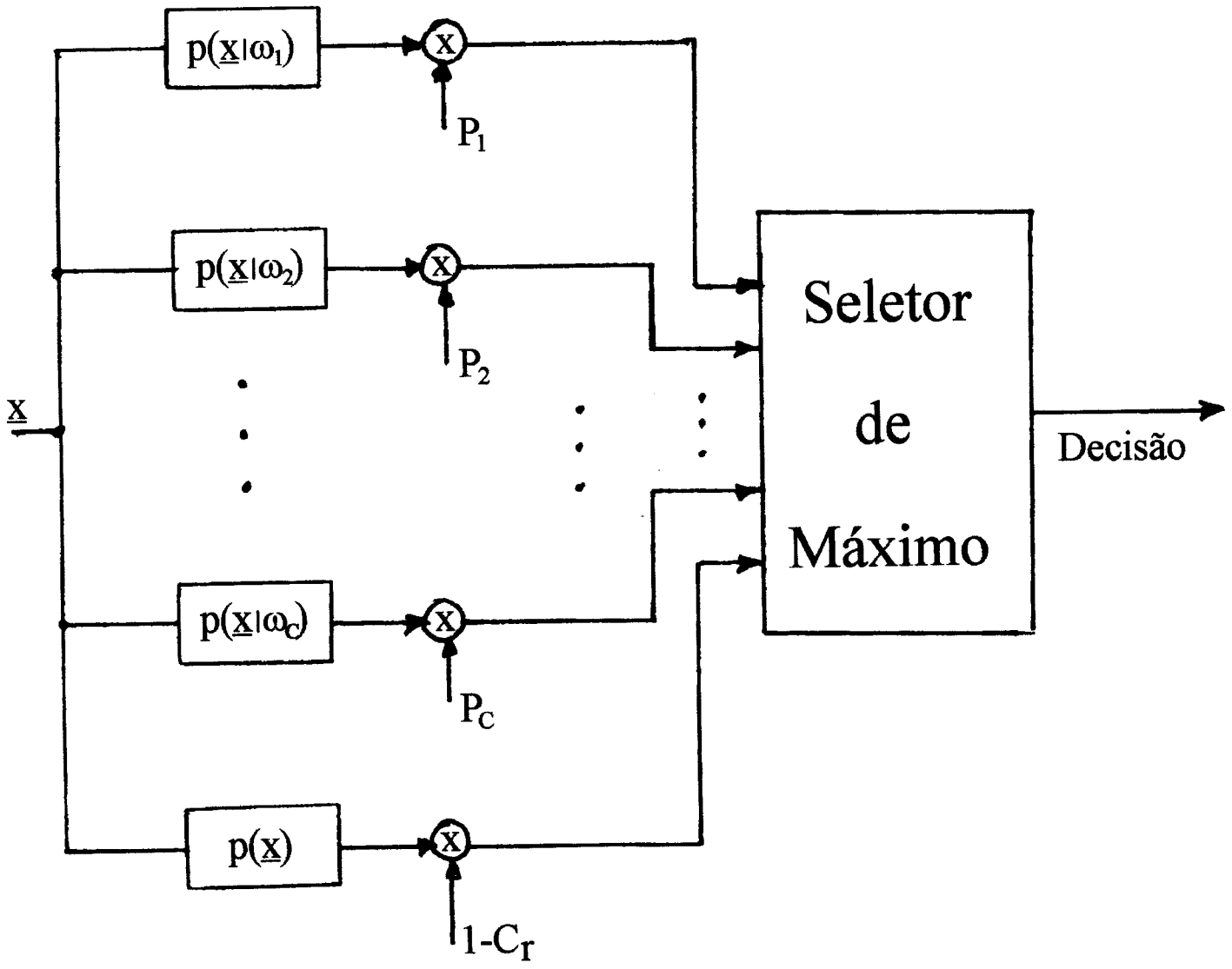


Fig. 3 – Classificador de Bayes para função custo simétrica

$$\omega^*(\underline{x}) = \left\{ \quad \omega_i \quad \text{se} \quad d_i(\underline{x}) \geq d_j(\underline{x}) \quad ,\forall j, \text{ com } j,i=0, 1,\ldots,c \right.$$

ou seja,

$$\omega^*(\underline{x}) = \omega_i \quad \text{se} \quad d_i(\underline{x}) = \max_{j=0,1,..,c} d_j(\underline{x})$$

TAXA OU PROBABILIDADE DE ERRO

Se um dado vetor de atributos $\underline{x}$ é associado à classe ω_i, tem-se uma probabilidade de erro de classificação $e_i(\underline{x})$

$$e_i(\underline{x}) = 1 - P(\omega_i|\underline{x}) \qquad i = 1,\ldots,c \qquad (26)$$

Para firmar a intuição é interessante pensar nos casos extremos $P(\omega_i|\underline{x}) = 1$, em que $e_i(\underline{x})=0$, ou $P(\omega_i|\underline{x})=0$, em que $e_i(\underline{x})=1$. Notar que essa é uma definição geral, sendo válida para função custo e regra de decisão arbitrárias. O valor esperado desta probabilidade sobre todos os vetores $\underline{x}$ pertencentes à região Ω_i de decisão para a classe ω_i é a probabilidade de classificação errada em ω_i, denotada E_i. Esta é a probabilidade de se estar cometendo um erro ao atribuir um vetor $\underline{x}$ arbitrário à classe ω_i:

$$E_i = \int_{\Omega_i} e_i(\underline{x})\, p(\underline{x})\, d\underline{x} = \int_{\Omega_i} [1 - P(\omega_i|\underline{x})]\, p(\underline{x})\, d\underline{x} \qquad (27)$$

Deve-se ressaltar que E_i não é uma probabilidade de erro condicionada à ocorrência da classe ω_i, mas sim a probabilidade de erro de se classificar um vetor de atributos na classe ω_i. Como a classificação de um vetor $\underline{x}$ só pode ocorrer nas classes mutuamente exclusivas ω_1, ω_2, ..., ω_c, segue que a probabilidade global de erro, ou taxa de erro, é a soma das probabilidades

de erro E_i em cada classe:

$$E = \sum_{i=1}^{c} E_i = \sum_{i=1}^{c} \int_{\Omega_i} \left[1-P(\omega_i|\underline{x})\right] p(\underline{x})d\underline{x} \qquad (28)$$

onde Ω_i é a região de aceitação associada à classe ω_i ; a expressão entre colchetes é a probabilidade condicional de erro $e_i(\underline{x})$; E_i é a média desta probabilidade para todo $\underline{x} \in \Omega_i$ e portanto é a probabilidade de classificação errada em ω_i .

Pode-se chegar a uma expressão alternativa para a taxa de erro, partindo-se da determinação da probabilidade de erro dado que a classe que ocorreu foi a ω_i:

$$P(\text{erro}|\omega_i) = \int_{\substack{\text{região de} \\ \Omega \text{ que não inclui } \Omega_i}} p(\underline{x}|\omega_i)d\underline{x} = 1 - \int_{\Omega_i} p(\underline{x}|\omega_i)d\underline{x} \qquad (29)$$

$$i = 1, 2, \ldots, c$$

A probabilidade global de erro é calculada somando as probabilidades conjuntas $P_i P(\text{erro}|\omega_i)$ para todas as classes de 1 a c :

$$E = \sum_{i=1}^{c} P_i \, P(\text{erro}|\omega_i) = \sum_{i=1}^{c} P_i \left[1 - \int_{\Omega_i} p(\underline{x}|\omega_i)d\underline{x} \right] \qquad (30)$$

onde Ω_i é a região de aceitação da classe ω_i; a expressão entre colchetes é a probabilidade de erro dado que a classe correta é ω_i.

Deve-se notar que se houver utilização de região de rejeição teremos $\Omega_a = \bigcup_{i=1,..,c} \Omega_i \neq \Omega$, sendo que $\Omega = \left(\bigcup_{i=1,..,c} \Omega_i \right) \cup \Omega_o = \Omega_a \cup \Omega_o$.

É fácil ver que (28) e (30) são equivalentes. De fato, de (30):

$$E = \sum_{i=1}^{c} P_i - \sum_{i=1}^{c} P_i \int_{\Omega_i} p(\underline{x}|\omega_i)d\underline{x} \ , \qquad \text{de onde}$$

$$E = 1 - \sum_{i=1}^{c} \int_{\Omega_i} P(\omega_i|\underline{x})p(\underline{x})d\underline{x} \qquad \text{e portanto}$$

$$E = \sum_{i=1}^{c} \int_{\Omega_i} p(\underline{x})d\underline{x} - \sum_{i=1}^{c} \int_{\Omega_i} P(\omega_i|\underline{x})p(\underline{x})d\underline{x}$$

o que prova o que queríamos.

Infelizmente o cálculo da probabilidade de erro é, em geral, extremamente difícil e em raríssimos casos é que se consegue chegar a uma expressão em forma explícita ("closed form"). Na prática a taxa de erro é geralmente estimada a partir de um conjunto de amostras (de vetores) com classificação conhecida (conjunto de teste) uma vez que raramente se conhecem as informações probabilísticas necessárias para a determinação pelas fórmulas (quer resolvidas analiticamente ou numericamente).

De posse da formalização da taxa ou probabilidade de erro, podemos perguntar qual será o classificador que minimiza este quantificador de desempenho.

CLASSIFICADOR PARA MÍNIMA TAXA DE ERRO

Faremos inicialmente definições duais às feitas nas fórmulas (26), (27) e (28), ao se enfocar probabilidades de acerto ao invés de probabilidades de erro. A probabilidade de acerto ao se classificar um dado $\underline{x}$ em ω_i é

$$a_i(\underline{x}) = P(\omega_i|\underline{x}) \qquad i=1,\ldots,c$$

A probabilidade de acerto ao se atribuir um vetor à classe ω_i é

$$A_i = \int_{\Omega_i} a_i(\underline{x})\ p(\underline{x})d\underline{x} = \int_{\Omega_i} P(\omega_i|\underline{x})p(\underline{x})d\underline{x}$$

A probabilidade de classificação correta ou probabilidade de acerto ou taxa de acerto é

$$\mathbf{A} = \sum_{i=1}^{c} \int_{\Omega_i} P(\omega_i|\underline{x})p(\underline{x})d\underline{x} \qquad (31)$$

A mínima taxa de erro é obtida para a máxima taxa de acerto

$$\min E \Leftrightarrow \max_{\Omega_i} \sum_{i=1}^{c} \int_{\Omega_i} P(\omega_i|\underline{x})p(\underline{x})d\underline{x}$$

Esta é obtida quando cada Ω_i é escolhido como o domínio onde $P(\omega_i|\underline{x}) \geq P(\omega_j|\underline{x})$, $\forall j$, pois se não o for, $P(\omega_i|\underline{x})$ não seria o máximo em alguma parte de Ω_i e neste caso seria possível aumentar **A** escolhendo o $P(\omega_j|x)$ $(j \neq i)$ majorante naquela parte de Ω_i. A Fig. 4 ilustra o fato para um caso de 2 classes e dimensão d=1, observando-se que no caso de se selecionar o limiar x_{L2} (ao invés do ótimo x_{L1}) temos

$$\int_{\Omega_{1a}} P(\omega_1|x)p(x)dx + \int_{\Omega_{2a}} P(\omega_2|x)p(x)dx = \int_{\Omega_1} P(\omega_1|x)p(x)dx + \int_{\Omega_{1a}-\Omega_1} P(\omega_1|x)p(x)dx +$$

$$\int_{\Omega_2} P(\omega_2|x)p(x)dx - \int_{\Omega_{1a}-\Omega_1} P(\omega_2|x)p(x)dx < \int_{\Omega_1} P(\omega_1|x)p(x)dx + \int_{\Omega_2} P(\omega_2|x)p(x)dx$$

A desigualdade provém do fato que

$$\int_{\Omega_{1a}-\Omega_1} P(\omega_1|x)p(x)dx < \int_{\Omega_{1a}-\Omega_1} P(\omega_2|x)p(x)dx.$$

Formalizando o resultado acima, temos o classificador de mímina taxa de erro:

$$\omega(\underline{x}) = \omega_i \quad \text{se} \quad \underline{x} \in \Omega_i \,, \quad \text{com} \quad \Omega_i = \left\{ \underline{x} \mid P(\omega_i|\underline{x}) \geq P(\omega_j|\underline{x}),\ j=1,\ldots,c \right\}$$

ou, mais simplesmente,

$$\omega(\underline{x}) = \omega_i \quad \text{se} \quad P(\omega_i|\underline{x}) \geq P(\omega_j|\underline{x}) \qquad j=1,2,\ldots,c$$

Conclui-se, portanto, que a regra de decisão de Bayes com função custo simétrica e sem a existência de classe de rejeição, coincide com a regra de decisão para mínima taxa de erro, ambas resultando na regra de decisão de máxima probabilidade a posteriori (MAP).

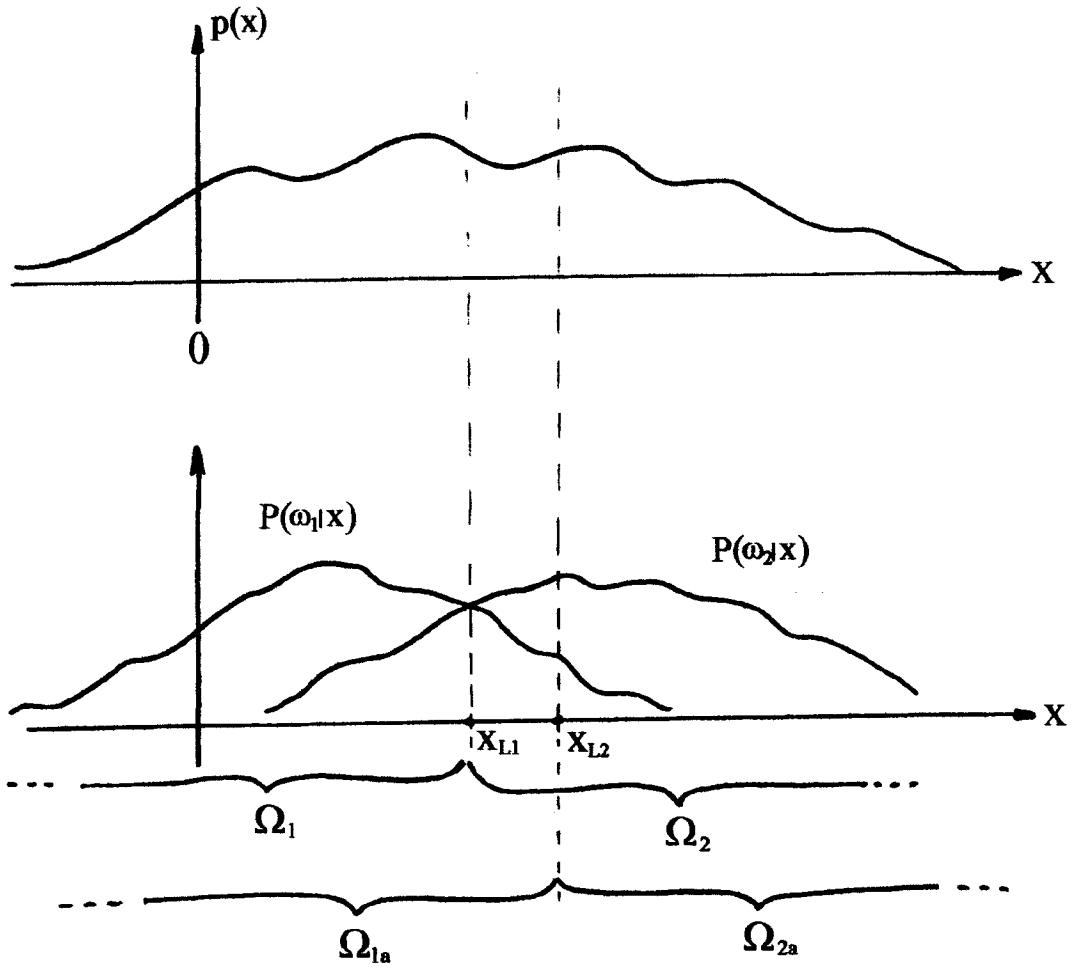


Fig.. 4 – Exemplo unidimensional, com duas classes, em que se mostra que deve ser escolhido o limiar x_{L1}, e não x_{L2} (arbitrário $\neq x_{L1}$), para minimizar a taxa de erro.

RELAÇÃO ENTRE CUSTO TOTAL E TAXA DE ERRO

Na teoria já apresentada, foram empregados dois diferentes quantificadores de desempenho para classificadores arbitrários: o custo total R (expressão 5) e a taxa de erro (expressão 28 ou 30). No caso de função custo arbitrária, um dado classificador apresentará, em geral, regiões diferentes Ωi para cada atribuição de classe e valores diferentes para o custo total e para a taxa de erro. Um classificador ótimo para um dos quantificadores (p.ex. R) não será, em geral, ótimo para o outro (E). Entretanto, no caso de um classificador arbitrário (não precisa obedecer a qualquer critério de otimalidade), sem a utilização de classe de rejeição, em que se utiliza a função custo simétrica para o cálculo de R, temos R = E, ou seja, os dois quantificadores coincidem. Uma demonstração pode ser vista no que segue:

$$R = \int_{\Omega} \sum_{j=1}^{c} C(\omega(\underline{x})|\omega_j)\ P(\omega_j|\underline{x})\ p(\underline{x})d\underline{x} = \sum_{j=1}^{c} \int_{\Omega-\Omega_j} P(\omega_j|\underline{x})\ p(\underline{x})\ d\underline{x} =$$

$$= \sum_{j=1}^{c} \left[\int_{\Omega} P(\omega_j|\underline{x})p(\underline{x})d\underline{x} - \int_{\Omega_j} P(\omega_j|\underline{x})p(\underline{x})d\underline{x} \right]$$

e como $P(\omega_j|\underline{x})p(\underline{x}) = P_j\ p(\underline{x}|\omega_j)$, temos

$$R = \sum_{j=1}^{c} P_j - \sum_{j=1}^{c} \int_{\Omega_j} P(\omega_j|\underline{x})p(\underline{x})d\underline{x}$$

Por outro lado, temos:

$$E = \sum_{j=1}^{c} \left[\int_{\Omega_j} p(\underline{x})d\underline{x} - \int_{\Omega_j} P(\omega_j|\underline{x})p(\underline{x})d\underline{x} \right] =$$

$$= \sum_{j=1}^{c} \int_{\Omega_j} \sum_{i=1}^{c} P_i p(\underline{x}|\omega_i) d\underline{x} \; - \; \sum_{j=1}^{c} \int_{\Omega_j} P(\omega_j|\underline{x}) p(\underline{x}) d\underline{x}$$

e como

$$\sum_{j=1}^{c} \int_{\Omega_j} \sum_{i=1}^{c} P_i \, p(\underline{x}|\omega_i) d\underline{x} \; = \; \sum_{i=1}^{c} P_i \sum_{j=1}^{c} \int_{\Omega_j} p(\underline{x}|\omega_i) d\underline{x} \; =$$

$$= \sum_{i=1}^{c} P_i \int_{\Omega} p(\underline{x}|\omega_i) d\underline{x} \; = \; \sum_{i=1}^{c} P_i$$

temos $E = R.$

A taxa de erro não é um quantificador tão completo quanto o custo total pois: (1) o custo total leva em conta os casos em que se quer dar importância diferente para os diferentes erros do tipo i/j; (2) no caso de haver região de rejeição, o classificador pode apresentar uma taxa de erro baixa em parte graças à associação de um grande número de vetores de atributos à classe de rejeição. Entretanto, este classificador provavelmente terá um desempenho indesejável uma vez que sua tarefa é de classificar (corretamente) os vetores de entrada e não simplesmente "jogar fora" uma grande quantidade destes. Como se pode atribuir um custo a este evento de rejeição, o custo total é um quantificador mais interessante neste caso.

Derivaremos a seguir uma relação entre o custo total, a taxa de erro e a taxa de rejeição para o classificador de Bayes obtido para funcão custo simétrica. Define-se a mínima probabilidade de erro condicionada a $\underline{x}$ como $e^*(\underline{x})$

$$e^*(\underline{x}) \overset{\Delta}{=} \min_{i=1,\ldots c} e_i(\underline{x}) = 1 - \max_{i=1,..c} P(\omega_i|\underline{x}) \qquad (32)$$

que independe da função custo utilizada. Entretanto, para regra de decisão de Bayes com função custo simétrica, $e^*(x)$, dado em (32), é a probabilidade de erro de classificação associada ao vetor $\underline{x}$. Com isto, uma outra forma de

escrever a Regra de Decisão de Bayes para função custo simétrica é :

$$\omega(\underline{x}) = \begin{cases} \omega_i & \text{se} \quad e^*(\underline{x}) = e_i(\underline{x}) \leq C_r \\ \omega_o & \text{se} \quad C_r < e^*(\underline{x}) \end{cases} \tag{33}$$

A taxa de erro associada ao classificador de Bayes com função custo simétrica, também chamada de taxa de Bayes é (vide (28)):

$$E^* = \sum_{i=1}^{c} E_i^* = \sum_{i=1}^{c} \int_{\Omega_i} e^*(\underline{x}) p(\underline{x}) d\underline{x} = \int_{\Omega_a} e^*(\underline{x}) p(\underline{x}) d\underline{x} \tag{34a}$$

ou ainda

$$E^* = \int_{\Omega_a} \{1 - \max_{i=1,..,c} [P(\omega_i|\underline{x})]\} \, p(\underline{x}) d\underline{x} \tag{34b}$$

Pode-se facilmente relacionar o custo total ou risco de Bayes com a taxa de erro no caso de se utilizar função custo simétrica com opção de rejeição, conforme apresentado a seguir:

$$R^* = \int_{\Omega} C^*(\underline{x}) \, p(\underline{x}) \, d\underline{x}$$

onde $C^*(\underline{x}) = \begin{cases} 1 - \max P(\omega_i|\underline{x}) & \text{se} \quad 1 - \max P(\omega_i|\underline{x}) \leq C_r, \; i=1,\ldots,c \\ C_r & \text{em caso contrário} \end{cases}$

e portanto

$$R^* = \int_{\Omega_a(C_r)} e^*(\underline{x}) \, p(\underline{x}) \, d\underline{x} + \int_{\Omega_o(C_r)} C_r \, p(\underline{x}) \, d\underline{x}$$

onde se tentou ressaltar que as regiões de aceitação e rejeição dependem do custo de rejeição C_r. A primeira integral nada mais é que a taxa (de erro) de Bayes e a segunda é proporcional à taxa de rejeição associada à regra de Bayes, TR^*:

$$R^* = E^* + C_r \, TR^*$$

notando-se que todos os termos dependem exclusivamente do custo de rejeição C_r (uma vez adotada a regra de decisão de Bayes). Para enfatizar este fato reescrevemos a última expressão como

$$R^*(C_r) = E^*(C_r) + C_r\, TR^*(C_r) \tag{35}$$

Fixado um certo valor para $TR^*(C_r)$, mediante a escolha de um valor adequado para C_r, não há classificador com taxa de erro menor que $E^*(C_r)$.

Uma nota à parte: no caso de não haver região de rejeição e lembrando que $P(\omega_i|\underline{x})\ p(\underline{x}) = P_i\ p(\underline{x}|\omega_i)$, tem-se

$$E^* = 1 - \int_{\Omega} \max_{i=1,..,c} [P_i\ p(\underline{x}|\omega_i)]\ d\underline{x} \tag{36}$$

E finalmente, de (21) e de (34b) com $\Omega_a = \Omega$ conclui-se que

$$E^* \leq (c-1)/c \tag{37}$$

O EXEMPLO DE DUAS CLASSES COM A FUNÇÃO CUSTO SIMÉTRICA

Caso na dada aplicação com duas classes não seja muito indesejável que o classificador rejeite uma certa proporção de vetores de atributos, pode-se utilizar para ausência de classificação um custo C_r de valor suficientemente baixo (ativando a opção de rejeição). Em termos matemáticos isto seria especificado como $C_r < 1/2$ (vide (21)). Neste caso a regra de decisão de Bayes fica:

$$\omega(\underline{x}) = \begin{cases} \omega_1 & \text{se } P(\omega_1|\underline{x}) \geq 1-C_r > \frac{1}{2} \\ \omega_2 & \text{se } P(\omega_2|\underline{x}) \geq 1-C_r \\ \omega_o & \text{caso contrário} \end{cases} \tag{38}$$

A regra de decisão acima também pode ser escrita em termos da razão de verossimilhança $p(\underline{x}|\omega_1)/p(\underline{x}|\omega_2)$, lembrando que o ℓn desta razão de verossimilhança muitas vezes é de maior utilidade:

$$\omega(\underline{x}) = \begin{cases} \omega_1 & \text{se} \quad \dfrac{p(\underline{x}|\omega_1)}{p(\underline{x}|\omega_2)} \geq \dfrac{P_2}{P_1}\dfrac{1-C_r}{C_r} \\ \omega_2 & \text{se} \quad \dfrac{p(\underline{x}|\omega_1)}{p(\underline{x}|\omega_2)} \leq \dfrac{P_2}{P_1}\dfrac{C_r}{1-C_r} \\ \omega_o & \text{caso contrário} \end{cases} \tag{39}$$

onde há 2 limiares, $P_2(1-C_r)/P_1C_r$ e $P_2C_r/P_1(1-C_r)$,que dividem o espaço em 3 regiões; 2 de aceitação e 1 de rejeição. Para provar (39), basta partir de (38) com

$$P(\omega_1|\underline{x}) = \frac{p(\underline{x}|\omega_1)P_1}{p(\underline{x}|\omega_1)P_1+p(\underline{x}|\omega_2)P_2}.$$

Por outro lado, caso seja indesejável haver ausência de classificação, ou seja, não se deseja rejeitar nenhum vetor de entrada, devemos atribuir um custo alto para uma rejeição, que no caso de duas classes redunda em $C_r > 1/2$. Neste caso a regra de decisão de Bayes fica:

$$\omega(\underline{x}) = \begin{cases} \omega_1 & \text{se} \quad P(\omega_1|\underline{x}) \geq P(\omega_2|\underline{x}) \\ \omega_2 & \text{se} \quad P(\omega_1|\underline{x}) < P(\omega_2|\underline{x}) \end{cases} \tag{40}$$

ou ainda

$$\omega(\underline{x}) = \begin{cases} \omega_1 & \text{se} \quad \dfrac{p(\underline{x}|\omega_1)}{p(\underline{x}|\omega_2)} \geq \dfrac{P_2}{P_1} \\ \omega_2 & \text{se} \quad \dfrac{p(\underline{x}|\omega_1)}{p(\underline{x}|\omega_2)} < \dfrac{P_2}{P_1} \end{cases} \tag{41}$$

onde P_2/P_1 é o limiar de decisão para a razão de verossimilhança, obtendo-se de (41) uma partição do espaço de atributos. A Fig.5a mostra duas funções densidade de probabilidade, cada uma condicionada a uma das classes. Para fins de ilustração tomamos ambas as densidades com variância unitária, sendo

a média da primeira igual a -1.5 e da segunda igual a 1.5. As probabilidades de classe foram supostas ser $P_1 = 0.3$ e $P_2 = 0.7$. A função densidade de probabilidade de $\underline{x}$, independente de classe, é vista na Fig. 5b. Para o classificador de Bayes de mínima taxa de erro (ou mínimo risco médio para função custo simétrica) deve-se tomar como região Ω_1 para classificar x em ω_1 aquela em que $P_1 p(x|\omega_1) > P_2 p(x|\omega_2)$, o que pode ser visto na Fig. 5c. Alternativamente, pode-se achar Ω_1 como sendo a região em x tal que $P(\omega_1|x) > P(\omega_2|x)$, o que pode ser visto na Fig. 5d, notando-se que é obtida a mesma região Ω_1 e o mesmo limiar como na Fig. 5c. Outra forma ainda é selecionar ω_1 se $e_1(x) < e_2(x)$, conforme visto na Fig. 5e. Por fim, pode-se utilizar a razão de verossimilhança e selecionar ω_1 se a razão for maior que o limiar P_2/P_1 , conforme visto na Fig. 5f. A área sob $p(\underline{x}|\omega_1)$ à direita do limiar indica a probabilidade de erro condicionada ao fato de ter ocorrido a classe ω_1. Aproveitaremos este exemplo unidimensional para ilustrar o problema da determinação da taxa de erro E :

$$E = P(\text{erro}|\omega_1).P(\omega_1) + P(\text{erro}|\omega_2).P(\omega_2) \tag{42}$$

$$E = \underbrace{\int_{\Omega_2} p(\underline{x}|\omega_1).P_1 \, d\underline{x}}_{E_2} + \underbrace{\int_{\Omega_1} p(\underline{x}|\omega_2).P_2 \, d\underline{x}}_{E_1} \tag{43}$$

CASO GAUSSIANO

No caso Gaussiano ou normal, em que $p(\underline{x}|\omega_i)$ é $N(\underline{\mu}_i, \Sigma_i)$, a regra de decisão de Bayes fica mais simples conforme visto a seguir. Temos $p(\underline{x}|\omega_i)$ normais com média $\underline{\mu}_i$ e matriz de covariância Σ_i , onde

$$\underline{\mu}_i = E\left[\underline{X}|\omega_i \right] \qquad \text{e} \qquad \Sigma_i = E\left[(\underline{X}-\underline{\mu}_i)(\underline{X}-\underline{\mu}_i)^T \, |\omega_i \right]$$

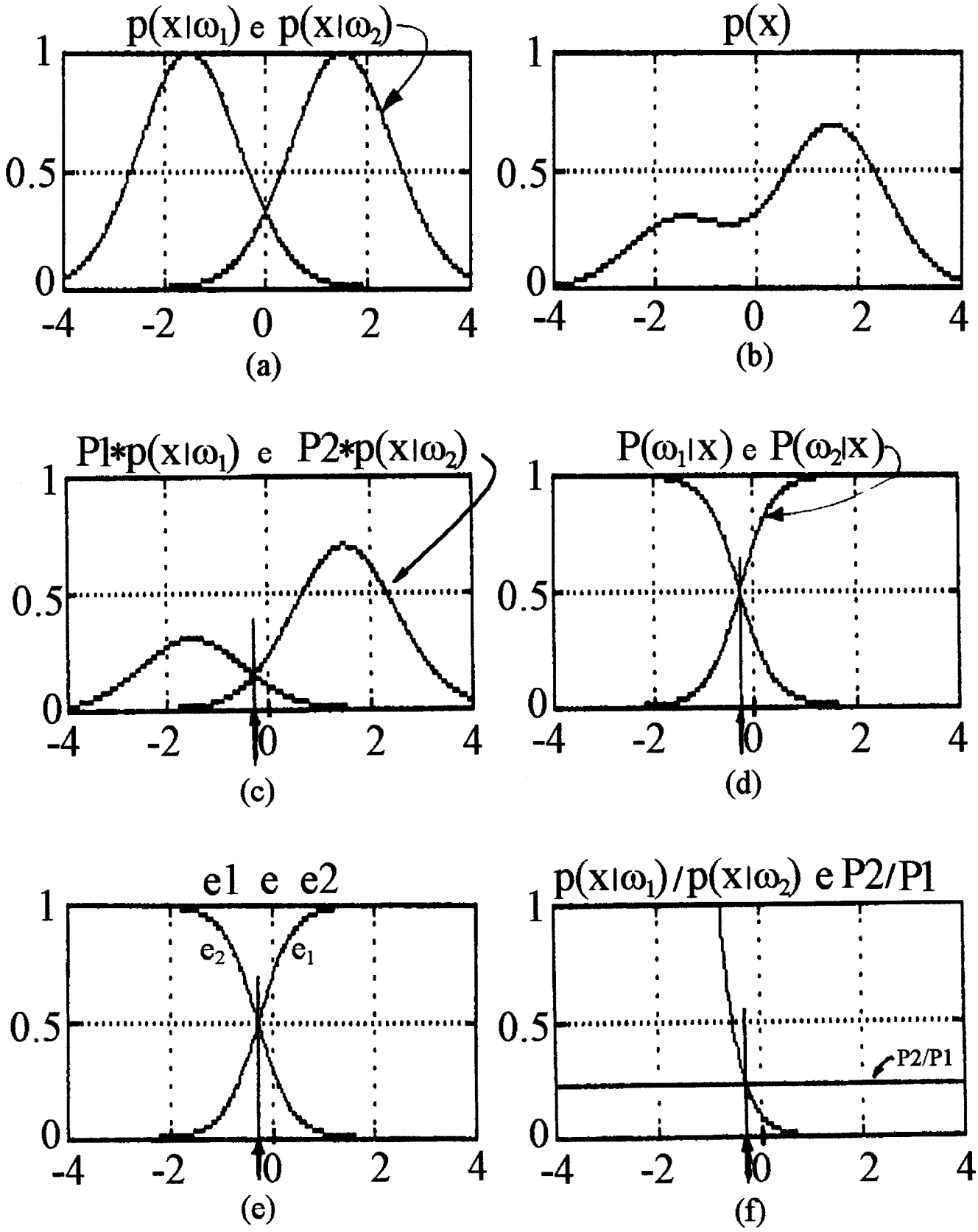

p(x|ω1) e p(x|ω2)
1
0.5
0
-4
-2
0
2
4
(a)
p(x)
(b)
P1*p(x|ω1) e P2*p(x|ω2)
(c)
P(ω1|x) e P(ω2|x)
(d)
e1 e e2
e2
e1
(e)
p(x|ω1)/p(x|ω2) e P2/P1
P2/P1
(f)

Fig. 5 Exemplo unidimensional com 2 classes Gaussianas, com P_1=0.3, P_2=0.7, $p(x|\omega_1) \equiv N(-1.5,1)$ e $p(x|\omega_2) \equiv N(1.5,1)$. A Fig. 5a mostra estas funções densidade condicionadas a cada classe. A função densidade de probalidade de x (independente de classe) é vista na Fig. 5b. Para o classificador de Bayes de mínima taxa de erro (ou de mínimo custo total para função custo simétrica) deve-se tomar como região Ω_1 (para classificar x em ω_1) aquela em que $P_1.p(x|\omega_1) > P_2.p(x|\omega_2)$, o que pode ser visualizado na Fig. 5c. Alternativamente, pode-se achar Ω_1 como sendo a região em x tal que $P(\omega_1|x) > P(\omega_2|x)$, o que pode ser visto na Fig. 5d, notando-se que se obtém a mesma região Ω_1 e mesmo limiar como na Fig. 5c. Outra forma ainda é selecionar ω_1 se $e_1(x) < e_2(x)$, conforme visto na Fig. 5e. Por fim, pode-se usar a razão de verossimilhança, selecionando-se ω_1 se o valor da razão for maior que o limiar P_2/P_1, conforme exemplificado na Fig. 5f.

$$p(\underline{x}|\omega_i) = \frac{1}{(2\pi)^{d/2}\sqrt{\det \Sigma_i}} \exp\left[-\frac{1}{2}(\underline{x}-\underline{\mu}_i)^T\Sigma_i^{-1}(\underline{x}-\underline{\mu}_i) \right] \tag{44}$$

Apliquemos a regra de decisão (41) utilizando o ℓn da razão de verossimilhança.

$$h(\underline{x}) \overset{\Delta}{=} \ell n \frac{p(\underline{x}|\omega_1)}{p(\underline{x}|\omega_2)} = \ell n \left[\frac{\sqrt{\det \Sigma_2}}{\sqrt{\det \Sigma_1}} \frac{\exp\left[-\frac{1}{2}(\underline{x}-\underline{\mu}_1)^T\Sigma_1^{-1}(\underline{x}-\underline{\mu}_1) \right]}{\exp\left[-\frac{1}{2}(\underline{x}-\underline{\mu}_2)^T\Sigma_2^{-1}(\underline{x}-\underline{\mu}_2) \right]} \right] \tag{45}$$

que resulta em :

$$h(\underline{x}) = \frac{1}{2}\ell n\left(\frac{|\Sigma_2|}{|\Sigma_1|} \right) - \frac{1}{2}(\underline{x}-\underline{\mu}_1)^T\Sigma_1^{-1}(\underline{x}-\underline{\mu}_1) + \frac{1}{2}(\underline{x}-\underline{\mu}_2)^T\Sigma_2^{-1}(\underline{x}-\underline{\mu}_2) \tag{46}$$

ou seja, a regra de decisão de Bayes fica, para o caso da opção de rejeição estar desativada:

$$\omega(\underline{x}) = \begin{cases} \omega_1 \quad \text{se} \quad h(\underline{x}) \geq \ell n \dfrac{P_2}{P_1} \\ \omega_2 \quad \text{se} \quad h(x) < \ell n \dfrac{P_2}{P_1} \end{cases} \tag{47}$$

Notamos que a fronteira ou separatriz ("boundary") de decisão ($h(\underline{x})=\ell n P_2/P_1$), neste caso, é uma forma quadrática em $\underline{x}$ (é uma superfície de 2ª ordem no espaço $x_1, x_2, \ldots x_d$).

⊙⊙⊙⊙⊙ Exemplo: Tomemos o caso em que há 2 distribuições normais, com $\underline{\mu}_1 = [-2 \quad 2]^T$, $\underline{\mu}_2 = [2 \quad 2]^T$, $\Sigma_1 = \begin{bmatrix} 1 & 0{,}8 \\ 0{,}8 & 1 \end{bmatrix}$, $\Sigma_2 = \begin{bmatrix} 0{,}5 & 0 \\ 0 & 1 \end{bmatrix}$ As probabilidades a priori são $P_1 = 2 \cdot P_2$ e suporemos que $C_r > 1/2$ (não se deseja opção de rejeição). Aplicando a fórmula (46), obtém-se, após várias passagens algébricas, a equação da separatriz:

$$7x_1^2 - 40x_1x_2 + 252x_1 - 144x_2 + 16x_2^2 + 236.57 = 0$$

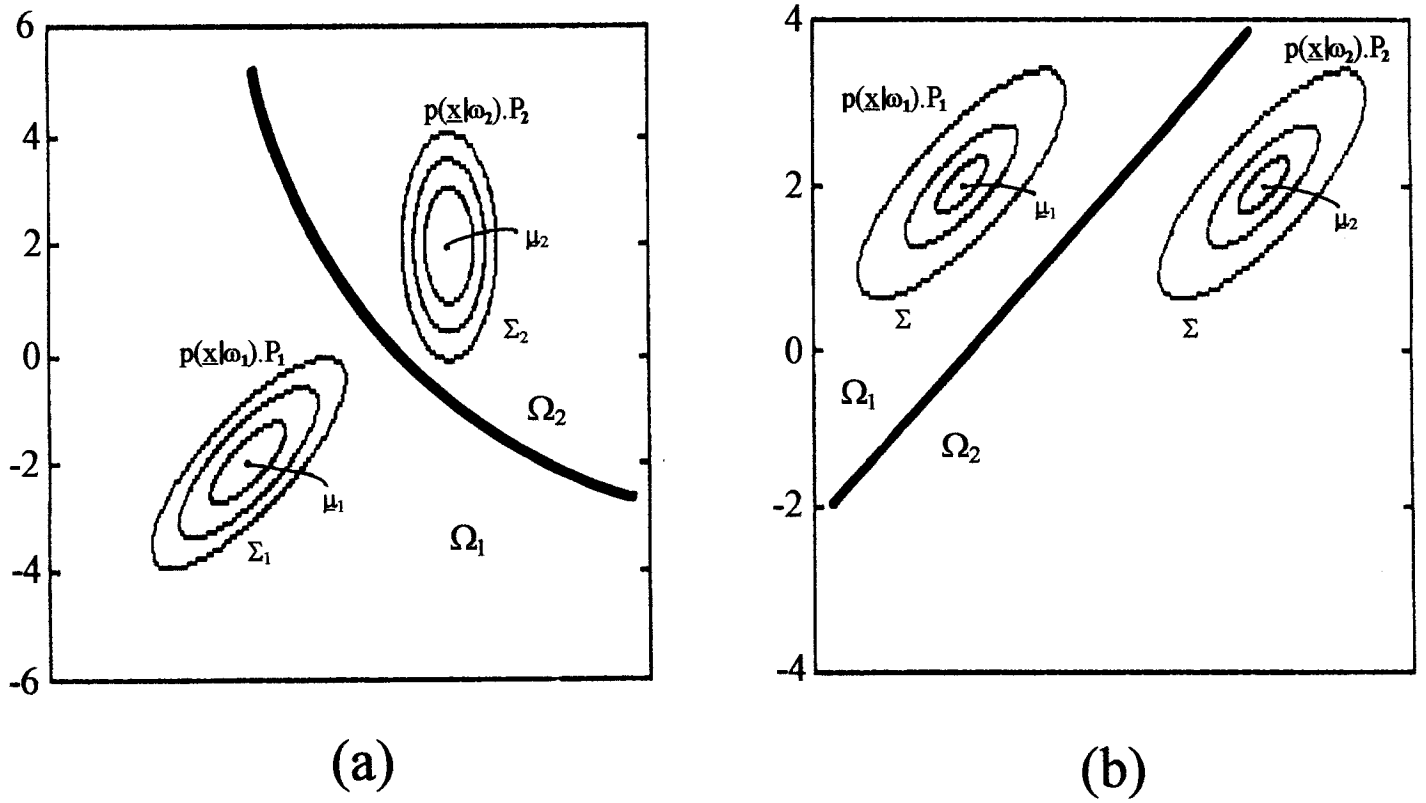


Fig. 6 – Ilustrações de fronteiras de decisão para um exemplo bi-dimensional com duas classes, em que $C_r > \frac{1}{2}$: (a) $\Sigma_1 \neq \Sigma_2$; (b) $\Sigma_1 = \Sigma_2$

A Fig. 6a mostra curvas de nível para as duas funções densidade e a respectiva separatriz imposta pelo classificador de Bayes (para função custo simétrica).

ooooo

Tomemos agora o caso em que $\Sigma_1 = \Sigma_2 = \Sigma$. A expressão de $h(\underline{x})$ em (47) fica:

$$h(\underline{x}) = \underbrace{(\underline{\mu}_1 - \underline{\mu}_2)^T \Sigma^{-1}}_{\text{constante}} \underline{x} + \underbrace{\frac{1}{2}\left[\underline{\mu}_2^T \Sigma^{-1} \underline{\mu}_2 - \underline{\mu}_1^T \Sigma^{-1} \underline{\mu}_1 \right]}_{\text{constante}} \qquad (48)$$

Neste caso a fronteira de decisão é uma forma linear (hiperplano) em $\underline{x}$, conforme ilustrado na Fig. 6b para um exemplo de dimensão 2 em que as matrizes de covariância são iguais.

Para o caso $C_r < 1/2$ conclusões semelhantes resultam, como apresentado a seguir. Definindo limiares T_1 e T_2 como

$$T_1 \overset{\Delta}{=} \ell n \left(\frac{P_2}{P_1} \; \frac{1-C_r}{C_r} \right)$$

$$T_2 \overset{\Delta}{=} \ell n \left(\frac{P_2}{P_1} \; \frac{C_r}{1-C_r} \right)$$

a regra de decisão de Bayes fica

$$\omega(\underline{x}) = \begin{cases} \omega_1 & \text{se} \quad h(\underline{x}) \geq T_1 \\ \omega_2 & \text{se} \quad h(\underline{x}) \leq T_2 \\ \omega_o & \text{se} \quad T_2 < h(\underline{x}) < T_1 \end{cases}$$

A separatriz ou fronteira entre a região Ω_1, definida por $\{ \underline{x} \mid h(\underline{x}) \geq T_1 \}$, e a região de rejeição Ω_o é dada por $h(\underline{x}) = T_1$. A separatriz entre a região Ω_2 e Ω_o é dada por $h(\underline{x}) = T_2$. Nota-se que as 2 separatrizes são paralelas pois a função em $\underline{x}$ é a mesma em ambas. Para $p(\underline{x}|\omega_i)$ normal (i=1,2) com $\Sigma_1 \neq \Sigma_2$ temos um classificador quadrático correspondendo ao exemplo da Fig.7a, e

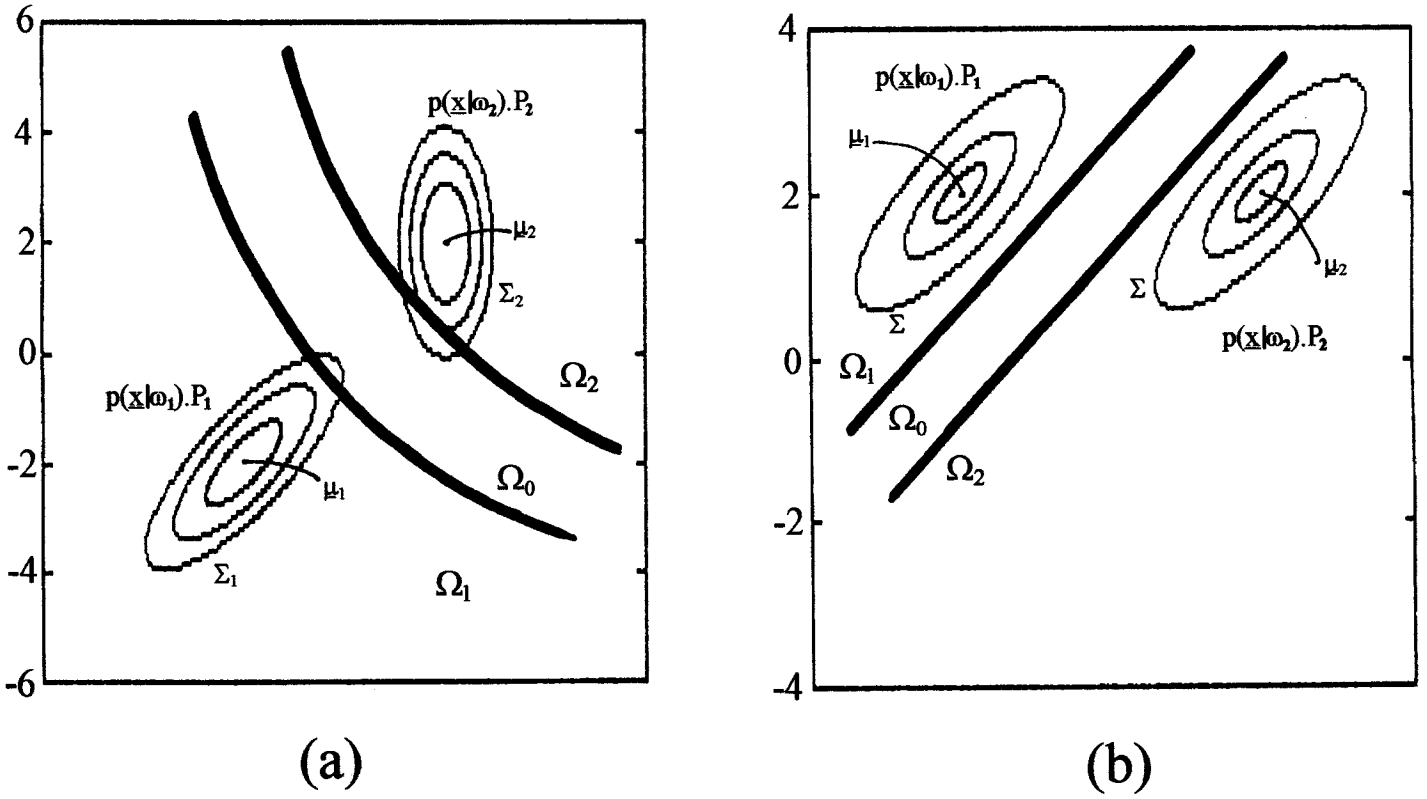


Fig. 7 -Ilustrações de fronteiras de decisão para um exemplo bi-dimensional com duas classes, em que $C_r < \frac{1}{2}$: (a) $\Sigma_1 \neq \Sigma_2$, (b) $\Sigma_1 = \Sigma_2$

para $\Sigma_1 = \Sigma_2$ temos um classificador linear correspondente ao exemplo da Fig.7b.

Uma interpretação útil é obtida expressando a regra de decisão de Bayes para o caso Normal não em termos da razão de verossimilhança mas sim em termos da função de verossimilhança. Por simplicidade, faremos isto para $C_r > 1/2$ e $\Sigma_1 = \Sigma_2 = \Sigma$:

$$\omega(\underline{x}) = \begin{cases} \omega_1 & \text{se} \quad \ell n\ p(\underline{x}|\omega_1) + \ell n\ P_1 > \ell n\ p(\underline{x}|\omega_2) + \ell n\ P_2 \\ \omega_2 & \text{em caso contrário} \end{cases}$$

Se $P_1 = P_2$, basta calcular a distância de Mahalanobis (vide Capítulo (pg 119) sobre Medidas de Distância) $(\underline{x}-\underline{\mu}_1)^T\Sigma^{-1}(\underline{x}-\underline{\mu}_1)$ de $\underline{x}$ a cada um dos vetores médios e classificar de acordo com a menor distância. No caso de se ter $P_1 \neq P_2$ há um deslocamento em favor da classe mais provável.

Um exemplo de matriz de covariância de interesse prático é aquele em que o vetor $\underline{x}$ tem como elementos as amostras x(nT) de um sinal x(t) amostrado a cada T segundos. Se o sinal é um trecho de uma função amostral de um processo aleatório estacionário, e se a função de autocovariância é ρ^n , n = 0,1,...,(ou seja, a variância é unitária) então a matriz de covariância será

$$\Sigma = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{d-1} \\ \rho & 1 & \rho & \cdots & \rho^{d-2} \\ \cdots & \cdots & \cdots & & \\ \rho^{d-1} & \cdots & \cdots & & 1 \end{bmatrix}$$

onde ρ é o coeficiente de correlação entre amostras adjacentes. No caso Gaussiano, necessita-se obter a inversa de Σ, que no caso sendo analisado é fácil de ser obtida:

$$\Sigma^{-1} = \frac{1}{1 - \rho^2} \begin{bmatrix} 1 & -\rho & 0 & \ldots\ldots & 0 \\ -\rho & 1+\rho^2 & -\rho & \ldots\ldots & 0 \\ 0 & -\rho & & \ldots\ldots & 0 \\ & & & \ldots\ldots & \\ 0 & 0 & \ldots & 1+\rho^2 & -\rho \\ 0 & 0 & & -\rho & 1 \end{bmatrix}$$

$$| \Sigma | = (1 - \rho^2)^{d-1}$$

Deve-se observar que uma função de autocovariância ρ^n é obtida para um processo aleatório resultante da passagem de ruído branco através de um filtro digital passa-baixas (ou analógico em que a frequência de corte é bem menor que T^{-1}) de 1ª ordem.

Será que sempre a função custo simétrica é a mais apropriada ? Em termos de modelar a realidade não. Por exemplo, tomemos 2 classes e um exame laboratorial. A classe 1 é de pacientes com cardiopatia e a classe 2 é de pacientes sem doença cardíaca. Aqui claramente não convém se associar custos iguais de classificação incorreta $C(\omega_1|\omega_2)$ e $C(\omega_2|\omega_1)$ pois é menos danoso classificar erradamente o paciente na classe 1, pois neste caso ele irá passar por outros exames, do que na classe 2 pois neste caso ele poderá, p.ex., vir a sofrer um enfarte. No classificador de Bayes é necessário conhecer a distribuição probabilística a priori de cada classe bem como as densidades condicionadas a cada classe, além de se estipular a função custo. Outras duas abordagens - a minimax e a de Neyman-Pearson - exigem menos conhecimentos a priori (vide por exemplo, Tou e Gonzalez, 1974 pg. 118). É interessante mencionar que, para 2 classes, os 3 classificadores fornecem como regra de decisão a razão de verossimilhança, mudando apenas o limiar.

ESTIMAÇÃO NÃO PARAMÉTRICA DE FUNÇÃO DENSIDADE DE PROBABILIDADE

INTRODUÇÃO

Na abordagem por teoria de decisão de Bayes para o problema da classificação, é necessário se conhecer $P(\omega_i)$ e $p(\underline{x}|\omega_i)$; i=1, ...c. Para a estimação tanto de $p(\underline{x}|\omega_i)$ quanto de $P(\omega_i)$, deve-se dispor de um conjunto de padrões em que a classificação de cada padrão é conhecida. Tomam-se então os padrões da classe ω_i e com estes estimam-se $P(\omega_i)$ e $p(\underline{x}|\omega_i)$. Em certos casos, pode-se dispor de conhecimentos sobre a forma das distribuições envolvidas, quer porque se conhecem bem os mecanismos que geram os padrões ou porque se realizaram testes estatísticos não paramétricos (χ^2, Kolmogorov-Smirnov, ou outros). Nestes casos fica faltando apenas estimar os parâmetros da distribuição, e isto pode ser feito utilizando técnicas clássicas da teoria de estimação (p.ex., Hoel, Port e Stone, 1971; Roussas, 1973; Mood, Graybill e Boes, 1974; Bickel e Docksum, 1977).

Nos casos em que não há informação sobre a distribuição, pode ser feita uma estimação diretamente a partir dos dados, sendo este tipo de estimação denominado de estimação não paramétrica ou independente de distribuição ("distribution free"). $P(\omega_i)$ é de estimação simples ao passo que $p(\underline{x}|\omega_i)$ requer considerações especiais, conforme apresentaremos a seguir.

O estimador mais elementar de função densidade de probabilidade é o histograma, mas este apresenta inconvenientes como:

(*) sensibilidade à escolha da origem de coordenadas para colocar a "malha ou grade" ("bins") que discretizam a variável ou variáveis, conforme ilustrado nas Figs. 1a e 1b, em que foram geradas 70 amostras de N(0,1) e mais 70 amostras de N(-1,1). A diferença entre as duas figuras é que na Fig. 1b as amostras mínima e máxima sofreram um acréscimo de 0.1 em seu valor, o que equivale a deslocar a grade, uma vez que a rotina utilizada (hist do pacote Matlab) coloca os extremos da grade nos valores extremos dos dados.

(*) dificuldade na escolha do fator de alisamento, dado pela largura da malha ("bin-width"), embora todos os métodos apresentem esse problema .

(*) dificuldade na escolha da direção da malha no caso multi-dimensional.

(*) derivada infinita (descontinuidades) em todas as regiões de transição na malha.

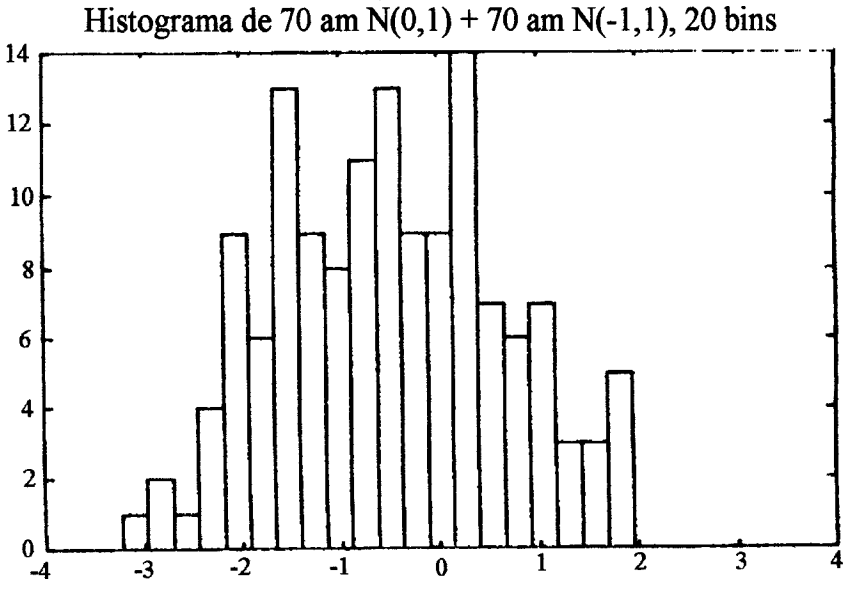
(*) número excessivo de cálculos (M^d subdivisões ou "bins" no espaço de d variáveis, onde M é o número de subdivisões em cada uma das d dimensões; por exemplo, para M=10 e d=6 temos 1 milhão de subdivisões).

(*) número excessivo de amostras para se obter uma estimativa de razoável qualidade.

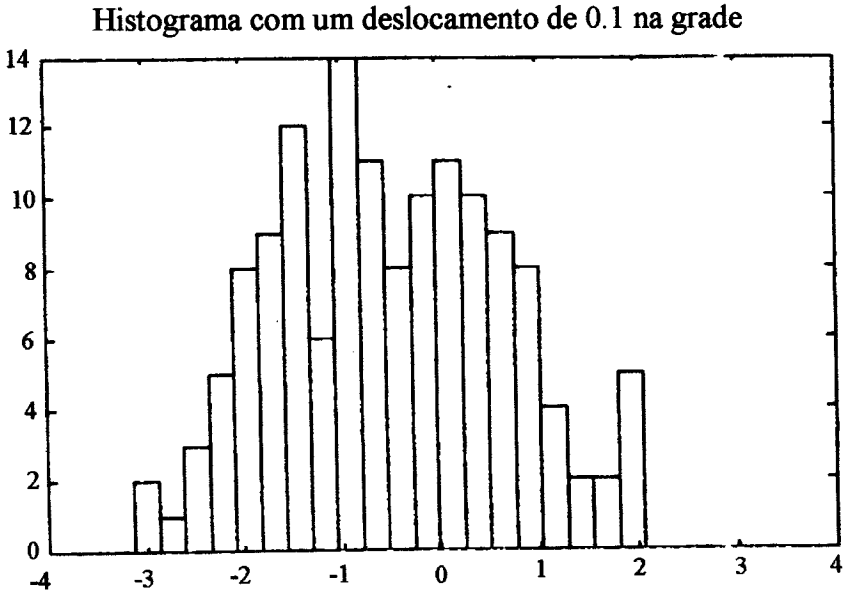
Dadas as inconveniências do histograma como estimador da função densidade de probabilidade, foram criados métodos melhores de estimação como, por exemplo, o da função peso ou "kernel" e o dos vizinhos mais próximos.

MÉTODO DA FUNÇÃO PESO

A maior parte de nossa análise será para o caso unidimensional para que tenhamos maior simplicidade de apresentação.



(a)



(b)

Fig. 1- Dois histogramas obtidos com 20 bins a partir das mesmas 140 amostras obtidas da mistura (70 amostras de cada) de uma distribuição N(0,1) com N(-1,1). A única diferença de (a) para (b) foi um deslocamento de 0,1 unidades na abscissa do histograma.

Sejam X_1 , X_2 , ... , X_N variáveis aleatórias i.i.d. (independentes e identicamente distribuidas) com função de distribuição F(x) e função densidade de probabilidade p(x). Temos portanto

$$F(x) = \text{Prob}(X \leq x) = \int_{-\infty}^{x} p(a)da \qquad (1)$$

Um estimador para F(x) pode ser obtido utilizando-se o conceito de frequência relativa aplicado sobre o conjunto de amostras x_1 , x_2 ,..., x_N :

$$F_N(x) = \{ \text{número de amostras com valor} \leq x \}/N \qquad (2)$$

onde deve-se ressaltar que $x \in \mathbb{R}$ ou a um subconjunto de $\mathbb{R}$. Duas propriedades normalmente apresentadas por bons estimadores são: vício nulo e variância tendendo assintoticamente a zero para $N \to \infty$). Iremos verificar se o estimador proposto satisfaz a estas propriedades.

Dispomos de N amostras e para o estimador dado em (2), temos que $N \cdot F_N(x)$, para um valor fixo para x, é o número de sucessos de "certo tipo" (valor $\leq$ x). Portanto, $N \cdot F_N(x)$ é uma variável aleatória binomial, pois cada experimento é do tipo Bernoulli e independente dos demais. Temos então que se

$$P(X_i \leq x) = p$$

então

$$E\,[F_N(x)] = (1/N) \cdot E\,[\text{número de } X_i \leq x\,] = (1/N) \cdot N \cdot p \qquad (3)$$

onde $N \cdot p$ é a média da Binomial(N,p). De (3) segue que

$$E\,[F_N(x)] = P(X \leq x) = F(x) \qquad (4)$$

e portanto conclui-se que o estimador $F_N(x)$ não é viciado. Podemos facilmente obter a variância do estimador:

$$\text{var}\,[F_N(x)] = (1/N^2) \cdot N \cdot p \cdot (1-p) = (1/N) \cdot F(x) \cdot (1-F(x)) \qquad (5)$$

de onde se conclui que $F_N(x)$ é consistente em média quadrática. Vale relembrar que um estimador θ_N de θ é consistente em média quadrática se $E\,[(\theta_N - \theta)^2] \longrightarrow 0$ para $N \longrightarrow \infty$. Essa consistência em média quadrática implica que (e vice-versa) tanto o vício quanto a variância do estimador tendem a zero para N tendendo a infinito, isto vindo do fato que

$$E\,[(\theta_N - \theta)^2] = \text{var}\ (\theta_N) + (\theta - E\,[\theta_N])^2$$

Pode-se ter a falsa impressão de que será fácil obter um bom estimador também para a função densidade de probabilidade p(x) uma vez que esta depende da função de distribuição. Tentaremos construir um estimador da forma mais direta possível a partir de F(x), lembrando que

$$p(x) = \frac{d}{dx}\,\text{Prob}(X \leq x) = \frac{dF(x)}{dx} = \lim_{h \to 0} \frac{F(x+h) - F(x-h)}{2h}$$

o que sugere usarmos o estimador

$$p_N(x) = [F_N(x+h) - F_N(x-h)]/2h \qquad (6)$$

Lembrando que $E\,[F_N(x+h)] = F(x+h)$ segue que

$$E\,[p_N(x)] = (1/2h)\cdot\int_{x-h}^{x+h} p(a)\,da \qquad (7)$$

Da expressão (7) percebe-se que, em geral, nem para N tendendo a infinito temos convergência do vício para zero.

O estimador em (6) pode ser escrito diretamente em função das amostras

$x_1, x_2, \ldots, x_N$

$$p_N(x) = (1/2h)\cdot\left(\frac{\text{número de amostras com valor} \leq x+h}{N} - \frac{\text{número de amostras com valor} \leq x-h}{N}\right)$$

ou ainda

$$p_N(x) = (1/2h)\cdot\left(\frac{\text{número de amostras com valor} \in (x-h, x+h]}{N}\right) \qquad (8)$$

onde deve-se ressaltar que x toma valores em $\mathbb{R}$ ou um seu subconjunto. Esta é uma diferença fundamental entre esse estimador e o estimador por histograma que faz uma quantização arbitrária da variável x (ou das variáveis, no caso de dimensão maior que 1). Para determinar o vício do estimador $p_N(x)$, utilizaremos a expressão (7). Antes, obteremos uma expansão em série de Taylor para p(a) em torno de a=x, truncando-a em seguida:

$$p(a) \cong p(x) + (a-x)\cdot dp(t)/dt|_{t=x} + 0.5\cdot(a-x)^2\cdot d^2p(t)/dt^2|_{t=x} \qquad (9)$$

Como a fórmula (7) requer uma integral na variável a, esta é calculada abaixo para cada um dos termos da expressão (9)

$$\int_{x-h}^{x+h} (a-x)da = 0$$

$$\int_{x-h}^{x+h} 0.5\cdot(a-x)^2 da = h^3/3$$

de onde se conclui que

$$E\,[p_N(x)] \cong p(x) + \frac{1}{2h}\cdot\frac{h^3}{3}p''(x) \qquad (10)$$

e portanto o vício é aproximadamente $\frac{h^2}{6}\cdot p''(x)$. Supondo h fixo, vemos que o vício não tende a zero quando N tende a infinito, ou seja, novamente concluímos que o estimador será viciado, em geral, para qualquer valor de N. Para diminuir o vício basta utilizar valores de h menores.

Forneceremos a seguir uma expressão aproximada para a variância, que foi obtida na suposição que i h é suficientemente pequeno para podermos desprezar o vício que é de ordem h^2 e ii $p(x) >> 2\cdot h\cdot p^2(x)$

$$\text{var}\,[p_N(x)] \cong \frac{p(x)}{2Nh}$$

Deste resultado, concluimos que i para diminuir a variância devemos aumentar o valor de h, o que infelizmente piora o vício e portanto há que se chegar a um compromisso entre vício e variância baixos, ii se fixarmos o valor de h e

fizermos N tender a infinito, temos a variância tendendo a zero. Mas como o estimador é viciado, não se tem a consistência (em média quadrática) do estimador. Como o vício é proporcional a h^2, se tivermos h tendendo a zero quando N tende a infinito, então o estimador $p_N(x)$ é assintoticamente não viciado. Se, além do mais, lim Nh = ∞ para N → ∞ então

$$\lim_{N\to\infty} \quad E\{[p_N(x) - p(x)]^2\} = 0$$

e nestas condições o estimador tem boas propriedades assintóticas.

Além do erro médio quadrático, outra medida de qualidade para estimadores de função densidade de probabilidade é a média da integral do erro quadrático

$$E\left[\int [p_N(x) - p(x)]^2 dx\right]$$

que é uma medida global de qualidade uma vez que resulta um número e não uma função de x como no caso do erro médio quadrático. Essa medida de qualidade pode ser utilizada por exemplo para tentar responder a uma pergunta básica: que valor de h empregar na prática? No caso particular de uma distribuição normal com variância σ^2, um valor ótimo para h é (Silverman, 1986)

$$h_o = 1.06 \cdot \sigma \cdot (N)^{-1/5}$$

Para casos mais genéricos, principalmente quando não se tem informação sobre o tipo de função densidade de probabilidade que rege as amostras disponíveis, não há uma solução em forma fechada uma vez que o valor ótimo para h fica dependendo da integral da segunda derivada ao quadrado da função densidade desconhecida. Silverman (1986) apresenta uma discussão de algumas abordagens possíveis para a escolha de h em aplicações práticas.

Uma generalização do que foi exposto acima pode ser obtida escrevendo-se $p_N(x)$ como (vide expressão (8))

$$p_N(x) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h} \right) \qquad (11)$$

onde a função K(x) é uma função peso ("kernel") satisfazendo às seguintes propriedades:

(i) $0 \leq K(x) \leq \infty$

(ii) $K(x) = K(-x)$

(iii) $\int_{-\infty}^{\infty} K(x)dx = 1$

(iv) $\lim_{|x| \to \infty} |x \cdot K(x)| = 0$

O caso já analisado do estimador dado por (8) estaria associado à função peso

$$K(x) = \begin{cases} 1/2 & \text{se } x \in [-1,1) \\ 0 & \text{caso contrario} \end{cases}$$

que é uma função peso descontínua (função peso retangular). Um exemplo de função peso contínua é a função densidade de probabilidade normal de média 0 e variância 1, indicada por N(0,1). Deve-se observar que a propriedade (iii) impõe que a função peso seja uma densidade de probabilidade. A função peso retangular fornece estimativas que são descontínuas (e portanto não diferenciáveis) o que em algumas aplicações pode ser inconveniente. Isto não acontece com funções peso contínuas como no caso da normal. Expressões semelhantes para vício e variância podem ser obtidas para o estimador mais geral dado em (11) (Silverman, 1986)

$$\text{vício} \cong 0.5 \cdot h^2 \cdot p''(x) \cdot \int t^2 K(t)\, dt$$

$$\text{variância} \cong \frac{p(x)}{N \cdot h} \int K^2(t)dt$$

onde novamente o compromisso entre baixo vício e baixa variância deve ser notado. Com a generalização que permite usar funções peso diferentes da

retangular (que é a mais intuitiva), pode-se obter estimadores mais ajustados para determinada aplicação prática.

Silverman (1982, 1986) apresenta uma técnica para cálculo eficiente de histogramas para uma única variável. Ela se baseia no fato que o estimador por função peso nada mais é do que a convolução da função peso com um trem de impulsos de Dirac localizados nas amostras (isto é, tomando o eixo de abscissas como sendo x, há um impulso $\delta(x-x_i)$ em cada amostra x_i). Como operacionalmente é melhor trabalhar no domínio da frequência, onde a convolução se transforma em um produto, pode-se tirar proveito da rapidez dos algoritmos de FFT para fazer um cálculo eficiente de estimadores de função densidade de probabilidade. Se selecionarmos uma função peso normal, sua transformada de Fourier também é normal e portanto bastaria fazer o produto no domínio da frequência por essa função. Ao invés desta, pode-se utilizar outros janelas espectrais. Na Fig. 2, mostramos, através de um exemplo, o uso de uma janela espectral triangular adotada apenas por simplicidade. Inicialmente, a Fig. 2a apresenta o histograma, em que os valores de contagem adjacentes foram unidos por meio de segmentos de reta. Nesta figura não se utilizou nenhuma janela espectral e portanto não há alisamento do histograma, como pode-se notar. Em linha contínua mostramos a função densidade teórica, que é uma normal N(0,1). Foram geradas 200 amostras e utilizada uma subdivisão do eixo de abscissas igual a 64. Na Fig. 2b se fez uma multiplicação no domínio de frequência por uma janela triangular, centrada na frequência zero e chegando ao valor zero na abscissa igual a 32 (metade do número de amostras usadas para o cálculo da FFT), isto correspondendo a um filtro passa-baixas relativamente suave. Já se nota uma melhoria, embora

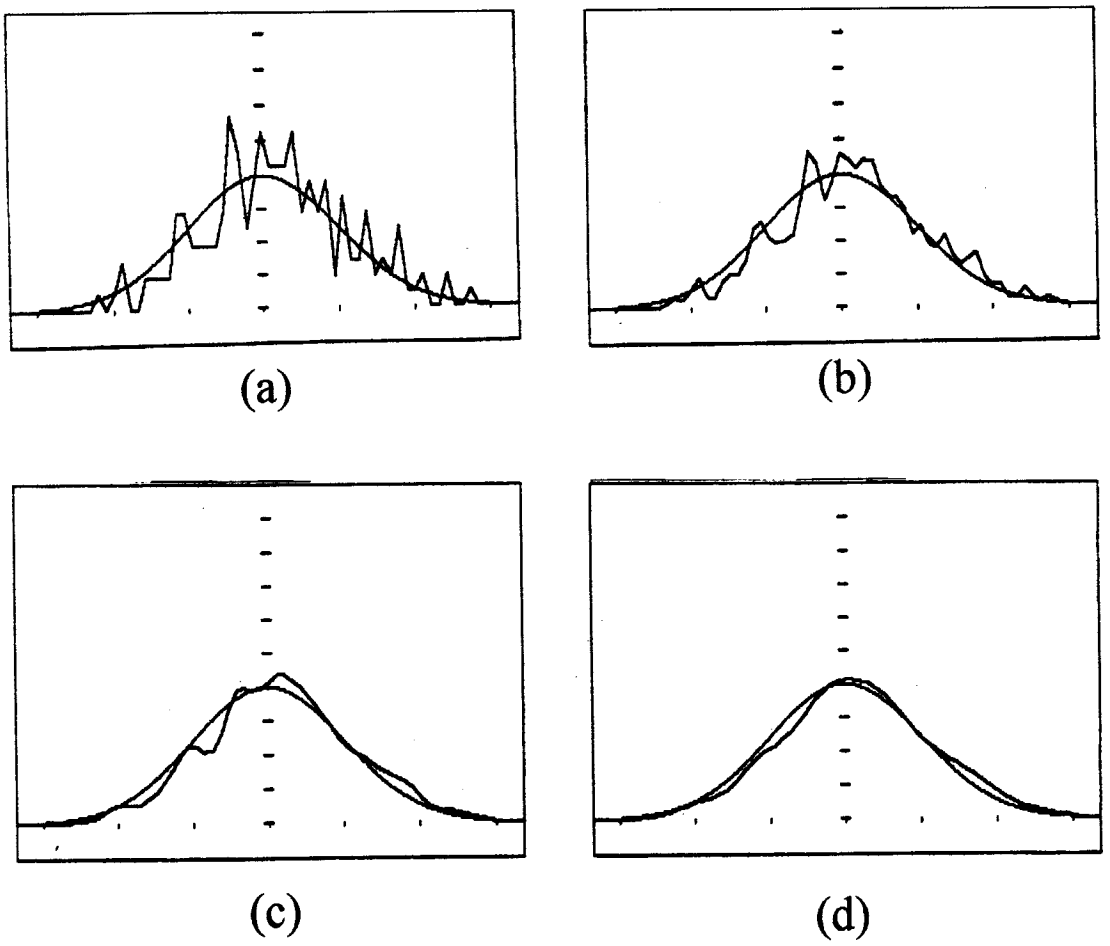


Fig. 2 – Estimação de função densidade, para uma única variável, utilizando o método da função peso. O controle do alisamento é feito no domínio da frequência, conforme explicado no texto.

ainda haja muita variabilidade. Nas Figs.2c e 2d a janela triangular atinge o valor zero para abscissas iguais a 17 e 10, respectivamente, ou seja, frequências altas serão zeradas ou atenuadas em maior grau. Nota-se que as estimativas da função densidade resultam extremamente melhores.

Para o caso multidimensional define-se uma função peso que é agora função de um vetor $\underline{x}$. Impõe-se novamente que a função peso seja uma função densidade de probabilidade (integral em $\mathbb{R}^d$ igual a 1). O estimador através de função peso fica

$$p_N(\underline{x}) = \frac{1}{N \cdot h^d} \sum_{i=1}^{N} K\,[(\underline{x} - \underline{x}_i)/h] \qquad (12)$$

onde os $\underline{x}_i$ são as amostras vetoriais disponíveis. Um exemplo de função peso é a função densidade normal multivariada com vetor médio nulo e matriz de covariância identidade. A simetria radial da função peso que ocorre nesse exemplo é geralmente empregada também em outras funções peso. Caso os dados tenham direções privilegiadas, pode-se fazer uma transformação para descorrelacionar os mesmos utilizando a matriz de covariância amostral (vide capítulos seguintes) e então a estimação de função densidade com função peso simétrica dará melhores resultados. Através de derivações semelhantes às feitas para o caso unidimensional, pode-se (Silverman, 1986) obter expressões para o vício e a variância do estimador dado por (12)

$$\text{vício} \cong 0.5 \cdot h^2 \cdot \alpha \cdot \text{tr}[H(p(\underline{x}))]$$

$$\text{variância} \cong (N)^{-1} \cdot (h)^{-d} \cdot \beta \cdot p(\underline{x})$$

onde

$\alpha = \int t_1^2 \cdot K(\underline{t}) d\underline{t}$, com $\underline{t} = [t_1 \;\; t_2 \;\; \ldots \;\; t_d]^T$,

$\beta = \int K^2(\underline{t}) d\underline{t}$ e

$H[p(\underline{x})]$ é o Hessiano da função $p(\underline{x})$;devendo-se ressaltar que estes resultados valem para o caso em que a função peso tem simetria radial (simetria par em qualquer variável, independente do valor fixado para as demais variáveis). Da análise das expressões para o vício e a variância chega-se a conclusões semelhantes ao caso escalar, ou seja, o vício não tende a zero quando o número de amostras tende a zero, embora a variância tenda. Quando se diminui h o vício se reduz mas a variância aumenta. Novamente não há uma expressão útil para um valor ótimo para h pois este fica em função de derivadas da própria função densidade desconhecida.

Podem-se ressaltar alguns aspectos sobre a estimação da função densidade de probabilidade no caso multidimensional:

a) o tempo de computação para se achar uma estimativa para a função densidade pode ser extremamente grande, principalmente no caso de se utilizar uma função peso complicada.

b) mesmo para dimensionalidades não muito grandes já temos a nessecidade de dispor de um número enorme de amostras para que a qualidade da estimativa seja boa. Para um exemplo normal em que se deseja para a estimativa na origem um erro relativo médio quadrático $(E[p_N(\underline{0})-p(\underline{0})]^2/p^2(\underline{0}))$ menor que 0,1, Silverman (1986) fornece os valores 223, 43.700 e 842.000 para o número desejado de amostras para dimensão 4, 8 e 10, respectivamente.

c) em dimensões maiores há certos resultados que podem ser contra-intuitivos para os que tentam extrapolar conhecimentos que se aplicam para uma ou poucas dimensões. Por exemplo, para uma distribuição normal sobre um espaço de dimensão 8, centrada na origem, com variáveis não correlacionadas e variância unitária, conclui-se que a probabilidade de amostras ocorrerem

em um hipercubo $\{\underline{x} \mid |x_i| \leq 1, \forall\ i\}$ é aproximadamente 0,047, o que parece contra-intuitivo pois no caso unidimensional a probabilidade de se ter amostras no intervalo -1,1 é de aproximadamente 0,6826. Isto significa que, no caso multidimensional, pode-se ter muito poucas amostras em regiões em que a função densidade tem valor grande. Não se pode, pela mesma razão, achar que para valores baixos de função densidade não se terá amostras.

A técnica de estimação não paramétrica por função peso (ou "Kernel" em inglês) é por vezes denominada de técnica de estimação por janelas de Parzen.

MÉTODO DOS VIZINHOS MAIS PRÓXIMOS

O método de estimação básico por função peso acaba sendo ineficiente pois utiliza a mesma função peso quer haja uma grande ou pequena densidade de amostras em diferentes regiões. Isto é facilmente sanado pela abordagem do estimador por (ou de) vizinhos mais próximos abreviado comoEVMP (em inglês é o "nearest neighbor estimator" com a abreviação NN).

Dada a função densidade de probabilidade $p(\underline{x})$, a probabilidade de que uma amostra vai ocorrer em uma vizinhança L de $\underline{x}$ é

$$\theta = \int_L p(\underline{\alpha})d\underline{\alpha}$$

Se L for uma região pequena, com volume V , então

$$\theta \approx p(\underline{x}).V \qquad \text{e portanto}$$

$$p(\underline{x}) \approx \theta/V$$

θ/V é uma versão alisada de $p(\underline{x})$, sendo igual ao valor médio de $p(\underline{\alpha})$ na

região L em torno de $\underline{\alpha} = \underline{x}$. Pode-se estimar θ a partir das amostras, calculando-se a proporção k_N/N de amostras que caíram em L , onde k_N é o número de amostras em L, e N é o número total de amostras. Desta forma temos o estimador

$$p_N(\underline{x}) = \frac{k_N}{NV_N}$$

onde os índices N em k e V indicam dependência com N.

Esse estimador é equivalente a um estimador por função peso, em que a função peso é $K(\underline{x}) = 1/v$ se $\|\underline{x}\|_2 \leq r$ e 0 em caso contrário, onde v é o volume de um hiperesfera de raio r, e $\|\underline{x}\|_2$ é a norma Euclidiana $(x_1^2+x_2^2+\ldots+x_d^2)^{1/2}$.

Se fixarmos um valor constante para V_N obtemos um estimador que apresenta a desvantagem de ter a largura h da função peso equivalente fixa em todo o espaço. Neste caso, em regiões com baixo número de amostras obtém-se uma estimativa com muita variabilidade, apresentando picos estreitos se alternando com regiões planas de valor nulo, ao passo que em regiões com maior densidade de amostras pode ser que as variações intrínsecas da função densidade de probabilidade sejam alisadas.

A abordagem mais interessante fixa k_N e determina o volume V necessário para englobar as k_N amostras mais próximas do ponto $\underline{x}$ em estudo. Devido ao uso de k amostras vizinhas, o método tem o nome k-NN. Prova-se que, se e só se: $\lim_{N\to\infty} k_N = \infty$ e $\lim_{N\to\infty} \frac{k_N}{N} = \infty$ então

$$p_N(\underline{x}) \xrightarrow[n\to\infty]{\text{em prob.}} p(\underline{x})$$

em todos os pontos de continuidade de $p(\underline{x})$, onde o limite em probabilidade

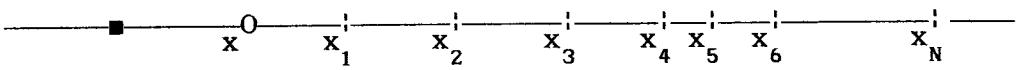
significa:

$$\text{Prob}\left[\left|p_N(\underline{x})-p(\underline{x})\right| > \varepsilon\right] \longrightarrow 0 \quad \text{para } N \rightarrow \infty, \quad \forall\, \varepsilon$$

Desta forma, em regiões de alta densidade de amostras as células terão volume pequeno resultando em alta resolução. Caso se utilize distância Euclideana, cada volume V será uma hiper-esfera com certo raio r. Caso se utilize a distância "sup" d_∞ ($d_\infty(\underline{x},\underline{y}) = \max_{1 \le i < d} |x_i - y_i|$), temos um hiper- cubo para o volume V. Para se ter vício nulo e variâcia não extremamente grande deve-se usar o estimador $Pn(\underline{x})=(k_n-1)/NV_n$ com $k_n>2$ (Fugunaga,1990).

O estimador de função densidade de probabilidade EVMP tem decaimento muito lento nas caudas, o que faz com que a sua integral seja infinita e portanto não seja um estimador de função densidade de probabilidade de verdade. Mostremos este fato da integral divergir para o caso unidimensional:

Suponhamos as amostras $x_1\ x_2\ \ldots\ x_N$ já ordenadas conforme esquematizado abaixo



Para $x < x_1$ tem-se largura de célula ("bin width") $2(x_k - x)$ e portanto nesta região

$$\int_{-\infty}^{x_1} \frac{k_N}{N\ 2(x_k-x)}\,dx = \int_{x_k-x_1}^{\infty} \frac{k_N}{2\ Nu}\,du = \frac{k_N}{2N}\ln u \Big|_{x_k-x_1}^{\infty} = +\infty$$

Na outra cauda o mesmo vai ocorrer. O decaimento em ambas as caudas segue a lei 1/x, que é muito lento. Uma outra propriedade inadequada do EVMP é que

ele não é suave, pois apresenta descontinuidades em sua derivada. No caso de dimensão igual a 1, as descontinuidades são devidas a picos finos formados pela confluência de 2 retas (vide Fig. 2.10 pg. 20 de Silverman, 1986). Mas este é um problema menor do que o ocorre no histograma tradicional, onde a derivada é infinita em vários pontos. O problema da integral divergente, devido a caudas decaindo com $||\underline{x}||^{-d}$ no caso multidimensional, pode ou não causar dificuldades práticas. Caso se utilize o estimador só na região fora das caudas ("cauda" significando região em que não há amostras e portanto apenas se extrapola a função densidade), poder-se-á obter um desempenho adequado para o classificador de Bayes.

O teorema mencionado da convergência do EVMP nos faz perguntar como escolher k_N. Uma escolha popular é $k_N = \sqrt{N}$ e se supusermos que $p_N(\underline{x})$ é uma razoável estimativa de $p(\underline{x})$ teremos

$$V_N \approx 1/(\sqrt{N}\ p(\underline{x})).$$

As idéias apresentadas nesse método de estimação de função densidade de probabilidade são muito úteis também para uma abordagem direta da tarefa de classificação, levando à regra de decisão de vizinhos próximos, apresentada em um capítulo seguinte.

FUNÇÕES DE DECISÃO

INTRODUÇÃO

Uma abordagem importante em classificação de padrões consiste na utilização de funções de decisão ou funções discriminantes. Estas funções são mapeamentos $\mathbb{R}^d \rightarrow \mathbb{R}$ onde d é a dimensão do espaço atributos. As regras de decisão são então baseadas nos valores associados a cada função de decisão. As funções de decisão lineares, devido à sua simplicidade, são bastante populares, além de serem ótimas no caso de distribuições normais com mesma matriz de covariância.

FUNÇÕES DE DECISÃO LINEARES

Uma função de decisão linear, ou função discriminante linear, tem a forma geral

$$d(\underline{x}) = \underline{v}_o^T \cdot \underline{x} + v_{d+1} = v_1 x_1 + v_2 x_2 + \ldots + v_d x_d + v_{d+1} \qquad (1)$$

mas, para facilitar a notação, definem-se os vetores de padrões (ou de atributos) $\underline{x}_e$ e peso $\underline{v}$ aumentados:

$$\underline{x}_e = [\, x_1 \; x_2 \; \ldots \; x_d \; 1 \,]^T \qquad \text{e}$$

$$\underline{v} = [\, v_1 \; v_2 \; \ldots \; v_d \; v_{d+1} \,]^T \qquad \text{e com isto}$$

$$d(\underline{x}) = \underline{v}^T \cdot \underline{x}_e \qquad (2)$$

Note que $d(\underline{x})$ é o produto escalar de $\underline{v}$ com $\underline{x}_e$.

No caso de duas classes, a função de decisão é suposta ter a propriedade:

$$d(\underline{x}) = \underline{v}^T \cdot \underline{x}_e \quad \begin{cases} > 0 & \text{se } \underline{x}_e \in \omega_1 \\ < 0 & \text{se } \underline{x}_e \in \omega_2 \end{cases}$$

onde utilizamos um abuso de notação ao escrever $\underline{x}_e \in \omega_1$ (ou ω_2), mas em termos de interpretação acreditamos não haver qualquer dúvida.

No caso de c classes temos os seguintes casos:

Caso 1 Cada classe é separável das outras classes por uma única superfície de decisão. Há c funções de decisão tais que

$$d_i(\underline{x}) = \underline{v}_i^T \cdot \underline{x}_e \quad \begin{cases} > 0 & \text{se } \underline{x}_e \in \omega_i \\ < 0 & \text{se } \underline{x}_e \notin \omega_i \end{cases} \qquad i = 1, 2, \ldots, c$$

onde $\underline{v}_i = [v_{i1} \; v_{i2} \ldots . v_{id+1}]^T$ é o vetor peso associado à i-ésima função de decisão. A superfície de decisão ou fronteira para a classe ω_i é definida por $d_i(\underline{x}) = 0$. No exemplo da Fig. 1 , as funções de decisão são:

$$d_1(\underline{x}) = -x_1 + x_2 - 1 \;, \quad d_2(\underline{x}) = x_1 + x_2 - 4 \quad \text{e} \quad d_3(\underline{x}) = -x_1 - 4x_2 + 2$$

Deve-se notar que a função $d_3(\underline{x})$ é $-x_1-4x_2+2$ e não x_1+4x_2-2 (como talvez fizéssemos instintivamente se apenas olhássemos para a equação da reta que define a fronteira $d_3(\underline{x}) = 0$) uma vez que deve-se também atentar para o lado positivo imposto à função (onde está Ω_3 em relação à fronteira).

Outro aspecto a observar é que, no Caso 1, a partição do espaço de atributos pelas funções de decisão gera dicotomias do tipo ω_i/não ω_i, bem como regiões com classificação indeterminada (RCI). Obviamente, qualquer amostra que porventura ocorra sobre uma fronteira também terá uma classificação indeterminada. As regiões de classificação válida são todas convexas e conexas.

⊙⊙⊙⊙⊙ Exemplo: Dadas as funções de decisão fornecidas acima, classificar o padrão $\underline{x}_o = [\; 0{,}5 \quad 3 \;]^T$.

Temos $\left.\begin{array}{l} d_1(\underline{x}_o) = \quad 1{,}5 > 0 \\ d_2(\underline{x}_o) = -0{,}5 < 0 \\ d_3(\underline{x}_o) = -10{,}5 < 0 \end{array}\right\}$ e como $d_1(\underline{x}_o) > 0$ tem-se $\underline{x}_o \in \omega_1$.

⊙⊙⊙⊙⊙

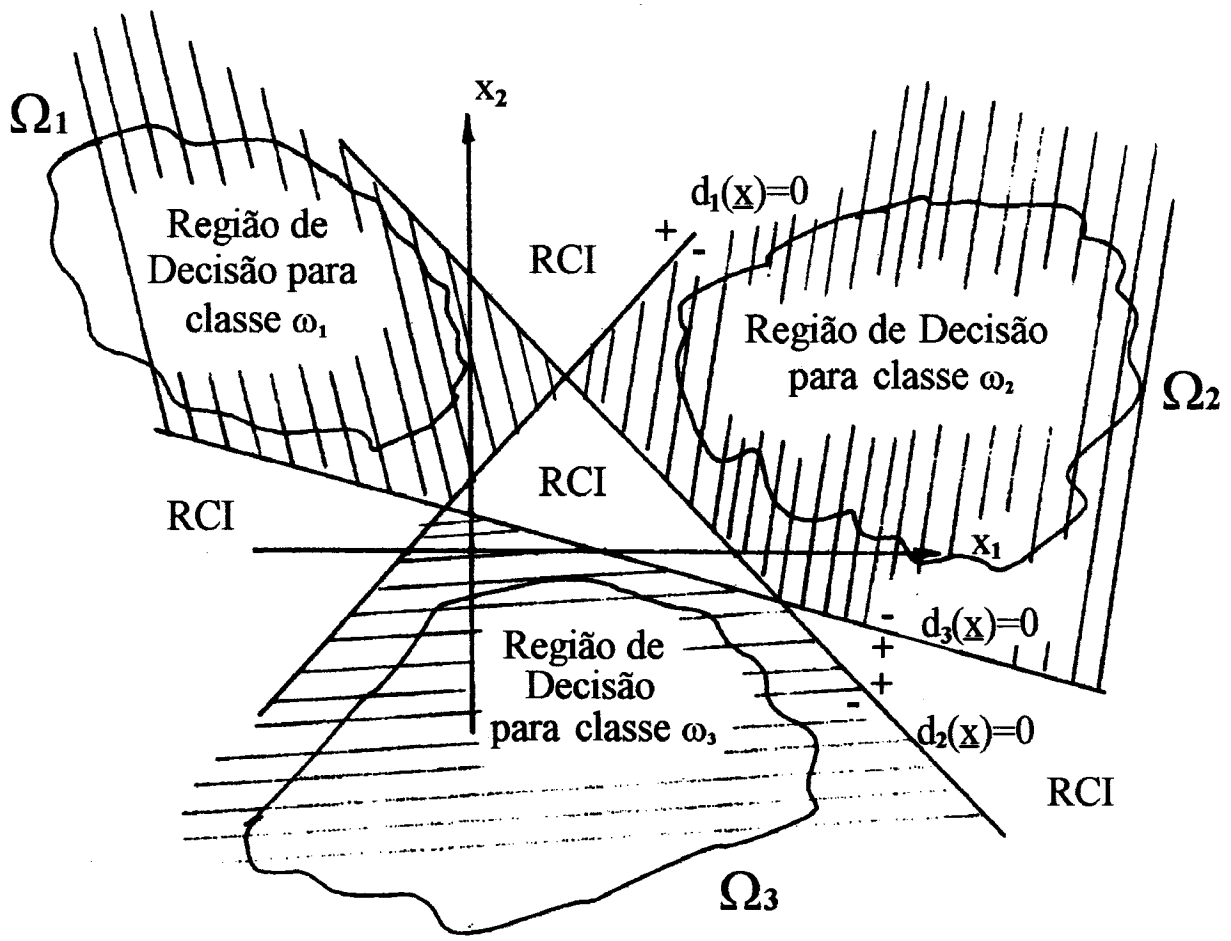


Fig. 1 – Fronteiras e regiões de decisão para um exemplo no Caso 1. Há várias regiões de classificação indeterminada (RCI).

Caso 2 Cada classe é separável de cada outra classe por uma superfície de decisão distinta. As classes são separáveis aos pares e portanto há c(c-1)/2 = $\binom{c}{2}$ superfícies de decisão. As funções de decisão são especificadas como $d_{ij}(\underline{x}) = \underline{v}_{ij}^T \cdot \underline{x}_e$ com $d_{ji}(\underline{x}) = -d_{ij}(\underline{x})$. Se $\underline{x} \in \omega_i$ então

$$d_{ij}(\underline{x}) > 0 \qquad \forall \; j \neq i$$

No exemplo da Fig. 2 as funções de decisão são:

$$d_{12}(\underline{x}) = -x_1 - x_2 + 5, \quad d_{13}(\underline{x}) = -x_1 + 3, \quad d_{23}(\underline{x}) = -x_1 + x_2,$$

$$d_{21}(\underline{x}) = -d_{12}(\underline{x}) \; , \quad d_{31}(\underline{x}) = -d_{13}(\underline{x}) \; , \quad d_{32}(\underline{x}) = -d_{23}(\underline{x}).$$

Neste exemplo teremos as decisões

$$\omega_1 \quad \text{se} \quad d_{12}(\underline{x}) > 0 \quad \text{e} \quad d_{13}(\underline{x}) > 0$$

$$\omega_2 \quad \text{se} \quad d_{21}(\underline{x}) > 0 \quad \text{e} \quad d_{23}(\underline{x}) > 0$$

$$\omega_3 \quad \text{se} \quad d_{31}(\underline{x}) > 0 \quad \text{e} \quad d_{32}(\underline{x}) > 0$$

No Caso 2 também há a possibilidade de regiões com classificação indeterminada (RCI), além das próprias fronteiras. No exemplo da Fig. 2 ela é definida por $d_{13}(\underline{x}) > 0$, $d_{21}(\underline{x}) > 0$ e $d_{32}(\underline{x}) > 0$. As dicotomias são do tipo ω_i/ω_j. Cada região de classificação válida é convexa e conexa.

⊙⊙⊙⊙⊙ Exemplo: Dadas as funções de decisão acima, classificar o vetor $\underline{x}_0 = [\ 4 \quad 5\ ]^T$. Deve-se calcular os valores das funções de decisão em $\underline{x}_o$:

$$d_{12}(\underline{x}_0) = -4 \qquad d_{21}(\underline{x}_0) = 4$$

$$d_{13}(\underline{x}_0) = -1 \qquad d_{31}(\underline{x}_0) = 1$$

$$d_{23}(\underline{x}_0) = 1 \qquad d_{32}(\underline{x}_0) = -1$$

Como d_{21} e d_{23} são positivos então conclui-se que $\underline{x}_0$ deve ser atribuído à classe ω_2 .

⊙⊙⊙⊙⊙

Casos práticos podem envolver uma mistura dos Casos 1 e 2.

Caso 3 Há c funções de decisão $d_j(\underline{x}) = \underline{v}_j^T \cdot \underline{x}_e$, j = 1, ...c, tais que se $\underline{x} \in \omega_i$

$$d_i(\underline{x}) > d_j(\underline{x}) \qquad \forall \quad j \neq i$$

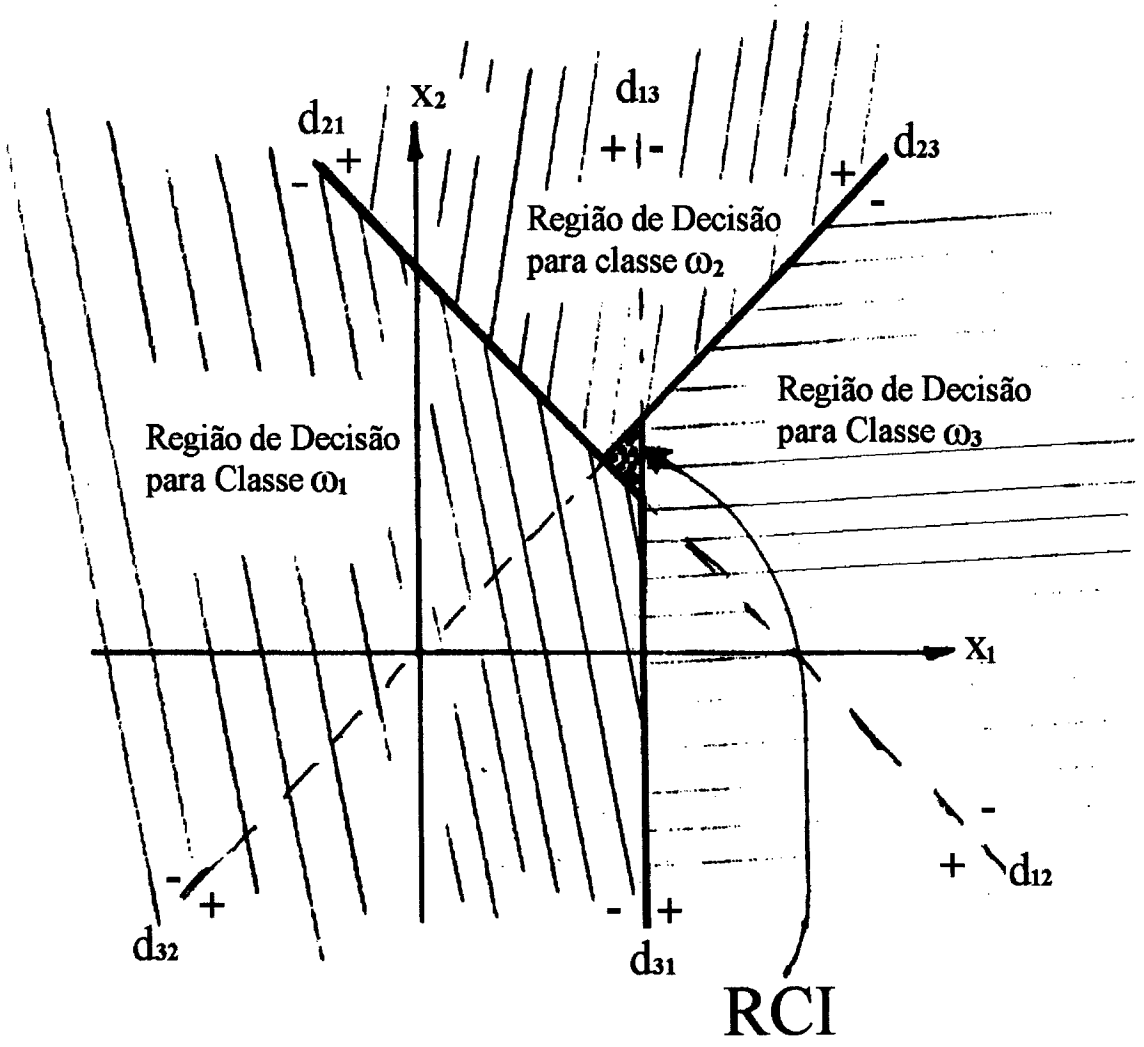


Fig. 2 – Fronteiras e regiões de decisão para um exemplo no Caso 2. Neste exemplo há uma região com classificação indeterminada (RCI).

Este foi o caso exemplificado na Introdução e é fácil verificar que recai diretamente no Caso 2 já visto, bastando fazer

$$d_{ij}(\underline{x}) = d_i(\underline{x}) - d_j(\underline{x}) = \underline{v}_i^T \underline{x}_e - \underline{v}_j^T \underline{x}_e = \underline{v}_{ij}^T \underline{x}_e$$

onde neste caso $\underline{v}_{ij} = \underline{v}_i - \underline{v}_j$.

Se $d_i(\underline{x}) > d_j(\underline{x}) \quad \forall \quad j \neq i$ então $d_{ij}(\underline{x}) > 0 \quad \forall \quad j \neq i$.

Deve-se enfatizar que o Caso 3 recai no Caso 2, mas não vice-versa, pois no Caso 2 as funções de decisão d_{ij} são, a princípio, todas linearmente independentes, ao passo que no Caso 3 elas são linearmente dependentes. Devido a isto, não há região de classificação indeterminada, excetuando-se as fronteiras. Este fato decorre naturalmente da regra de decisão que procura, para um dado $\underline{x}$, o máximo dos $d_i(\underline{x})$ para i=1,2,...,c. Novamente as regiões de classificação são convexas e conexas.

⊙⊙⊙⊙⊙ Exemplo: Tomemos d = 2 e c = 3, com

$$d_1(\underline{x}) = x_1+x_2+2 \; ; \; d_2(\underline{x}) = -4x_1+3x_2-3 \; ; \; d_3(\underline{x}) = x_1-2x_2-4$$

As fronteiras entre as 3 regiões são determinadas por

$$d_{12}(\underline{x}) = 0 \quad \Rightarrow \quad d_1(\underline{x})-d_2(\underline{x}) = 5x_1-2x_2+5 = 0$$

$$d_{13}(\underline{x}) = 0 \quad \Rightarrow \quad d_1(\underline{x})-d_3(\underline{x}) = 3x_2 + 6 = 0$$

$$d_{23}(\underline{x}) = 0 \quad \Rightarrow \quad d_2(\underline{x})-d_3(\underline{x}) = -5x_1+5x_2+1 = 0$$

A Fig. 3 esquematiza as fronteiras de decisão para este exemplo. É fácil ver que a última equação de fronteira é uma combinação linear das anteriores e portanto a solução

$$\underline{x}^* = [\, x_1^* \quad x_2^* \,] \qquad \text{de} \qquad \begin{bmatrix} 5 & -2 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} x_1^* \\ x_2^* \end{bmatrix} = \begin{bmatrix} -5 \\ -6 \end{bmatrix}$$

também satisfaz a equação $-5x_1+5x_2+1 = 0$. Conclui-se (neste caso especial) que as retas que caracterizam as 3 fronteiras passam todas por um mesmo ponto do plano, deixando de haver região com classificação indeterminada. Apenas as fronteiras de decisão é que são de classificação indeterminada.

⊙⊙⊙⊙⊙

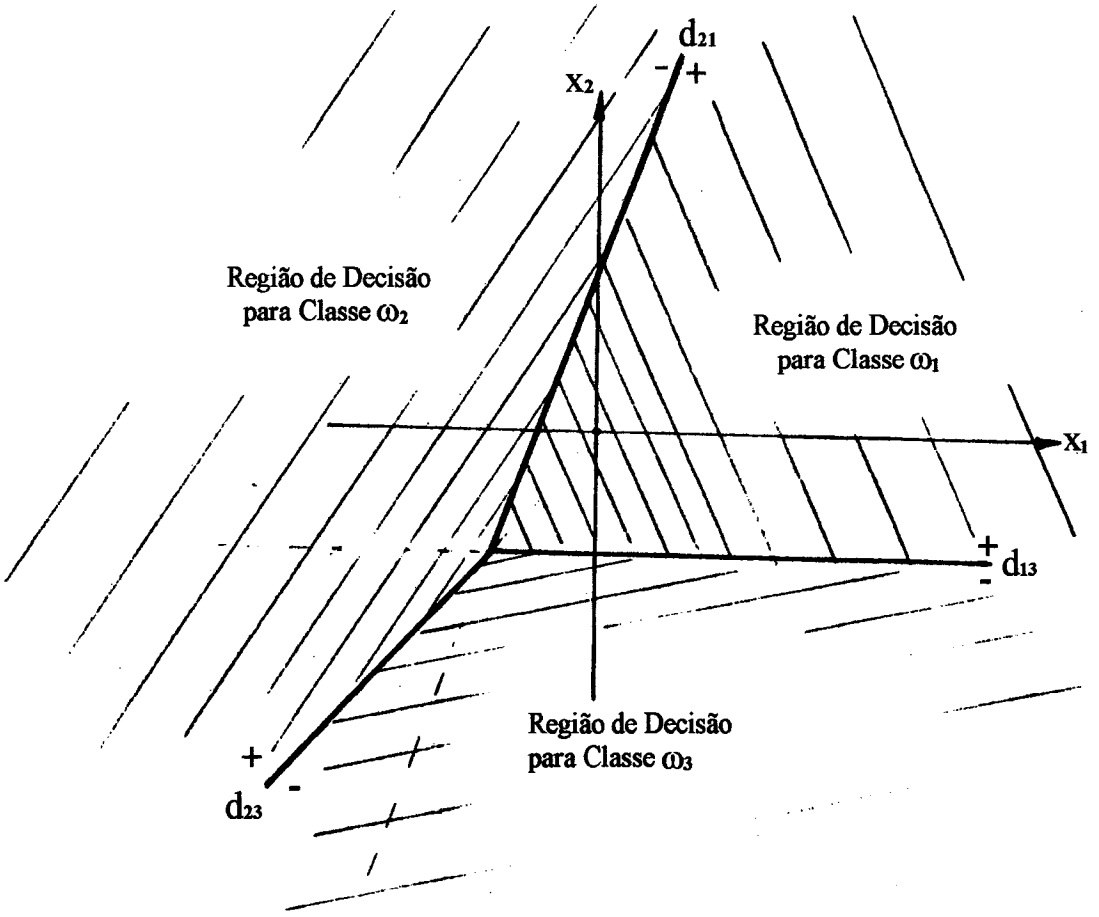


Fig. 3 – Fronteiras e regiões de decisão para um exemplo no Caso 3. Neste exemplo especial (em que há 3 classes e a dimensão é 2) as três fronteiras passam por um mesmo ponto.

⊙⊙⊙⊙⊙ Exemplo: Para não se ficar com uma idéia errada que no Caso 3 as fronteiras sempre passam por um mesmo ponto comum, esboça-se na Fig. 4 um exemplo em que d = 2 e c = 4.

⊙⊙⊙⊙⊙

É claro que em termos práticos é mais conveniente utilizar as funções $d_i(\underline{x})$ ao invés das $d_{ij}(\underline{x})$ pois o número destas últimas é muito maior do daquelas. O Caso 3 é o mais recomendável pois não há regiões de classificação indeterminada (exceto as fronteiras).

Sempre que for possível classificar corretamente todos os padrões do conjunto de treinamento com funções de decisão lineares como nos Casos 1 a 3 diz-se que as classes são linearmente separáveis. Um classificador de padrões baseado em funções de decisão lineares é chamado "máquina linear".

PROPRIEDADES GEOMÉTRICAS

Nos casos 1 e 2 já discutidos de separabilidade entre classes vimos que as fronteiras entre classes eram obtidas igualando-se as funções de decisão a zero. No Caso 3 basta utilizar as funções $d_{ij}(\underline{x}) = d_i(\underline{x}) - d_j(\underline{x})$, e com isto a fronteira é calculada fazendo-se $d_{ij} = 0$.

A análise será feita com

$$d(\underline{x}) = v_1 x_1 + v_2 x_2 + \ldots + v_d x_d + v_{d+1} = 0$$

ou

$$d(\underline{x}) = \underline{v}_0^T \cdot \underline{x} + v_{d+1} = 0 \qquad (3)$$

onde $\underline{v}_0 = [v_1\ v_2\ \ldots . v_d]^T$, como feito também em (1).

No caso de dimensão d=2 temos em (3) a representação de uma reta. Para dimensão d=3 temos a equação de um plano no espaço 3-D. Para d > 3 temos como separatriz, ou fronteira, um hiperplano no espaço d-dimensional.

Seja $\underline{u}$ um vetor unitário normal ao hiper-plano em um ponto $\underline{p}$ deste

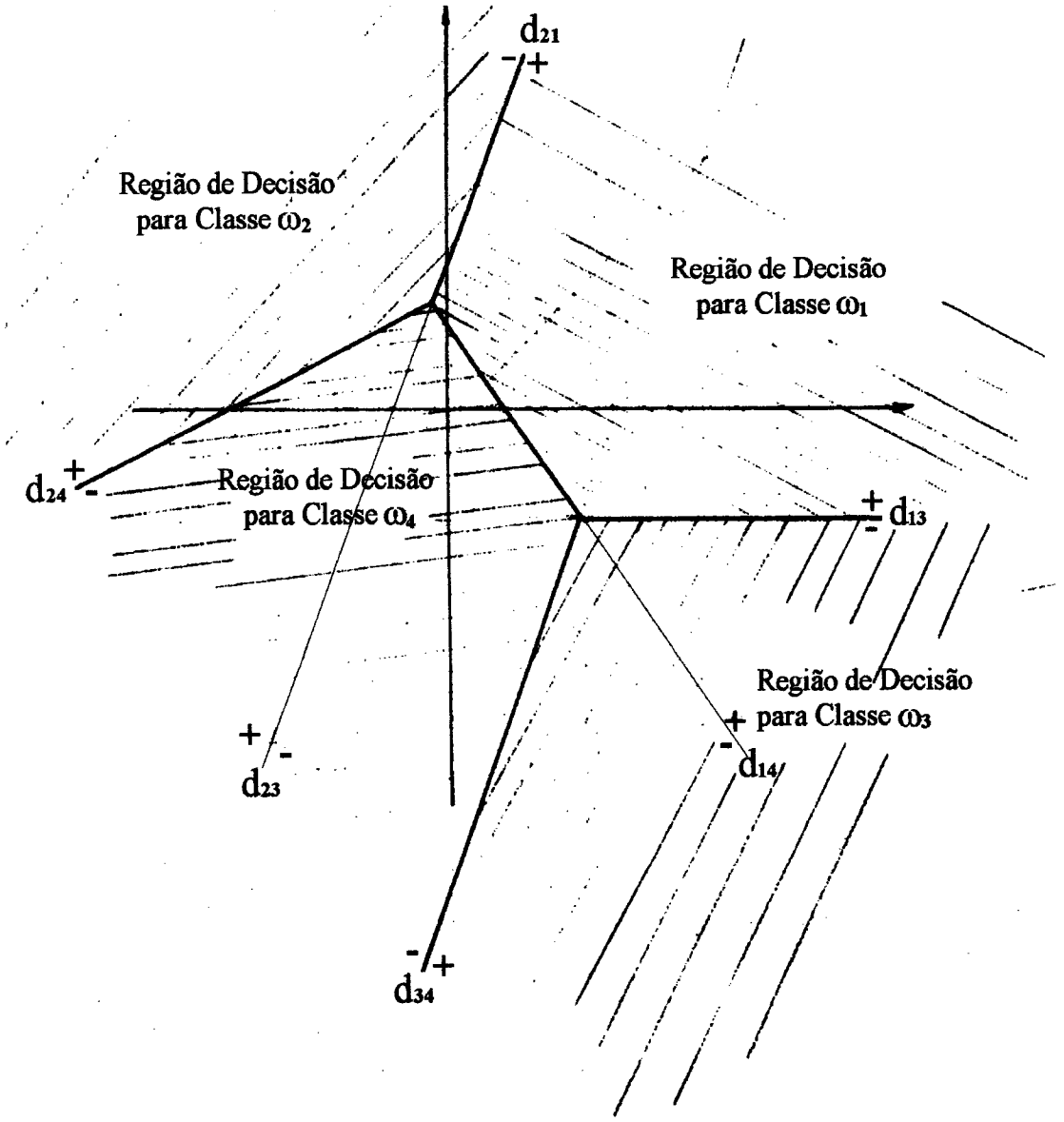


Fig. 4 – Exemplo para o Caso 3 em que há 4 classes, com funções de decisão $d_1(\underline{x})=x_1+x_2+2$; $d_2(\underline{x})=-4x_1+3x_2-3$; $d_3(\underline{x})=x_1-2x_2-4$; $d_4(\underline{x})=-2x_1-x_2+5$. Note que neste exemplo (como na grande maioria dos casos) as fronteiras não passam por um mesmo ponto.

(vide Fig. 5) e orientado para o lado positivo do hiperplano, e seja $\underline{x}$ pertencente ao hiperplano (i.e. $\underline{v}_0^T \cdot \underline{x} + v_{d+1} = 0$), então:

$$\underline{u}^T \cdot (\underline{x}-\underline{p}) = 0 \tag{4}$$

ou

$$\underline{u}^T \cdot \underline{x} = \underline{u}^T \cdot \underline{p} \tag{5}$$

Dividindo-se (3) por $\| \underline{v}_0 \| = \left(\sum_{i=1}^{d} v_i^2 \right)^{1/2} = (\underline{v}_0^T \, \underline{v}_0)^{1/2}$, obtém-se

$$\frac{\underline{v}_o^T \cdot \underline{x}}{\| \underline{v}_o \|} = - \frac{v_{d+1}}{\| \underline{v}_o \|} \tag{6}$$

De (6), como $\underline{p} \in$ hiperplano, temos

$$\frac{\underline{v}_0^T \cdot \underline{p}}{\| \underline{v}_o \|} = - \frac{v_{d+1}}{\| \underline{v}_o \|} \tag{7}$$

de (6)-(7) obém-se

$$\frac{\underline{v}_o^T \cdot (\underline{x}-\underline{p})}{\| \underline{v}_o \|} = 0 \tag{8}$$

de (4) e (8) obtém-se

$$\underline{u} = \frac{\underline{v}_o}{\| \underline{v}_o \|} \tag{9}$$

de onde se conclui que os coeficientes $v_1,\ldots,v_d$ definem a direção de um vetor ortogonal ao hiperplano que denominaremos H para simplificar o palavreado. Deve-se notar que em (9) poderíamos tanto ter escolhido o vetor $\underline{u} = \underline{v}_0 / \| \underline{v}_0 \|$ quanto $\underline{u} = -\underline{v}_0 / \| \underline{v}_0 \|$. Mostraremos que a escolha (9) implica que $\underline{u}$ aponta para a região em que $d(\underline{x}) > 0$, ou seja, em que $\underline{v}_0^T \cdot \underline{x}' > -v_{d+1}$ ($\underline{x}' \in$ região "positiva"):

seja $\underline{p} \in H$, ou seja, $\underline{v}_0^T \cdot \underline{p} = - v_{d+1}$

seja $\underline{z} = \underline{p} + \underline{u}$ que portanto $\notin H$

temos $\underline{v}_0^T \cdot (\underline{p}+\underline{u}) = \underline{v}_0^T \cdot \underline{p} + \underline{v}_0^T \cdot \underline{u} = - v_{d+1} + \underbrace{(\underline{v}_0^T \cdot \underline{v}_0)/\| \underline{v}_o \|}_{> 0} > - v_{d+1}$

e portanto $\underline{p}+\underline{u}$ pertence à região em que $d(\underline{x}) > 0$ e portanto $\underline{u}$ aponta para

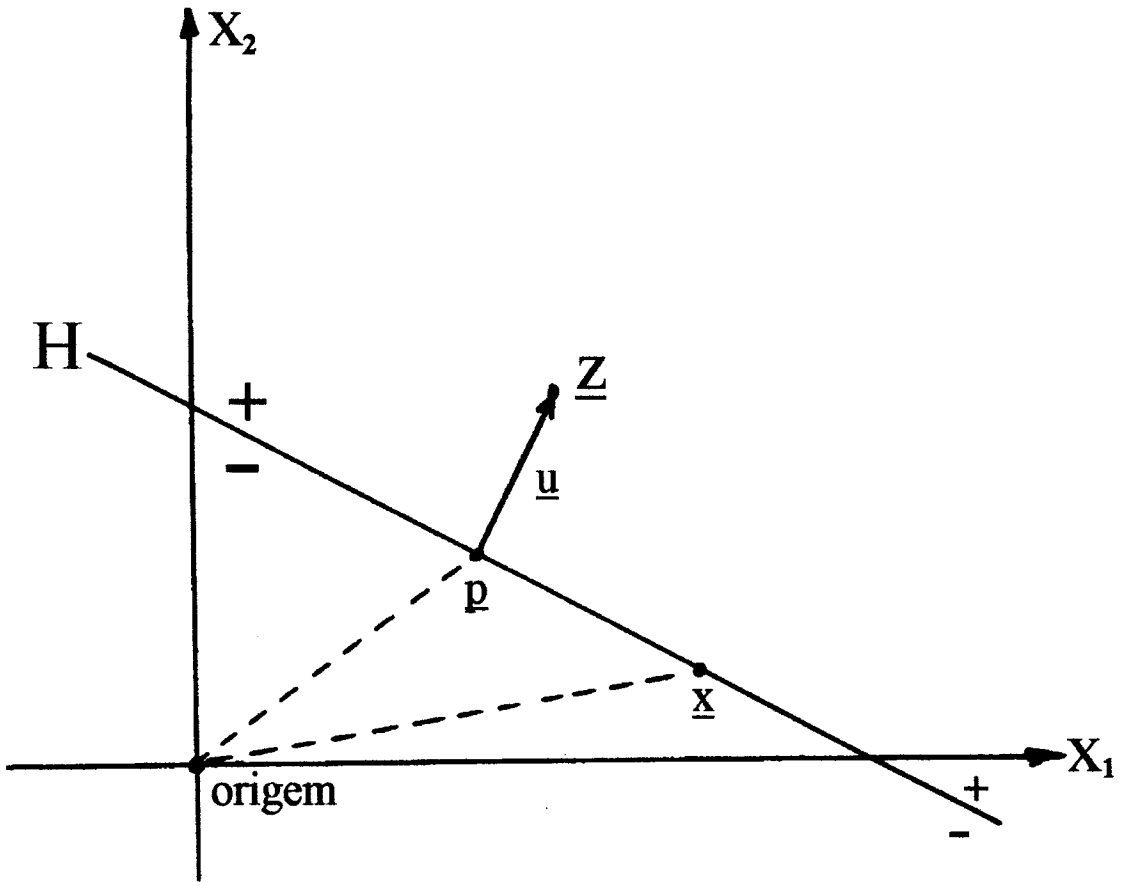


Fig. 5 - Ilustração para o caso bidimensional de uma fronteira linear e de certas propriedades geométricas.

esta região positiva do hiperplano.

A direção de $\underline{u}$ indica a direção da ortogonal do hiperplano. Se uma dada componente de $\underline{u}$ é nula, então H é paralelo ao eixo de coordenadas correspondente àquela componente. Como $\underline{u} = \underline{v}_0 / \| \underline{v}_0 \|$, basta inspecionar $\underline{v}_0$ para se ter idéia de algum paralelismo particular (vide Fig. 6).

O parâmetro v_{d+1} é denominado peso limiar e $\underline{v}_0$ é denominado vetor peso. Tomando uma função de decisão linear, o hiperplano correspondente divide o espaço em 2 regiões R_1 e R_2 . Chamemos de R_1 a região onde $d(\underline{x}) > 0$ e R_2 onde $d(\underline{x}) < 0$.

Variando v_{d+1}, deslocamos o hiperplano paralelamente. No caso em que $v_{d+1} = 0$, o hiperplano passa pela origem do espaço de atributos.

É interessante investigar a distância δ (ortogonal) algébrica de um ponto arbitrário $\underline{r} = [\ r_1 \ \ldots \ r_d\ ]^T$ ao hiperplano H definido por $d(\underline{x}) = 0$. Tomemos $\underline{r}_h$ como o ponto de projeção ortogonal de $\underline{r}$ em H. Temos então

$$\underline{r} = \underline{r}_h + \delta.\underline{u} \quad ; \quad \delta \in \mathbb{R} \tag{10}$$

$$d(\underline{r}_h) = 0 \Rightarrow \underline{v}_o^T \underline{r}_h + v_{d+1} = 0 \tag{11}$$

$$d(\underline{r}) = \underline{v}_o^T (\underline{r}_h + \delta.\underline{u}) + v_{d+1} = \underline{v}_0^T \underline{r}_h + v_{d+1} + \underline{v}_0^T \cdot \underline{u} \cdot \delta \tag{12}$$

Substituindo (9) e (11) em (12), obtém-se

$$\delta = \frac{d(\underline{r})}{\| \underline{v}_o \|} \tag{13}$$

com $\delta > 0$ se $\underline{r} \in R_1$ e $\delta < 0$ se $\underline{r} \in R_2$.

É interessante tomar o ponto particular que é a origem do espaço de atributos $\underline{y} = \underline{0}$. A sua distância algébrica a H é

$$\delta_0 = \frac{v_{d+1}}{\| \underline{v}_o \|} \tag{13}$$

e portanto, se $\delta_0 > 0$, ou seja $v_{d+1} > 0$, a origem pertence a R_1 e em caso contrário a R_2 . Se $\delta_0 = 0$ então a origem pertence a H .

Resumindo: $\underline{V}_o$ indica a orientação de H enquanto v_{d+1} indica a sua

localização em relação à origem.

As regiões de decisão quando se usam funções de decisão lineares são convexas e simplesmente conexas ("simply connected"), esta última propriedade significando que é uma região sem "buracos". Estas duas propriedades limitam um pouco a utilidade das máquinas lineares uma vez que tendem a favorecer um bom desempenho para os problemas em que as densidades $p(\underline{x}|\omega_i)$ são unimodais (mesmo nestas pode-se ter regiões não conexas para um classificador de Bayes e portanto não aproximáveis com o uso de funções de decisão lineares). Outra limitação que merece ser relembrada é que o desempenho pode ser ruim se as classes não forem linearmente separáveis (vide Fig. 7).

Uma outra visão da utilização de uma função de decisão linear $d(\underline{x}) = \underline{v}_0^T \underline{x} + v_{d+1}$ é que se está projetando o padrão $\underline{x}$ na direção do vetor $\underline{v}_0$ e comparando o valor da projeção com o peso limiar v_{d+1}. Este enfoque é utilizado na teoria do discriminante linear de Fisher (a ser estudada em um capítulo específico) em que se utilizam considerações estatísticas para se determinar, não iterativamente, o vetor peso $\underline{v}$.

TÉCNICAS ITERATIVAS PARA DETERMINAÇÃO DE UM CLASSIFICADOR LINEAR

Há várias técnicas iterativas, determinísticas ou probabilísticas, que podem ser utilizadas para a obtenção de vetores peso $\underline{v}_i$ que resultem em um classificador com desempenho adequado (dentro das condições do problema). O processo de ajuste iterativo, baseado em um conjunto de amostras com classificação conhecida, é denominado de treinamento ou aprendizado. O conjunto de amostras já classificadas é portanto denominado conjunto de treinamento.

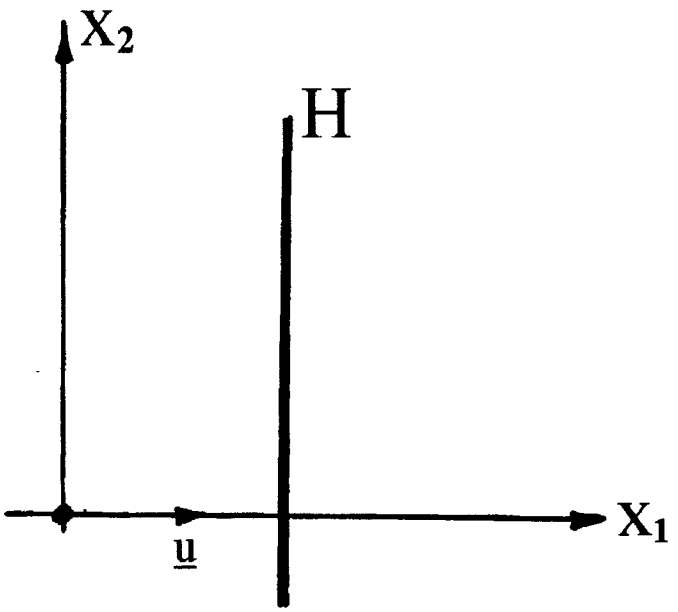


Fig. 6 – Exemplo em que $\underline{u} = [1\ 0]^T$ e a fronteira (uma reta) é paralela ao eixo x_2.

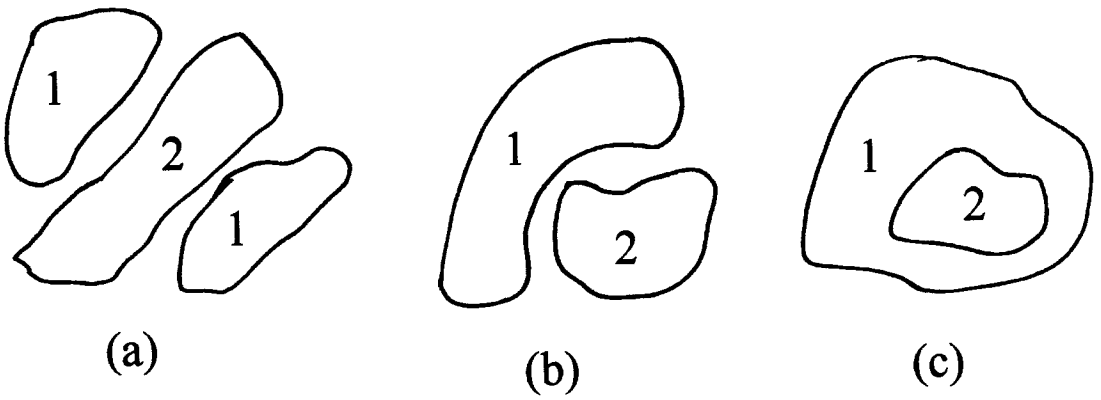


Fig. 7 – Exemplos de duas classes não linearmente separáveis: (a) classe ω_1 bimodal e classe ω_2 localizada entre duas modas; (b) e (c) classes unimodais.

Algoritmo Perceptron

O algoritmo Perceptron básico (Rosenblatt, 1962) é um precursor de muitos algoritmos já descritos na área de reconhecimento de padrões. O trabalho de Rosenblatt pode inclusive ser considerado como a pedra fundamental da atual área de redes neurais. Quando as classes são linearmente separáveis e se dispõe de um conjunto de treinamento com N amostras, o algoritmo Perceptron converge em um número finito de iterações (Sklansky e Wassel, 1981). Como o algoritmo Perceptron não utiliza nenhuma hipótese ou conhecimento a priori sobre a descrição probabilística das classes ele é um algoritmo determinístico.

No caso de duas classes, o vetor $\underline{v}$ da função de decisão linear $d(\underline{x}) = \underline{v}^T \cdot \underline{x}_e$ deve ser tal que

$$\underline{v}^T \cdot \underline{x}_e \begin{cases} >0 \text{ se } \underline{x} \in \omega_1 \\ <0 \text{ se } \underline{x} \in \omega_2 \end{cases} \qquad (15)$$

onde $\underline{x}_e = [\underline{x}^T \; 1]^T$.

O algoritmo básico Perceptron consiste na correção do valor do vetor $\underline{v}$ na iteração k, indicado por $\underline{v}(k)$, para tentar chegar nas imposições de (15). Para fins de descrição do algoritmo os elementos do conjunto de treinamento são denotados, $\underline{x}(n)$, n = 0,1,..,N-1, e acrescidos de uma dimensão com valor fixo em 1, obtendo-se elementos $\underline{x}_e(n) = [\underline{x}^T(n) \; 1]^T$. Há três possibilidades para a correção de um dado $\underline{v}(k)$:

1) $\underline{v}(k+1) = \underline{v}(k) + \gamma \cdot \underline{x}_e(k)$ se $\underline{v}^T(k) \cdot \underline{x}_e(k) < 0$ para $\underline{x}(k) \in \omega_1$

2) $\underline{v}(k+1) = \underline{v}(k) - \gamma \cdot \underline{x}_e(k)$ se $\underline{v}^T(k) \cdot \underline{x}_e(k) > 0$ para $\underline{x}(k) \in \omega_2$

3) $\underline{v}(k+1) = \underline{v}(k)$ em caso contrário

onde γ é um fator de correção positivo, e inicia-se o processo com um valor arbitrário para $\underline{v}(0)$. Com o valor corrigido de $\underline{v}$ obtemos uma melhoria no valor da função de decisão, no sentido que seu valor se torna mais próximo

de satisfazer (15). Isto é ilustrado para o caso em que $\underline{x}(k) \in \omega_1$ mas $\underline{v}^T(k)\cdot\underline{x}_e(k) < 0$, tendo-se então, com o valor corrigido de $\underline{v}$, um valor para a função de decisão igual a

$$\underline{v}^T(k)\cdot\underline{x}_e(k) + \gamma\cdot\underline{x}_e^T(k)\cdot\underline{x}_e(k)$$

ou seja, o valor da função de decisão com o $\underline{v}$ corrigido passa a ser mais positivo, uma vez que γ é positivo. A regra de treinamento ou de correção de erro (de classificação) apresentada é denominada de regra de incremento proporcional ou de recompensa-punição ("reward and punishment"). Nesta última nomenclatura, a punição ocorre nos dois casos em que há erro e a recompensa na realidade é uma ausência de punição. No caso de classes linearmente separáveis, o algoritmo deve ser iterado até que não haja alteração no valor de $\underline{v}$ durante N iterações seguidas, ou equivalentemente, até que todas as amostras do conjunto de treinamento sejam corretamente classificadas. Isto significa que se após as primeiras N iterações ainda não houve convergência, deve-se utilizar os elementos do conjunto de treinamento novamente, e assim por diante, o que em termos matemáticos pode ser expresso como $n = k_{[\text{mod } N]}$.

No caso de c classes linearmente separáveis, pode-se pensar nos 3 casos de funções de decisão lineares apresentados anteriormente. No caso 1, para o treinamento, basta tomar para cada função de decisão $d_i(\underline{x})$ a dicotomia ω_i/não ω_i. Para o caso 2, tomam-se as classes aos pares e para o caso 3 pode-se utilizar o algoritmo apresentado a seguir.

No caso 3 há c funções de decisão lineares, $d_i(\underline{x}) = \underline{v}_i^T\cdot\underline{x}_e$, i=1,2,...,c, de tal forma que se $\underline{x} \in \omega_i$ então $d_i(\underline{x}) > d_j(\underline{x})$ para qualquer $j \neq i$, com j,i = 1,2,...,c. Supondo que na k-ésima iteração do algoritmo para se determinar os vetores peso tem-se

$\underline{x}(k) \in \omega_i$ e

$d_i(\underline{x}(k)) \leq d_m(\underline{x}(k))$ para M valores de $m \neq i$, então

$$\begin{cases} \underline{v}_i(k+1) = \underline{v}_i(k) + M\cdot\gamma\cdot\underline{x}_e(k) \\ \underline{v}_m(k+1) = \underline{v}_m(k) - \gamma\cdot\underline{x}_e(k) \end{cases} \quad \text{se } d_i(\underline{x}) \leq d_m(\underline{x}) \ , \ m\neq i$$

$$\underline{v}_j(k+1) = \underline{v}_j(k) \quad \text{se } d_i(\underline{x}) > d_j(\underline{x}) \ , \ j\neq i$$

onde γ é uma constante positiva. Observações semelhantes às feitas para o caso de duas classes também valem aqui (por exemplo, os vetores iniciais $\underline{v}_j(0)$ são arbitrários).

FUNÇÕES DE DECISÃO LINEARES GENERALIZADAS

Pode-se, a partir de funções de decisão lineares obter funções de decisão mais gerais. Isto pode ser feito, por exemplo, através de funções $\phi_i(\underline{x}):\mathbb{R}^d \rightarrow \mathbb{R}$, como formalizado abaixo:

$$d(\underline{x}) = v_1\phi_1(\underline{x}) + v_2\phi_2(\underline{x}) + \ldots + v_d\phi_d(\underline{x}) + v_{d+1} \tag{16}$$

$$\therefore \quad d(\underline{x}) = \sum_{i=1}^{d+1} v_i\cdot\phi_i(\underline{x}) \qquad \text{com } \phi_{d+1}(\underline{x}) = 1 \tag{17}$$

Uma vez calculados os valores $\phi_i(\underline{x})$, tem-se um vetor de elementos reais $\underline{z}$ e então $d(\underline{x}) = \underline{v}^T\cdot\underline{z}$ e portanto se cada padrão $\underline{x}$ é transformado no correspondente padrão mapeado $\underline{z}$ passamos a ter uma função de decisão linear.

⊙⊙⊙⊙⊙ Exemplo: Caso de $2^{\underline{a}}$ ordem com função de decisão quadrática

$$d(\underline{x}) = v_{11}x_1^2 + v_{12}x_1x_2 + v_{22}x_2^2 + v_1x_1 + v_2x_2 + v_3 \tag{18}$$

que é claramente não linear em x_1 e x_2. Se tomarmos

$$\underline{v} = [v_{11}\ v_{12}\ v_{22}\ v_1\ v_2\ v_3]^T \text{ e } \underline{z} = [x_1^2\ \ x_1x_2\ \ x_2^2\ \ x_1\ \ x_2\ \ 1]^T$$

teremos

$$d(\underline{x}) = \underline{v}^T\cdot\underline{z} \tag{19}$$

Ainda podemos notar que (18) contém 3 termos quadráticos, 2 termos

lineares e 1 constante, o que sugere escrever

$$d(\underline{x}) = \underline{x}^T A \underline{x} + \underline{b}^T \underline{x} + c$$

onde $A = \begin{bmatrix} v_{11} & 0.5v_{12} \\ 0.5v_{12} & v_{22} \end{bmatrix}$ $\underline{b} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$ $c = v_3$

Se A for a matriz identidade então a parte quadrática da função de decisão define uma circunferência. No caso d-dimensional seria uma hiperesfera.

⊙⊙⊙⊙⊙

As funções $\phi_i(\underline{x})$ podem ser dos mais variados tipos, incluido polinômios de Hermite, de Legendre, de Laguerre (vide seção 2.7 de Tou & Gonzalez, 1974).

COMPLEMENTOS sobre TÉCNICAS ITERATIVAS no PROJETO de CLASSIFICADORES LINEARES

Dada uma função f(w) o gradiente df(w)/dw, por vezes indicado $\nabla f(\underline{w})$, é um vetor que aponta na direção em que há a maior taxa de acréscimo em f(w). Caso a função tenha um mínimo então $-\nabla f(\underline{w})$ aponta para a direção que causa o maior decréscimo em f(w).

No caso de 2 classes linearmente separáveis temos para uma regra de decisão linear com taxa de erro nula

$$\underline{w}^T\underline{x} > 0 \text{ se } \underline{x} \text{ " } \in \text{ " } \omega_1 \qquad ou \qquad \underline{w}^T\underline{x} < 0 \text{ se } \underline{x} \text{ " } \in \text{ " } \omega_2 \tag{1}$$

Para facilitar a notação, invertemos o sinal de todas as amostras da classe ω_2 e portanto $\underline{w}^T(-\underline{x})>0$ para $\underline{x}$ "$\in$" ω_2 e portanto um classificador que redunda em taxa de erro nula satisfaz

$$\underline{w}^T\underline{x} > 0, \forall \underline{x}$$

e como normalmente se tem um conjunto de treinamento $Q_N=\{\underline{x}_1, \underline{x}_2,...,\underline{x}_N\}$ tem-se N desigualdades $\underline{w}^T\underline{x}_i>0$ i=1,...,N.

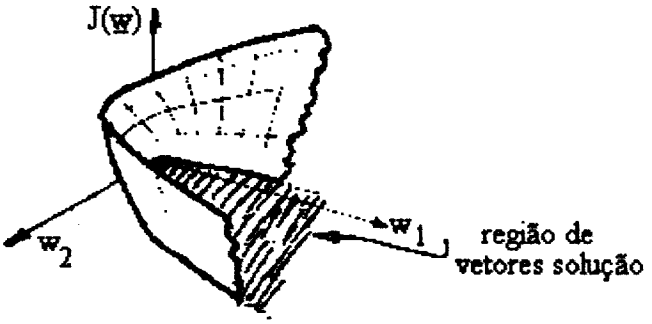
Função Critério para o Perceptron

Devemos resolver $\underline{w}^T\underline{x}_i>0$ e uma primeira idéia de função critério seria o número de amostras com classificação errada. Se desejamos uma busca por gradiente, as descontinuidades desta proposta mostram que não é adequada.

A Função Critério Perceptron é

$$J(\underline{w}) = \sum_{\underline{x} \in \chi} (-\underline{w}^T \underline{x}) \tag{2}$$

onde χ é o conjunto de vetores cuja classificação por $\underline{w}$ está errada, ou seja $\underline{w}^T\underline{x} \leq 0$. Portanto $J(\underline{w})$ é sempre não-negativo, sendo nulo quer para $\underline{w}$ na fronteira ou para vetores $\underline{w}$ que são solução.



Para a função critério Perceptron

$$\nabla J(\underline{w}) = \sum_{\underline{x} \in \chi} (-\underline{x}) \tag{3}$$

e como a correção sobre $\underline{w}(k)$ para obter $\underline{w}(k+1)$ é

$$\underline{w}(k+1) = \underline{w}(k) - \gamma(k) \nabla J(\underline{w}) \mid_{\underline{w}=\underline{w}(k)} \tag{4}$$

onde $\gamma(k)$ é um fator de correção positivo, tem-se

$$\underline{w}(k+1) = \underline{w}(k) + \gamma(k) \sum_{\underline{x} \in \chi_k} \underline{x}$$

onde χ_k é o conjunto de vetores de Q_N que são classificados erradamente por $\underline{w}(k)$.

Caso a cada iteração para $\underline{w}(k)$ se tome um vetor $\underline{x}_i$ de Q_N (que é associado a $\underline{x}(k)$) e numerando os elementos assim obtidos de 1 até ∞ (pode-se repetir ciclicamente os elementos de Q_N) temos:

$$\underline{w}(k+1) = \underline{w}(k) + \gamma(k)\underline{x}(k) \tag{6}$$

se $\underline{x}^T(k)\ \underline{w}(k)<0$

(onde continuamos com a convenção de trocar o sinal de todos os vetores associados a ω_2)

que é o algoritmo básico Perceptron se tomarmos $\gamma(k)$ constante para $\forall k$.

Uma formulação alternativa é tomar:

$$J(\underline{w}) = \frac{1}{2}\left(|\underline{w}^T\underline{x}| - \underline{w}^T\underline{x}\right) \tag{7}$$

que é positivo para $\underline{w}^T\underline{x}<0$ e nulo para $\underline{w}^T\underline{x}\geq 0$, sendo então adequado para se buscar por descida de gradiente um $\underline{w}$ tal que $\underline{w}^T\underline{w}>0$ para $\underline{x}\in Q_N$ (note que $\underline{w}^T\underline{x}=0$ para $\underline{x}\in Q_N$, com $\underline{x} \neq \underline{0}$, significa que $\underline{w}=\underline{0}$, o que não nos interessa). Com este novo $J(\underline{w})$ temos

$$\nabla J(\underline{w}) = \frac{1}{2}\left[\underline{x}\cdot sign\left(\underline{w}^T\underline{x}\right) - \underline{x}\right] \tag{8}$$

onde $sign(\alpha) = \begin{cases} 1 \text{ se } \alpha > 0 \\ -1 \text{ se } \alpha \leq 0 \end{cases}$ e com isto a atualização em $\underline{w}(k)$ fica

$$\underline{w}(k+1) = \underline{w}(k) + \frac{\gamma(k)}{2}\left[\underline{x}(k) - \underline{x}(k)sign(\underline{w}^T(k).\underline{x}(k))\right] \tag{9}$$

ou seja

$$\underline{w}(k+1) = \underline{w}(k) + \gamma(k)\begin{cases} 0 \text{ se } \underline{w}^T(k)\underline{x}(k) > 0 \\ \underline{x}(k) \text{ se } \underline{w}^T(k)\underline{x}(k) \leq 0 \end{cases} \tag{10}$$

que é exatamente o algoritmo básico Perceptron se $\gamma(k)$ for constante.

Função Critério Erro Médio Quadrático

Ao invés de procurar $\underline{w}$ para satisfazer

$$\underline{x}_i^T \underline{w} > 0, \qquad i = 1, 2, ..., N \tag{11}$$

procuraremos $\underline{w}$ para que

$$\underline{x}_i^T \underline{w} = b_i, \qquad i = 1, 2, ..., N \quad e \quad b_i > 0 \tag{12}$$

o que é equivalente. Definimos

$$J(\underline{w}) = \frac{1}{2} \sum_{i=1}^{N} (\underline{w}^T \underline{x}_i - b_i)^2 \tag{13}$$

que é não negativo, com valor mínimo atingido quando $\underline{x}_i^T \underline{w} = \underline{w}^T \underline{x}_i = b_i$. Agrupando os vetores $\underline{x}_i$ na matriz X e as constantes b_i no vetor $\underline{b}$:

$X = \begin{bmatrix} \underline{x}_1^T \\ \underline{x}_2^T \\ \vdots \\ \underline{x}_N^T \end{bmatrix}$; $\underline{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_N \end{bmatrix}$, e com esta notação temos, ao invés de (12)

$$\underset{(Nxd)}{X} \underline{w} = \underline{b} \tag{14}$$

e ao invés de (13):

$$J(\underline{w}) = \frac{1}{2} \|X\underline{w} - \underline{b}\|_2^2 = \frac{1}{2} \left[\underline{w}^T X^T X \underline{w} - \underline{w}^T X^T \underline{b} - \underline{b}^T X \underline{w} + \underline{b}^T \underline{b}\right] \tag{15}$$

Podemos examinar um solução não iterativa, bastando fazer $\frac{\partial J(\underline{w})}{\partial \underline{w}} = \underline{0}$ resultando

$$X^T X \underline{w}^* - X^T \underline{b} = \underline{0} \tag{16}$$

$$\text{ou} \quad \underline{w}^* = (X^T X)^{-1} X^T \underline{b} \tag{17}$$

supondo X^TX invertível. A matriz $(X^TX)^{-1}X^T$é denominada a pseudo-inversa de X, indicada por $X^\sharp$, devendo-se notar que $X^\sharp$ tem dimensão d×N ao passo que X tem dimensão N×d.

Uma solução iterativa pode ser encontrada usando busca por gradiente que tem a vantagem de não depender do fato de X^TX ser singular e evita a manipulação de grandes matrizes. O gradiente de $J(\underline{w})=\frac{1}{2}\|X\underline{w}-\underline{b}\|_2^2$ é $X^T(X\underline{w}-\underline{b})$ e portanto o algoritmo de busca por descida de gradiente fica:

$$\underline{w}(k+1) = \underline{w}(k) - \gamma(k) X^T (X\underline{w}(k) - \underline{b}) \tag{18}$$

Nesta expressão tem-se a explicitação que o conjunto de treinamento (linhas de X) é utilizado de uma vez. A simplificação para o caso de utilização sequencial das amostras produz a regra de Widrow-Hoff, ou regra LMS ou LMSE:

$$\underline{w}(k+1) = \underline{w}(k) + \gamma(k) \left[b_k - \underline{w}^T(k) \underline{x}_k \right] \underline{x}_k \tag{19}$$

com $\gamma(k)$ uma sequência decrescente para se garantir a convergência (p.ex. $\gamma(k)=\gamma/k$, com $\gamma \in \Re^+$).

Este algoritmo converge para uma reta, quer as classes sejam separáveis ou não, o que é uma melhoria em relação ao Perceptron. Entretanto, no caso das classes serem separáveis não se pode garantir a convergência para uma reta que de fato separe as classes, o que é indesejável. Isto ocorre devido ao fato de se impor a priori um valor fixo (positivo) para $\underline{b}$ quando o que se pode afirmar é que para cada $\underline{w}$ que separa perfeitamente as 2 classes existe

um vetor $\underline{r}$, com $r_i > 0$, tal que $X\underline{w}=\underline{r}$. Portanto não devemos pré-fixar o vetor $\underline{b}$, deixando-o livre na otimização. Temos

$$J(\underline{w},\underline{b}) = \frac{1}{2}\|X\underline{w} - \underline{b}\|_2^2 \tag{20}$$

onde se variam $\underline{w}$ e $\underline{b}$, só que $\underline{b}$ é variado condicionado a $\underline{b} > \underline{0}$ (isto é, $b_i > 0$ i=1,...N). Os gradientes são:

$$\frac{\partial J}{\partial \underline{w}} = X^T(X\underline{w} - \underline{b}) \tag{21}$$

e

$$\frac{\partial J}{\partial \underline{b}} = -X\underline{w} + \underline{b} \tag{22}$$

Como $\underline{w}$ não tem limitações no seu domínio de variação, então fazemos $\partial J/\partial \underline{w}=\underline{0}$ para achar o $\underline{w}$ ótimo:

$$\underline{w}^* = (X^T X)^{-1} X^T \underline{b} = X^{\sharp}\underline{b} \tag{23}$$

que será usado para relacionar $\underline{w}(k+1)$ com $\underline{b}(k+1)$:

$$\underline{w}(k+1) = X^{\sharp}\underline{b}(k+1) \tag{24}$$

Na busca por gradiente, devemos tomar cuidado para garantir que $\underline{b}(k)$ permaneça >0, ou seja usamos

$$\underline{b}(k+1) = \underline{b}(k) + 2\gamma(X\underline{w}(k) - \underline{b}(k)) \tag{25}$$

se o incremento $\delta\underline{b}(k)=2\gamma(X\underline{w}(k)-\underline{b}(k))$ for >0 e usamos $\underline{b}(k+1)=\underline{b}(k)$ em caso contrário. Definindo um vetor erro

$$\underline{e}(k) = X\underline{w}(k) - b(\underline{k}) \tag{26}$$

e $|\underline{e}(k)|$ como um vetor com módulo em todas componentes de $\underline{e}(k)$ tem-se

$$\underline{b}(k+1) = \underline{b}(k) + \gamma\,[\underline{e}(k) + |\underline{e}(k)|] \tag{27}$$

e como $\underline{w}(k+1) = X^{\sharp}\underline{b}(k+1)$:

$$\underline{w}(k+1) = \underline{w}(k) + \gamma X^{\sharp}\,[\underline{e}(k) + |\underline{e}(k)|] \tag{28}$$

As equações recursivas (27) e (28) ou (24) e (27) fornecem o algoritmo de Ho-Kashyap que é um algoritmo LMSE melhor do que o apresentado antes.

Se as classes são linearmente separáveis e se $0<\gamma<1$, então o algoritmo fornece uma solução em número finito de passos. Se ocorrer $\underline{e}(k)=\underline{0}$ então obtém-se uma solução com a indicação de que as 2 classes são linearmente separáveis. Caso elas não sejam, então a parada do algoritmo se dá quando um dos componentes de $\underline{e}(k)$ se torna negativo (um sinal de classes que não tem separabilidade linear).

REGRA DE DECISÃO DOS VIZINHOS MAIS PRÓXIMOS

INTRODUÇÃO

Na abordagem por teoria de decisão estatística é necessário se conhecer todas as distribuições de probabilidade. Entretanto, na metodologia por (ou de) vizinho mais próximo (VMP) este conhecimento não é necessário, sendo entretanto necessário se dispor de um grande número de padrões já classificados. É claro que de posse destes padrões já classificados pode-se efetuar uma estimação das distribuições probabilísticas desconhecidas para então aplicar a regra de decisão de Bayes. Uma alternativa é tentar fundir estas duas tarefas (estimação e classificação) em uma só, o que é feito nos classificadores por vizinho(s) mais próximo(s).

REGRA DE DECISÃO DO VIZINHO MAIS PRÓXIMO (1-NN)

É dado um conjunto de amostras (vetores de atributos ou padrões) já classificadas $Q_N = \{ \underline{x}_1 , \underline{x}_2 ,,,, \underline{x}_N \}$. Este será o conjunto de referência para o método dos VMPs. Dado um padrão $\underline{x}$ cuja classificação é desconhecida, a regra 1-NN classifica $\underline{x}$ na mesma classe que a do vetor $\underline{x}' \in Q_N$ mais próximo de $\underline{x}$, onde a proximidade é medida por exemplo pela distância Euclideana.

Um comentário é necessário para enfatizar que cada amostra vetorial do conjunto Q_N é obtida, de forma independente das demais, a partir da distribuição $p(\underline{x}) = \sum_{i=1}^{c} P_i \; p(\underline{x}|\omega_i)$. Isto significa que na prática não se deve

tomar o mesmo número de amostras de cada uma das c classes para formar Q_N, mas sim tomar as primeiras N amostras que forem surgindo da distribuição global, ou alternativamente, para cada vetor a ser acrescentado a Q_N, sortear a classe segundo os valores P_i, para então gerar um vetor da densidade $p(\underline{x}|\omega_i)$.

A regra de classificação do VMP irá acarretar uma taxa de erro maior do que a da regra de decisão de Bayes, mas, existe um teorema que diz que, para o caso de número infinito de padrões classificados, esta taxa não ultrapassa (sendo em geral menor que) o dobro da taxa de Bayes. Uma demonstração é apresentada a seguir. Sejam

$e_N(\underline{x},\underline{x}') \triangleq$ probabilidade condicional de erro na classificação de $\underline{x}$ tendo como VMP $\underline{x}'$ ($P[\text{erro}|\underline{x},\underline{x}']$

$e_N(\underline{x}) \triangleq$ probabilidade erro de classificação para o dado $\underline{x}$ ($P[\text{erro}|\underline{x}]$)

Temos então

$$e_N(\underline{x}) = \int e_N(\underline{x}\ ,\ \underline{x}')\ p(\underline{x}'|\underline{x})d\underline{x}' \qquad (1)$$

Como $\underline{x}'$ é o VMP de $\underline{x}$ então $p(\underline{x}'|\underline{x})$ tem um pico agudo em $\underline{x}$ e valores aproximadamente nulos nas vizinhanças de $\underline{x}$. Tem-se para $N \to \infty$ $p(\underline{x}'|\underline{x}) \to \delta(\underline{x}'-\underline{x})$. Dado $\underline{x}$, cuja classificação correta é $\theta \in \{\omega_1, \ldots \omega_c\}$, e dado $\underline{x}'$, seu VMP, com classificação $\theta' \in \{\omega_1, \ldots \omega_c\}$

$$P[\theta\ ,\ \theta'|\ \underline{x}\ ,\ \underline{x}'] = P[\theta\ |\ \underline{x}\ ]P[\theta'|\ \underline{x}'] \qquad (2)$$

(Pensar em termos da natureza gerando os $\underline{x}$, cada um em um certo estado ω_i)

Utilizando a regra de decisão do VMP , estará sendo cometido um erro sempre que $\theta \neq \theta'$ e portanto

$$e_N(\underline{x}, \underline{x}') = 1 - \sum_{i=1}^{C} P[\theta = \omega_i, \theta' = \omega_i | \underline{x}, \underline{x}']$$

$$= 1 - \sum_{i=1}^{C} P[\omega_i | \underline{x}].P[\omega_i | \underline{x}'] \quad (3)$$

Trabalhando com (1) e (3) e fazendo $N \to \infty$ (quando então $p(\underline{x}'|\underline{x}) = \delta(\underline{x}'-\underline{x})$) tem-se

$$\lim_{N\to\infty} e_N(\underline{x}) = \int \left[1 - \sum_{i=1}^{C} P(\omega_i|\underline{x})\, P(\omega_i|\underline{x}')\right] \delta(\underline{x}'-\underline{x})d\underline{x}'$$

$$= 1 - \sum_{i=1}^{C} P^2(\omega_i|\underline{x}) \quad (4)$$

A taxa média de erro ou probabilidade média de erro E é

$$E = \int e(\underline{x})p(\underline{x})d\underline{x} = \lim_{N\to\infty} E_N$$

onde $$E_N = \int e_N(\underline{x})p(\underline{x})d\underline{x}$$

e, contanto que se possa inverter a ordem de limite e integral (pode, se $p(\underline{x}) \neq 0$; se $p(\underline{x}) = 0$ eliminar da integral), tem-se

$$E = \int \left[1 - \sum_{i=1}^{C} P^2(\omega_i|\underline{x})\right] p(\underline{x})d\underline{x} \quad (5)$$

Desejamos obter alguma desigualdade relacionando E com E^* (taxa de erro de Bayes). Um limite inferior para E é o próprio E^*. Para achar um limite superior, deve-se determinar o máximo de E para um dado E^* e para isto deseja-se saber o mínimo valor de $\sum_{i=1}^{C} P^2(\omega_i|\underline{x})$ para um dado valor de $P(\omega_m|\underline{x})$, onde ω_m é a classe que maximiza $P(\omega_i|\underline{x})$ sendo portanto a classificação obtida pela regra de decisão de Bayes.

$$\sum_{i=1}^{c} P^2(\omega_i|\underline{x}) = P^2(\omega_m|\underline{x}) + \sum_{i\neq m} P^2(\omega_i|\underline{x}) \qquad (6)$$

O mínimo do 1º membro de (6) deve ser determinado sujeito às duas condições:

i $P(\omega_i|\underline{x}) \geq 0$

ii $\sum_{i\neq m} P(\omega_i|\underline{x}) = 1 - P(\omega_m|\underline{x}) = e^*(\underline{x})$, ou seja, tomando como ω_m aquela classe que maximiza $P(\omega_i|\underline{x})$ (regra de decisão de Bayes).

Podemos portanto utilizar multiplicador de Lagrange :

$$\frac{\partial}{\partial P(\omega_i|\underline{x})}\left[P^2(\omega_m|\underline{x}) + \sum_{\substack{i\neq m \\ p/\ i\neq m}} P^2(\omega_i|\underline{x}) + \lambda\left(\sum_{i\neq m} P(\omega_i|\underline{x}) - e^*(\underline{x}) \right) \right] = 0 \qquad (7)$$

$\therefore$ $2P^*(\omega_i|\underline{x}) + \lambda = 0$ $\quad \forall\, i\neq m$ onde a estrela indica "ótimo"

$\therefore$ os $P^*(\omega_i|\underline{x})$ são todos iguais para minimizar (6), $i\neq m$.

$\therefore$

$$P^*(\omega_i|\underline{x}) = \begin{cases} \dfrac{1-P(\omega_m|\underline{x})}{c-1} = \dfrac{e^*(x)}{c-1} & i\neq m \\ 1-e^*(\underline{x}) & i=m \end{cases} \qquad (8)$$

de (6) e (8) temos:

$$\sum_{i=1}^{c} P^2(\omega_i|\underline{x}) \geq [1-e^*(\underline{x})]^2 + \frac{(e^*(\underline{x}))^2}{c-1} = \sum_{i=1}^{c} (P^*(\omega_i|\underline{x}))^2 \qquad (9)$$

$$\therefore \quad 1 - \sum_{i=1}^{c} P^2(\omega_i|\underline{x}) \leq 2e^*(\underline{x}) - \frac{c}{c-1}(e^*(\underline{x}))^2 \qquad (10)$$

(10) em (5)

$$E \leq \int 2\, e^*(\underline{x})p(\underline{x})d\underline{x} - \underbrace{\frac{c}{c-1}\int (e^*(\underline{x}))^2 p(\underline{x})d\underline{x}}_{\beta^2} \qquad (11)$$

$\therefore$

$$E \leq 2\,E^{*} - \beta^{2} \tag{12}$$

e portanto para número infinito de amostras no conjunto Q_N, a taxa média de erro do classificador por VMP (1-NN) é no máximo o dobro da taxa de Bayes E^{*}. Podemos melhorar (11) :

$$\text{var}\,[e^{*}(\underline{x})] = \int \left(e^{*}(\underline{x}) - E^{*}\right)^{2} p(\underline{x})d\underline{x}$$

$$= \int (e^{*}(\underline{x}))^{2} p(\underline{x})d\underline{x} - (E^{*})^{2} \geq 0$$

e ∴

$$\int (e^{*}(\underline{x}))^{2} p(\underline{x})d\underline{x} \geq (E^{*})^{2}$$

ou seja

$$-\frac{c}{c-1}\int (e^{*}(\underline{x}))^{2} p(\underline{x})d\underline{x} \leq -\frac{c}{c-1}(E^{*})^{2} \tag{13}$$

(13) e (11)

$$\boxed{E^{*} \leq E \leq E^{*}\left[2 - \frac{c}{c-1}E^{*}\right]} \tag{14}$$

Esta desigualdade para E é mostrada graficamente na Fig. 1. O valor máximo para E^{*} foi derivado no capítulo sobre Teoria de Decisão de Bayes, sendo fornecido na expressão (37) do mesmo.

Para taxas de Bayes baixas o limite superior em (14) é praticamente $2E^{*}$. Deve-se manter em mente que o resultado em (14) só é válido para $N \to \infty$, o que limita um pouco o seu interesse prático pois o teorema não diz nada sobre a velocidade de convergência quando N tende a infinito. Outros resultados seriam necessários sobre o desempenho para número finito de amostras, mas não são fáceis de serem obtidos.

A expressão (14) pode ser utilizada de outra forma: obtendo-se uma desigualdade para a taxa de erro de Bayes em função daquela do método do VMP (1-NN). Obtém-se uma região delimitada por uma parábola

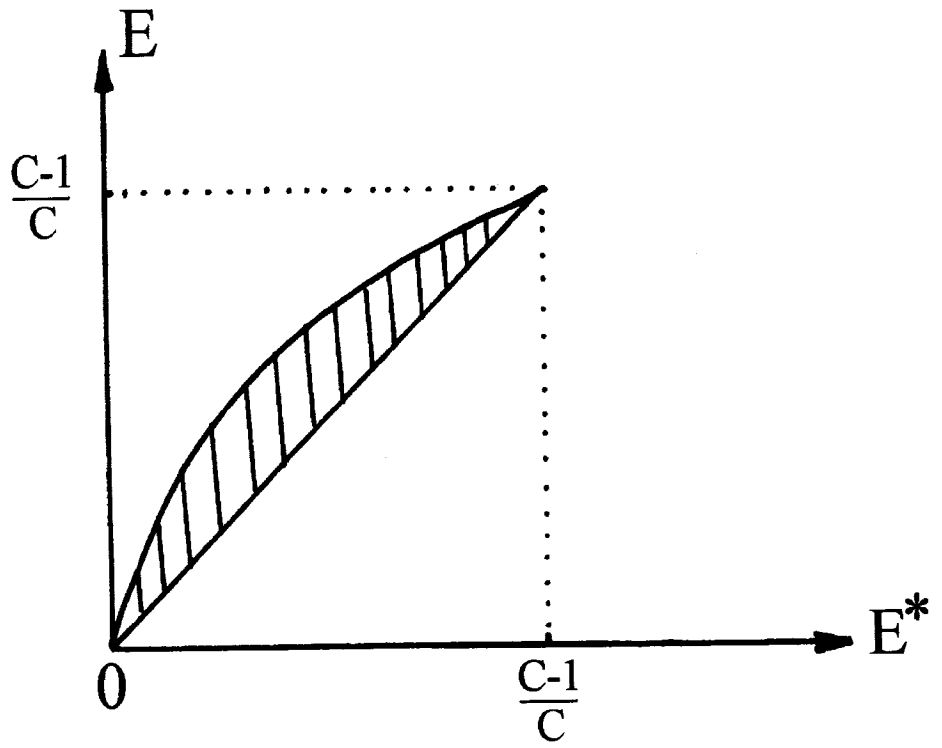


Fig. 1 – Gráfico da desigualdade da expressão (14)

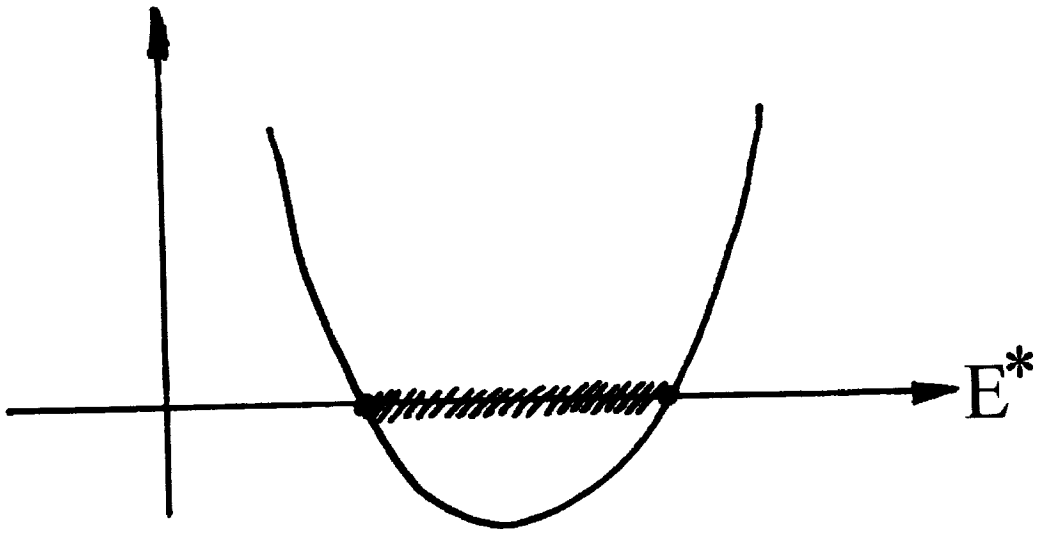


Fig. 2 – Parábola associada à desigualdade que impõe que a quadrática em E* seja não positiva. A região hachurada é a que nos interessa, sendo associada à expressão (15).

$$\left(\frac{c}{c-1}\right)(E^*)^2 - 2E^* + E \leq 0$$

que pode ser vista hachurada na Fig. 2. Segue que

$$\frac{c-1}{c}\left[1 - \sqrt{1 - E\,\frac{c}{c-1}}\right] \leq E^* \leq \frac{c-1}{c}\left[1 + \sqrt{1 - E\,\frac{c}{c-1}}\right] \tag{15}$$

mas como o majorante é maior que 1 então juntando (15) e (14), obtém-se:

$$\boxed{\frac{c-1}{c}\left[1 - \sqrt{1 - E\,\frac{c}{c-1}}\right] \leq E^* \leq E} \tag{16}$$

Note que, se $\frac{c-1}{c} \approx 1$ e E é pequeno (<<1) então

$$\boxed{E/2 \leq E^* \leq E} \tag{17}$$

As desigualdades (16) e (17) fornecem, para um número grande de amostras, uma forma de se estimar uma ordem de grandeza da taxa de erro para a regra de decisão de Bayes, isto sendo conseguido a partir do valor experimentalmente obtido para o erro E da regra de decisão do VMP. Esta informação pode, p.ex., ser utilizada para avaliar a capacidade de discriminação entre classes de um conjunto de atributos selecionado, ou seja pode-se utilizar esta informação para seleção ou extração de atributos. Como $E^* \leq E$, caso E resulte razoavelmente pequeno, então isto sugere que as classes estão adequadamente separadas no espaço das variáveis ou atributos selecionados.

REGRA DE DECISÃO DOS k VIZINHOS MAIS PRÓXIMOS (k-NN)

É dado o conjunto de referência $Q_N = \left\{ \underline{x}_1, \ldots, \underline{x}_N \right\}$ contendo amostras já classificadas. Dado um padrão $\underline{x}$ cuja classificação se deseja obter, a

regra k-NN classifica $\underline{x}$ na mesma classe que a classe mais frequente dos k padrões de Q_N mais próximos de $\underline{x}$, onde a proximidade se baseia em alguma medida de distância. Uma forma de se ter alguma intuição sobre essa regra de decisão é lembrando que o classificador de Bayes com mínima taxa de erro para um vetor arbitrário $\underline{x}$ é:

$$\text{decidir } \omega_i \quad \text{se} \quad P(\omega_i|\underline{x}) \geq P(\omega_j|\underline{x}) \qquad i,j = 1,\ldots,c$$

que equivale à desigualdade $P_i\ p(\underline{x}|\omega_i) \geq P_j\ p(\underline{x}|\omega_j)$, contanto que $p(\underline{x}) \neq 0$. Para estimar P_i a partir dos dados, basta tomar N_i/N onde N_i é o número de amostras, dentre as N, que são da classe ω_i. Para estimar $p(\underline{x}|\omega_i)$, pode-se utilizar a conceituação vista na estimação de função densidade pelo método dos vizinhos mais próximos em que se toma um volume V (pequeno), centrado em $\underline{x}$, e conta-se quantas amostras k_i há no seu interior. Desta forma a regra de decisão de Bayes aproximada fica:

$$\text{decidir } \omega_i \text{ se} \quad \frac{N_i}{N} \cdot \frac{k_i}{N_i \cdot V} \geq \frac{N_j}{N} \cdot \frac{k_j}{N_j \cdot V} \qquad j=1,2,\ldots c$$

onde pode-se interpretar o volume V (dependente de $\underline{x}$) como abarcando exatamente k amostras (para facilitar supomos que isto seja sempre possível, ou seja não há amostras equidistantes ao dado $\underline{x}$ que impeçam a exequibilidade de um dado valor de k) indistintamente das classes envolvidas, com $k = \sum_{i=1}^{c} k_i$.

Cancelando os termos iguais, obtemos a regra de decisão k-NN.

A regra k-NN deve ser preferida em relação ao 1-NN quando se dispõe de um número N grande de amostras em Q_N. Um outro caso em que ela deve ter primazia é nos casos em que o conjunto de referência pode apresentar um certo número de amostras com classificação errada.

A taxa média de erro pode ser estimada em termos de uma desigualdade mas a matemática é complicada e a expressão final não é de utilidade especial e portanto não será vista.

Aparentemente não há muito que se possa sugerir para a escolha de

bons valores para k em um caso prático, recomendando-se tentativa e erro, com a utilização de uma estimativa da taxa de classificações erradas. Via de regra a escolha de um k ímpar pode ser melhor para diminuir a ocorrência de empates.

REGRA DE DECISÃO DOS k VIZINHOS MAIS PRÓXIMOS, COM OPÇÃO DE REJEIÇÃO $\left((k,m)\text{-NN}\right)$

É semelhante ao k-NN exceto que é feita uma classificação apenas se o número de vezes em que a classe mais frequente dentre as classes associadas aos k VMPs aparece é maior ou igual a um certo nível m ($m \in \mathbb{Z}^+$). Caso contrário o padrão é rejeitado. Pensando em uma analogia com um processo de votação, o método permitiria a eleição do candidato mais votado, dentre os k vizinhos mais próximos, desde que obtivesse um número de votos maior do que m. Devido ao uso de dois parâmetros, esse método recebe o nome de (k,m) VMPs.

Uma expansão deste método pode ser obtida tomando-se o nível m como dependente da decisão a ser tomada, ou seja, decide-se por $\omega = \omega_i$ se pelo menos m_i ($m_i \in \mathbb{Z}^+$) VMPs dentre os k VMPs a $\underline{x}$ são da classe ω_i. Esta regra seria indicada como (k, m_i)-NN. Pode-se interpretar a regra de pelo menos duas formas: dentre os k vizinhos mais próximos, se apenas uma classe ω_i teve representatividade maior que m_i, então atribui-se a $\underline{x}$ esta classe. Se mais que uma classe teve representatividade maior que seu respectivo limiar então o desempate deve ser feito ao acaso ou utilizando um critério adicional como, por exemplo, o de escolher a classe que mais suplantou (em termos absolutos ou relativos) o respectivo limiar m_i. Alternativamente, pode-se partir da classe que maior representatividade teve dentre os k

vizinhos mais próximos e verificar se sua "votação" ultrapassou o respectivo limiar. Em caso afirmativo, então esta seria a classe atribuida ao $\underline{x}$. Em caso negativo, se passaria para a próxima classe mais votada, repetindo o teste com limiar, e assim por diante. Para maiores detalhes sobre os classificadores por vizinhos mais próximos, sugere-se consultar DeVijver e Kittler (1982).

ABORDAGENS PARA AGILIZAR O MÉTODO DE VIZINHOS MAIS PRÓXIMOS

O método dos VMPs tem a característica de requerer que todas as as amostras ou padrões de Q_N permaneçam armazenadas e que a cada novo padrão $\underline{x}$ se calculem os k VMPs. Quanto maior o número N de amostras, melhor será o desempenho do classificador mas maiores serão também as necessidades de memória e de tempo de processamento. A princípio, para cada $\underline{x}$ deve-se calcular sua distância ou proximidade a cada um dos N padrões de referência. Duas abordagens são possíveis para diminuir o tempo de processamento.

a - Diminuição do número de padrões do conjunto Q_N (com isto também se diminui a necessidade de memória).

b - Utilização de um algoritmo computacional eficiente para a localização do vizinho mais próximo como, p.ex., o algoritmo "branch and bound" e o "k-d tree".

Dentre as duas abordagens, a segunda é absolutamente necessária, a não ser que o número de amostras N seja pequeno. Apresentamos no que segue um exemplo interessante da primeira abordagem, que é a regra de decisão do vizinho mais próximo condensada (CNN).

Parte-se de 2 regiões de memória A e B , com B contendo o conjunto Q_N de padrões. Para iniciar, um padrão é transferido de B para A. Em termos

gerais, cada padrão de B é classificado segundo a regra VMP, usando como conjunto de referência o grupo de padrões existente em A. Se a classificação VMP coincidir com a classificação correta então o padrão é mantido em B , caso contrário é transferido para A. Quando todos os padrões de B tiverem sido analisados, o procedimento é repetido até que não haja mais transferências de B para A. Com isto, obtém-se um conjunto A de padrões classificados "representativo", e em geral com bem menos padrões do que os N originais. Nesse conjunto condensado de padrões, haverá uma tendência a agrupamento de amostras ou padrões nas vizinhanças das fronteiras de Bayes, ou seja, padrões longe da fronteira são descartados com bem maior frequência que os localizados nas vizinhanças das fronteiras.

Com esse método, o que se consegue é um conjunto A de padrões que classifica corretamente por VMP(1-NN) todos os padrões restantes (isto é, do conjunto B). O que o método não pode garantir é que o conjunto A obtido seja mínimo. Uma melhoria seria obtida suplantando a limitação do CNN em não permitir que um elemento uma vez armazenado em A possa vir a ser retirado de A, pois amostras posteriormente adicionadas a A podem tornar alguma(s) amostra(s) de A supérflua(s) pois seria(m) corretamente classificada(s) a partir das outras. Um aperfeiçoamento do que acabamos de apresentar é o método "reduced-nearest-neighbor" que aparentemente dá uma melhoria pequena mas a um custo computacional grande (Hand, 1981). Uma extensão bastante óbvia do que vimos acima é para o caso da regra de decisão dos k VMPs, obtendo-se por método análogo, a regra de decisão dos k vizinhos mais próximos condensada (k-CNN).

Para uma primeira leitura sobre algoritmos que aumentam a velocidade da técnica de classificação por VMP pode-se sugerir o artigo de Fukunaga e Narendra (1975) que apresenta a abordagem "branch and bound", e o artigo de Friedman, Bentley e Finkel (1977) que apresenta a abordagem "k-d tree".

CLASSIFICAÇÃO DE PADRÕES POR MÍNIMA DISTÂNCIA

INTRODUÇÃO

Neste breve capítulo será apresentada a abordagem de classificação por mínima distância, que é equivalente à técnica de classificação por vizinho mais próximo.

CLASSIFICAÇÃO POR MÍNIMA DISTÂNCIA COM UM ÚNICO PROTÓTIPO POR CLASSE

Em certas aplicações, cada classe pode ser caracterizada por um único protótipo. Suponhamos então que há c classes, cada uma caracterizada pelo protótipo $\underline{z}_i$ (i=1, 2,..c) e que se utiliza a distância Euclideana. Dado um padrão $\underline{x}$ a ser classificado, a sua distância Δ_i ao i-ésimo protótipo é

$$\Delta_i = d_2(\underline{x},\underline{z}_i) = \| \underline{x}-\underline{z}_i \|_2 = \sqrt{(\underline{x}-\underline{z}_i)^T (\underline{x}-\underline{z}_i)} \qquad (1)$$

Nestas condições, o classificador por mínima distância atribui o padrão $\underline{x}$ à classe ω_k caso $\Delta_k < \Delta_j$, $j \neq k$; j=1,...c.

Desenvolvendo (1) obtemos:

$$\Delta_i^2 = \| \underline{x}-\underline{z}_i \|_2^2 = \underline{x}^T\underline{x} - 2(\underline{x}^T\underline{z}_i - \frac{1}{2}\ \underline{z}_i^T\ \underline{z}_i) \qquad (2)$$

De (2) vemos que escolher o mínimo de Δ_i^2 , ou de Δ_i , é equivalente a escolher o máximo de $\underline{x}^T\underline{z}_i - \frac{1}{2}\ \underline{z}_i^T\ \underline{z}_i$. Com isto, definimos as funções de decisão

$$d_i(\underline{x}) = \underline{x}^T\underline{z}_i - \frac{1}{2}\ \underline{z}_i^T\ \underline{z}_i \qquad i = 1, 2,...,c$$

que são lineares (em x). Se fizermos

$$\underline{v}_{oi} = \underline{z}_i \qquad e \qquad v_{d+1,i} = -\frac{1}{2}\,\underline{z}_i^T\,\underline{z}_i$$

elas recaem na forma $d_i(\underline{x}) = \underline{v}_{oi}^T\,\underline{x} + v_{d+1,i}$ vista no capítulo sobre Funções de Decisão. A regra de decisão fica:

* classificar $\underline{x}$ em ω_k se $d_k(\underline{x}) > d_j(\underline{x})$, $j \neq k$, $j = 1,\ldots,c$. Este é equivalente ao Caso 3 apresentado no capítulo sobre Funções de Decisão e portanto obtém-se regiões de decisão conexas e convexas. Deve-se lembrar que a fronteira entre as classes i e j é obtida de $d_i(\underline{x})-d_j(\underline{x}) = 0$. Para o caso de classificação por mínima distância com 1 protótipo por classe a equação da fronteira é:

$$\underline{x}^T(\underline{z}_i-\underline{z}_j) - \frac{1}{2}\,(\underline{z}_i^T\underline{z}_i-\underline{z}_j^T\underline{z}_j) = 0$$

que é um hiperplano (no caso geral) mediatriz do segmento de reta que une $\underline{z}_i$ a $\underline{z}_j$. Para ver isto, basta notar que $\Delta_i^2 = \underline{x}^T\underline{x} - 2d_i(\underline{x})$ e, caso $d_i(\underline{x}) = d_j(\underline{x})$ resulta $\Delta_i^2 = \Delta_j^2$, e portanto, tem-se o lugar geométrico dos pontos equidistantes de $\underline{z}_i$ e $\underline{z}_j$.(para mostrar a ortogonalidade do hiperplano ao segmento $\underline{z}_i-\underline{z}_j$ basta notar de (1) que $\underline{v}_o=\underline{z}_i-\underline{z}_j$, que indica uma direção $\perp$ H).

Com 1 protótipo por classe o critério de classificação por mínima distância (Euclidiana) é equivalente ao classificador por vizinho mais próximo (1-NN) com proximidade definida pela distância Euclideana e com um conjunto de referência contendo apenas uma amostra por classe. Conclui-se que o classificador resultante pode ser escrito na forma de funções de decisão lineares. Obtém-se separatrizes lineares entre regões de decisão contigüas. No caso de 3 ou mais classes as regiões de decisão são definidas por trechos de hiperplanos.

A Fig. 1 apresenta um exemplo em que há 4 classes em um espaço de atributos bi-dimensional.

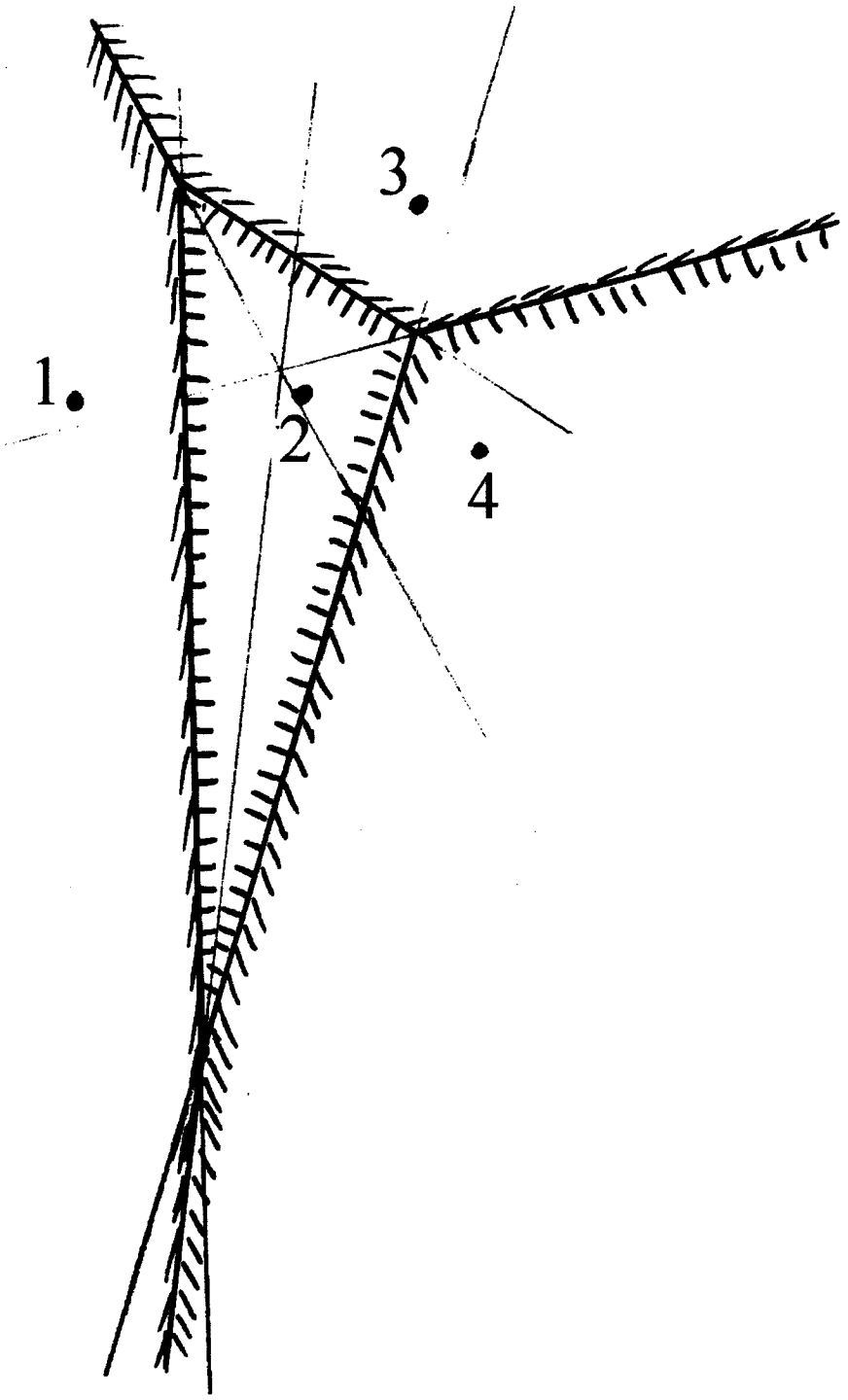


Fig. 1 – Exemplo de classificador por mínima distância em que cada uma das quatro classes é representada por um único protótipo.

CLASSIFICAÇÃO POR MÍNIMA DISTÂNCIA COM VÁRIOS PROTÓTIPOS POR CLASSE

Cada classe ω_i é representada por vários protótipos $\underline{z}_{i1}$, $\underline{z}_{i2}$,... $\underline{z}_{ik}$..., $\underline{z}_{in_i}$ onde n_i é o número de protótipos na classe ω_i .

Define-se a distância de uma amostra $\underline{x}$ à classe ω_i como sendo

$$\Delta_i(\underline{x}) = \min_k \| \underline{x} - z_{ik} \|_2$$

$$\text{com} \quad k = 1, 2, \ldots, n_i \quad \text{e} \quad i=1,\ldots,c.$$

Com isto a regra de classificação fica:

* atribuir $\underline{x}$ à classe ω_i se

$$\Delta_i(\underline{x}) < \Delta_j(\underline{x}) \quad \text{para todo} \quad j \neq i \ , \quad j = 1,\ldots,c$$

$$\text{onde:} \quad \Delta_i^2(\underline{x}) = \min_k \left(\underline{x}^T\underline{x} - 2(\underline{x}^T\underline{z}_{ik} - \frac{1}{2}\ \underline{z}_{ik}^T\ \underline{z}_{ik}) \right)$$

Observando-se que $\Delta_i \geq 0$ pode-se trabalhar com Δ_i^2 que é mais conveniente.

Define-se as funções de decisão como

$$d_i(\underline{x}) = \max_k \ (\underline{x}^T\underline{z}_{ik} - \frac{1}{2}\ \underline{z}_{ik}^T\underline{z}_{ik})$$

com $k = 1,2,\ldots, n_i$ e $i = 1,\ldots,c.$

Nesta formalização, a regra de classificação fica

* atribuir $\underline{x}$ à classe ω_i se

$$d_i(\underline{x}) > d_j(\underline{x}) \quad \text{para todo } j \neq i \ , \ j = 1,\ldots c$$

No caso de haver 2 classes com vários protótipos por classe, a fronteira não é um hiperplano como em um classificador linear. Ela é formada por trechos de hiperplanos (Fig. 2), resultando facilmente em regiões de decisão não convexas e desconexas (ou não conexas) para uma mesma classe (p.ex., Fig. 2a). Diz-se que se tem um classificador ou discriminante linear por trechos ("piecewise linear"). Pela própria definição desse classifica-

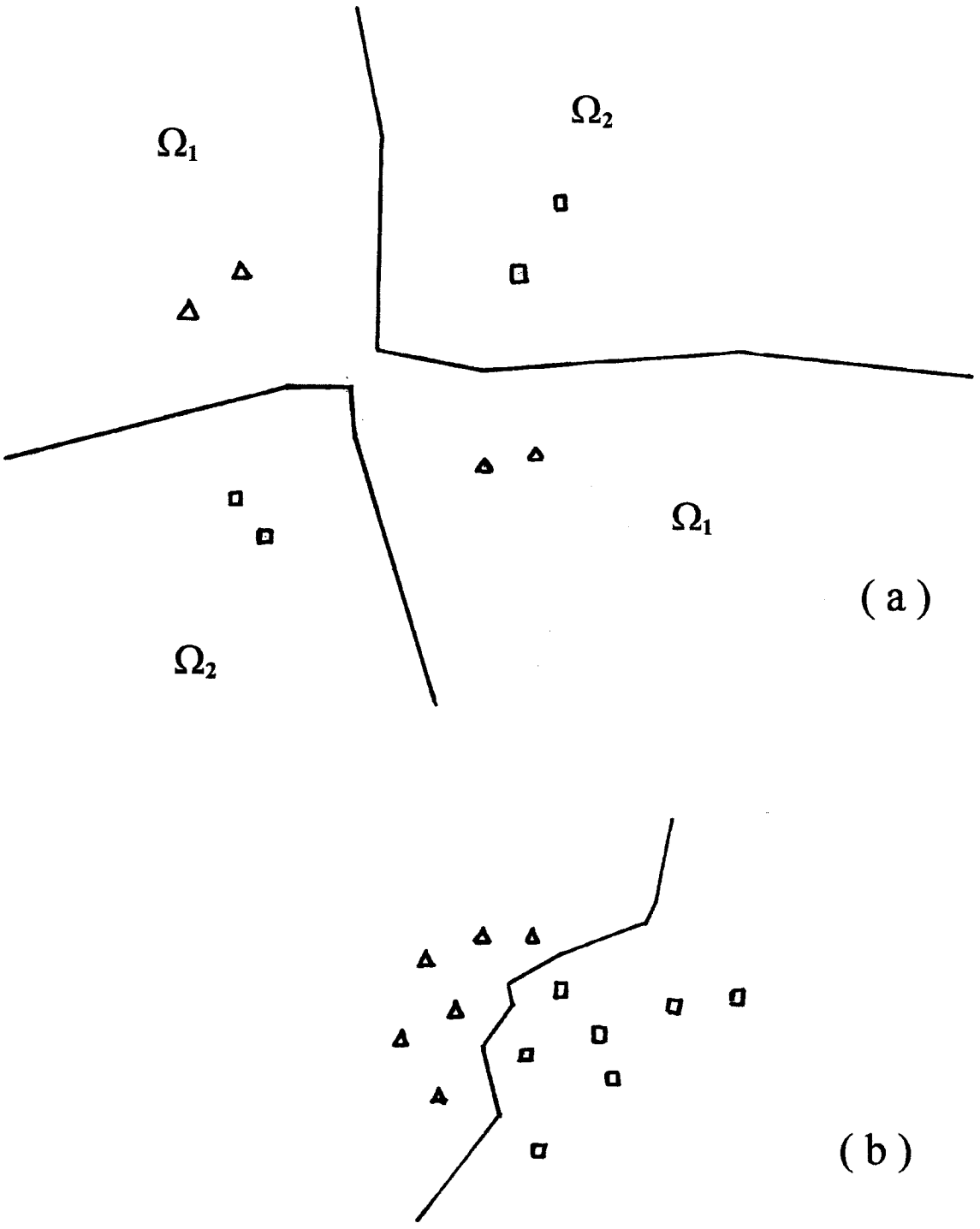


Fig. 2 – Exemplo de classificador por mínima distância em que cada uma das duas classes é representada por mais que 1 protótipo. No caso (a) resulta uma região de decisão desconexa para a classe ω_2, ao contrário de (b).

dor, ou pela sua construção geométrica, fica claro que excetuando-se as fronteiras, não há regiões de classificação indeterminada. As regiões de decisão podem ser interpretadas como a união de regiões convexas para o caso em que tomássemos cada amostra como sendo uma classe distinta. A união de regiões convexas adjacentes (i.e., que tem uma superfície de contacto) não é necessariamente convexa.

Dentro da filosofia do classificador de mínima distância é possível trabalhar com o classificador linear por trechos principalmente se o número de protótipos por classe é pequeno. Entretanto não há para os classificadores lineares por trechos algoritmos gerais para determinação das funções de decisão como existe para os classificadores ou máquinas lineares.

Deve-se notar que o classificador por mínima distância, para o caso de múltiplos protótipos por classe, é equivalente ao método do vizinho mais próximo tomado com a distância Euclideana.

DISCRIMINANTE E CLASSIFICADOR LINEAR DE FISHER

INTRODUÇÃO

Esta abordagem se originou nos trabalhos de R.A. Fisher, desenvolvidos entre 1930 e 1940. Ela se aplica ao problema de separar ou discriminar classes representadas quer por seus vetores médios e matrizes de covariância, quanto por amostras pré-classificadas. Adicionando-se um ponto de limiar sobre a reta correspondente à direção da projeção, obtém-se um classificador. Desta forma, a abordagem de Fisher corresponde a uma técnica de obtenção de funções de decisão lineares segundo um critério de otimalidade (baseado em estatísticas de primeira e segunda ordem). Sob uma outra perspectiva, a abordagem de Fisher provê uma extração de atributos, reduzindo a dimensionalidade do problema de classificação.

O CASO DE DUAS CLASSES OU POPULAÇÕES

São dados dois conjuntos de amostras ou vetores de atributos, um com elementos provenientes da classe ω_1 e o outro de ω_2 . Estes dois conjuntos são utilizados para se determinar uma função discriminante linear tal que as projeções dos elementos de ω_i(i=1,2) fiquem o mais próximo possível entre si e as projeções dos elementos de ω_1 fiquem o mais distante possível dos elementos de ω_2 . A função discriminante obtida pode então ser utilizada para classificar novas amostras ou vetores de atributos.

A abordagem de Fisher é de reduzir a dimensão do vetor de atributos para dimensão 1 efetuando um mapeamento $\mathbb{R}^d \rightarrow \mathbb{R}$ conveniente. Uma vez em $\mathbb{R}$,

passa-se ao problema de escolher um limiar para separar ω_1 de ω_2. No caso mais amplo em que pelo menos uma das classes se caracteriza por ocupar regiões não conexas no espaço de atributos (distribuição multi-modal), poderia ser necessário determinar mais do que um limiar para separar ω_1 de ω_2.

Definimos $\underline{X}|\omega_1$ como os vetores aleatórios provenientes da classe ω_1 e $\underline{X}|\omega_2$ da classe ω_2 . Indicamos os vetores médios condicionados às classes ω_1 e ω_2 por $\underline{\mu}_1 = E[\underline{X}|\omega_1]$ e $\underline{\mu}_2 = E[\underline{X}|\omega_2]$ e as matrizes de covariância condicionais por $\Sigma_i = E[(\underline{X}-\underline{\mu}_i)(\underline{X}-\underline{\mu}_i)^T|\omega_i]$; i=1, 2 .

O vetor aleatório $\underline{X}$ pode tomar valores de $\underline{X}|\omega_1$ ou de $\underline{X}|\omega_2$ com respectivas probabilidades P_1 e P_2 . Temos portanto

$$p_{\underline{X}}(\underline{\alpha}) = p_{\underline{X},\omega_1}(\underline{\alpha}, \omega_1) + p_{\underline{X},\omega_2}(\underline{\alpha}, \omega_2) \tag{1}$$

ou

$$p_{\underline{X}}(\underline{\alpha}) = p_{\underline{X}|\omega_1}(\underline{\alpha}|\omega_1)\cdot P_1 + p_{\underline{X}|\omega_2}(\underline{\alpha}|\omega_2)\cdot P_2 \tag{2}$$

De (2) segue:

$$E[\underline{X}] = \underline{\mu}_x = \underline{\mu}_1 P_1 + \underline{\mu}_2 P_2 \qquad \text{e} \tag{3}$$

$$\Sigma_X \triangleq E\left[(\underline{X}-\underline{\mu}_x)(\underline{X}-\underline{\mu}_x)^T\right] = E\left[\underline{X}\,\underline{X}^T|\omega_1\right]P_1 + E\left[\underline{X}\,\underline{X}^T|\omega_2\right]P_2 - \underline{\mu}_x\,\underline{\mu}_x^T$$

e portanto

$$\Sigma_x = (\Sigma_1+\underline{\mu}_1\underline{\mu}_1^T)P_1+(\Sigma_2+\underline{\mu}_2\underline{\mu}_2^T)P_2 - P_1^2\underline{\mu}_1\underline{\mu}_1^T - P_2^2\underline{\mu}_2\underline{\mu}_2^T - P_1P_2(\underline{\mu}_1\underline{\mu}_2^T+\underline{\mu}_2\underline{\mu}_1^T) \tag{4}$$

Como se percebe, a expressão (4) não nos conduz a uma expressão simples mesmo se $\Sigma_1 = \Sigma_2 = \Sigma$ e se $P_1 = P_2 = 1/2$.

O vetor $\underline{X}$ será projetado sobre uma direção conveniente fornecendo a variável aleatória Y :

$$Y = \underline{w}^T\,\underline{X} \tag{5}$$

com $\underline{w} = [w_1\ w_2\ \ldots . w_d]^T$. Escolhemos aqui a notação $\underline{w}$ ao invés de $\underline{v}_0$ como no

capítulo anterior apenas pela maior simplicidade de escrita.

Definimos $\underline{\mu}_{1y} = E\,[Y|\omega_1]$ e $\underline{\mu}_{2y} = E\,[Y|\omega_2]$ e portanto

$$\mu_{1y} = \underline{w}^T \underline{\mu}_1 \tag{6}$$

$$\mu_{2y} = \underline{w}^T \underline{\mu}_2 \tag{7}$$

Se calcularmos σ_y^2 ,veremos que mesmo para $\Sigma_1 = \Sigma_2$ e $P_1 = P_2 = 1/2$ resulta uma expressão complicada. Trabalharemos entretanto com $\sigma_{iy}^2 \triangleq E\left[(Y-\mu_{iy})^2|\omega_i\right]$ obtendo

$$\sigma_{iy}^2 = E\,[\,\underline{w}^T(\underline{X} - \underline{\mu}_i)(\underline{X} - \underline{\mu}_i)^T \underline{w} \;|\omega_i] = \underline{w}^T \Sigma_i \underline{w} \quad ; \quad i = 1,\ 2 \tag{8}$$

que para o caso de 2 classes com a mesma matriz de covariância $\Sigma_1 = \Sigma_2 = \Sigma$, fica $\sigma_{iy}^2 = \underline{w}^T \Sigma \underline{w}$.

Fisher propôs maximizar a função critério $J(\underline{w})$, supondo que as 2 classes têm a mesma matriz de covariância

$$J(\underline{w}) \triangleq \frac{(\mu_{1y}-\mu_{2y})^2}{\sigma_{iy}^2} = \frac{(\underline{w}^T\underline{\mu}_1 - \underline{w}^T\underline{\mu}_2)^2}{\underline{w}^T\Sigma\underline{w}} = \frac{[\underline{w}^T(\underline{\mu}_1 - \underline{\mu}_2)]^2}{\underline{w}^T\Sigma\underline{w}} =$$

$$= \frac{\underline{w}^T(\underline{\mu}_1-\underline{\mu}_2)(\underline{\mu}_1-\underline{\mu}_2)^T\underline{w}}{\underline{w}^T\Sigma\underline{w}} \quad ; \quad i=1 \text{ ou } 2 \tag{9}$$

obs: o índice i é para não haver confusão com s_y^2 que é a variância total de y para se determinar o $\underline{w}$ ótimo. Busca-se com isto uma direção tal que as projeções das médias das 2 classes sejam as mais distantes possíveis, ao mesmo tempo em que se minimiza a dispersão dentro de cada classe projetada (esta é, por hipótese, a mesma nas duas classes).

O máximo de (9) é obtido para

$$\underline{w}^* = c\,\Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_2) \tag{10}$$

para qualquer $c \neq 0$, sendo que c=1 é o mais conveniente. Resulta neste caso a função discriminante linear de Fisher :

$$Y = \underline{w}^{*T}\underline{X} = (\underline{\mu}_1 - \underline{\mu}_2)^T\,\Sigma^{-1}\,\underline{X} \tag{11}$$

Para demonstrar (10) necessitamos antes demonstrar 3 Lemas.

Lema 1 (Desigualdade de Cauchy-Schwarz)

Dados $\underline{a}$ e $\underline{b}$ dois vetores arbitrários de dimensão d x 1. Vale então

$$(\underline{a}^T\underline{b})^2 \leq (\underline{a}^T\underline{a})(\underline{b}^T\underline{b}) \tag{12}$$

com igualdade se e só se $\underline{a} = c\underline{b}$, para $c \in \mathbb{R}$.

Demonstração

Temos para α arbitrário $\in \mathbb{R}$, excluindo $\underline{a} = \underline{0}$ e/ou $\underline{b} = \underline{0}$ que claramente satisfazem (12),

$$(\underline{a} - \alpha\,\underline{b})^T(\underline{a} - \alpha\,\underline{b}) \geq 0$$

com igualdade somente para $\underline{a} = \alpha\,\underline{b}$. Portanto, $\underline{a}^T\underline{a} - 2\alpha\,\underline{a}^T\underline{b} + \alpha^2\,\underline{b}^T\underline{b} \geq 0$, que é uma expressão quadrática em α. Completamos o quadrado somando e subtraindo o escalar $(\underline{a}^T\underline{b})^2/\underline{b}^T\underline{b}$, obtendo

$$\underline{a}^T\underline{a} - \frac{(\underline{a}^T\underline{b})^2}{\underline{b}^T\underline{b}} + \frac{(\underline{a}^T\underline{b})^2}{\underline{b}^T\underline{b}} - 2\alpha(\underline{a}^T\underline{b}) + \alpha^2\,\underline{b}^T\underline{b} \geq 0$$

$$\therefore \quad \underline{a}^T\underline{a} - \frac{(\underline{a}^T\underline{b})^2}{\underline{b}^T\underline{b}} + \underbrace{(\underline{b}^T\underline{b})}_{>0}\ \underbrace{\left[\alpha - \frac{\underline{a}^T\underline{b}}{\underline{b}^T\underline{b}}\right]^2}_{\geq 0} \geq 0$$

como a desigualdade deve valer para todo α , tomemos $\alpha = \underline{a}^T\underline{b}/\underline{b}^T\underline{b}$

$$\therefore \quad (\underline{a}^T\underline{a})(\underline{b}^T\underline{b}) \geq (\underline{a}^T\underline{b})^2$$

com igualdade se e somente se $\underline{a} = c\underline{b}$ $(c \in \mathbb{R})$

Lema 2 (Desigualdade de Cauchy-Schwarz estendida)

São dados os vetores $\underline{a}$ e $\underline{b}$, arbitrários, de dimensão dx1, e seja Ψ uma matriz d x d simétrica positiva definida. Então vale:

$$(\underline{a}^T\underline{b})^2 \leq (\underline{a}^T\,\Psi\,\underline{a})(\underline{b}^T\,\Psi^{-1}\,\underline{b}) \tag{13}$$

com igualdade se e só se $\underline{a} = c\,\Psi^{-1}\,\underline{b}$ (ou $\underline{b} = c\,\Psi\,\underline{a}$), para algum $c \in \mathbb{R}$.

Demonstração

Para $\underline{a} = \underline{0}$ e/ou $\underline{b} = \underline{0}$ a expressão (13) é válida, e, portanto, a partir daqui, supomos $\underline{a} \neq \underline{0}$ e $\underline{b} \neq \underline{0}$. Como Ψ é uma matriz simétrica e portanto diagonalizável(e.g.,Searle,1982,pg290), então $\Psi = P\ \Psi_{diag} P^T$, onde P é a matriz d x d cujas colunas são os autovetores correspondentes aos autovalores λ_i de Ψ que compõe a matriz diagonal Ψ_{diag} . Define-se $\Psi^{1/2}$ como $\Psi^{1/2} \triangleq P\ \Psi_{diag}^{1/2} P^T$ onde $\Psi_{diag}^{1/2} = diag\left(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots \sqrt{\lambda_d}\right)$ com as propriedades que $\Psi^{1/2}$ é uma matriz simétrica, $\Psi^{1/2}\Psi^{1/2} = \Psi$; $\Psi^{1/2}.\Psi^{-1/2} = \Psi^{-1/2}\ \Psi^{1/2} = I$ com $\Psi^{-1/2} \triangleq (\Psi^{1/2})^{-1}$.

Do já visto segue que:

$$\underline{a}^T\underline{b} = \underline{a}^T\ \Psi^{1/2}\Psi^{-1/2}\underline{b} = (\Psi^{1/2}\underline{a})^T(\Psi^{-1/2}\underline{b})$$

e aplicando a desigualdade de Cauchy-Schwarz ao 2º membro temos:

$$(\underline{a}^T\underline{b})^2 \leq (\underline{a}^T\ \Psi\ \underline{a})(\underline{b}^T\ \Psi^{-1}\ \underline{b})$$

com igualdade se e só se $\underline{a} = c\ \Psi^{-1}\underline{b}$ ou $\underline{b} = c\ \Psi\ \underline{a}$, $c \in \mathbb{R}$.

Lema 3 (Lema da maximização)

São dados: Ψ uma matriz simétrica positiva definida e $\underline{b}$ um vetor dx1.

Então :

$$\max_{\underline{w} \neq \underline{0}} \frac{(\underline{w}^T\underline{b})^2}{\underline{w}^T\Psi\ \underline{w}} = \underline{b}^T\Psi^{-1}\underline{b} \qquad (14)$$

que ocorre para

$$\underline{w} = \underline{w}^* = c\ \Psi^{-1}\ \underline{b}\ , \quad \forall\ c \neq 0 \qquad (15)$$

Demonstração Da desigualdade de Cauchy-Schwarz estendida tem-se

$$(\underline{w}^T\underline{b})^2 \leq (\underline{w}^T\Psi\ \underline{w})(\underline{b}^T\Psi^{-1}\underline{b})$$

e como $\underline{w} \neq \underline{0}$ e Ψ é positivo definido segue que $\underline{w}^T\Psi\ \underline{w} > 0$ e pode-se, portanto, dividir tudo por $\underline{w}^T\Psi\ \underline{w}$ resultando

$$\frac{(\underline{w}^T\underline{b})^2}{\underline{w}^T\Psi\underline{w}} \leq \underline{b}^T\Psi^{-1}\underline{b}$$

com igualdade se e só se $\underline{w} = \underline{w}^* = c\ \Psi^{-1}\underline{b}$, $c \in \mathbb{R}$ c.q.d.

Agora fica imediato provar a expressão (10), bastando aplicar-se o

Lema 3 no contexto da expressão (9). Desta forma a projeção $Y = \underline{w}^{*T}\underline{X}$ tem a propriedade das médias projetadas das 2 classes estarem o mais afastadas possível e, ao mesmo tempo, o espalhamento de cada classe projetada ser o menor possível. Via de regra, esta técnica fornece melhores resultados para distribuições $p(\underline{x}|\omega_i)$ unimodais, conforme ilustrado na Fig. 1 para c=d=2 .

Determinação de limiar

Uma vez em dimensão 1, pode-se determinar limiares para classificação utilizando-se, por exemplo, a regra de Bayes no caso das distribuições de $Y|\omega_i$ serem conhecidas (além de P_i). Um discriminante linear de Fisher acrescido de um limiar (ou mais) se torna um classificador, que podemos chamar de classificador de Fisher. Como na prática pouco se conhece das distribuições, deve-se utilizar critérios mais simples para a determinação de limiar. Analisaremos separadamente dois casos em que se supõe que as duas distribuições são unimodais: i caso de matrizes de covariância iguais ($\Sigma_1=\Sigma_2=\Sigma$). Pode-se definir um ponto de limiar y_L, sobre a reta ótima de Fisher, que separa duas distribuições unimodais, como sendo

$$y_L = \underline{w}^{*T}(\underline{\mu}_1 P_2 + \underline{\mu}_2 P_1) = P_2\mu_{1y} + P_1\mu_{2y} \qquad (16)$$

Deve-se notar que a probabilidade P_2 multiplica μ_{1y} e não a probabilidade P_1, pois se P_2 for maior que P_1, o limiar y_L deve ficar mais distante de μ_{2y}, ou seja, mais próximo de μ_{1y}. No caso de classes equiprováveis, temos o valor do limiar $y_L=(\mu_{1y}+\mu_{2y})/2$. Quando os valores de P_1, P_2, μ_{1y} e μ_{2y} forem desconhecidos, pode-se estimá-los a partir dos dados.

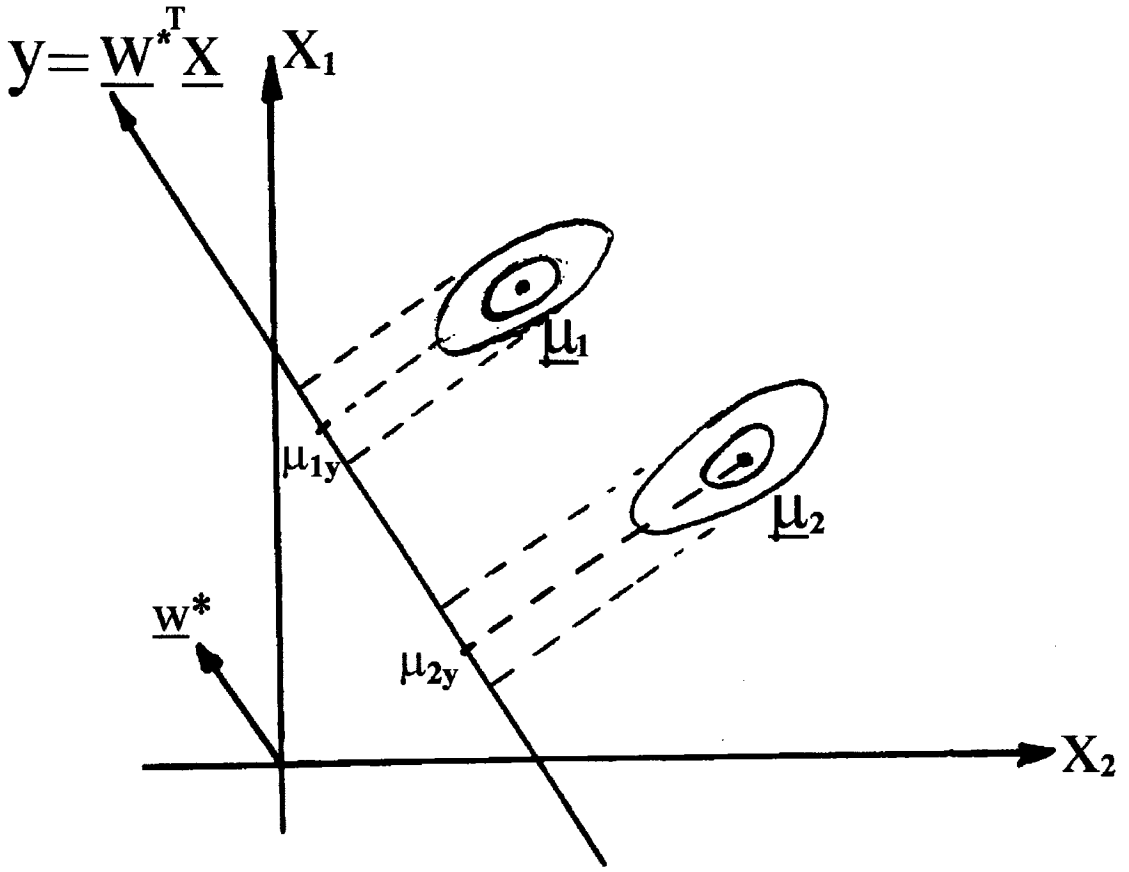


Fig. 1 – Exemplo do discriminante de Fisher para o caso bidimensional de distribuições unimodais com a mesma matriz de covariância.

Utilizando a projeção definida em (11), tem-se, para o caso de y_L dado por (16) a relação

$$\mu_{2y} < y_L < \mu_{1y} \tag{17}$$

pois, como

$$\mu_{1y}=E[Y|\omega_1] = (\underline{\mu}_1-\underline{\mu}_2)^T \Sigma^{-1} \underline{\mu}_1 \quad e \quad \mu_{2y}=E[Y|\omega_2] = (\underline{\mu}_1-\underline{\mu}_2)^T \Sigma^{-1} \underline{\mu}_2 \tag{18}$$

temos que $\mu_{1y}-\mu_{2y} = (\underline{\mu}_1-\underline{\mu}_2)^T\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2)$, que por ser uma forma quadrática positiva definida (lembrando que Σ é uma matriz de covariância) mostra que $\mu_{1y} > \mu_{2y}$. Como y_L é uma combinação linear de μ_{1y} e μ_{2y}, com parâmetros não negativos cuja soma é 1, fica demostrado o que queríamos.

ii caso de matrizes de covariância diferentes ($\Sigma_1 \neq \Sigma_2$). Pode-se definir uma matriz de covariância global como sendo $\Gamma = P_1\Sigma_1+P_2\Sigma_2$, passando-se a maximizar $[\underline{w}^T(\underline{\mu}_1-\underline{\mu}_2)]^2/\underline{w}^T\Gamma\underline{w}$, obtendo-se $\underline{w}^* = \Gamma^{-1}(\underline{\mu}_1-\underline{\mu}_2)$. O valor esperado da classe ω_i (i=1,2) projetado na direção $\underline{w}^*$ é dado por $\mu_{iy} = \underline{w}^{*T}\underline{\mu}_i$, ao passo que a variância σ^2_{iy} nessa mesma direção é dada por $\underline{w}^{*T}\Sigma_i\underline{w}^*$. Um ponto de limiar y_L, na direção $\underline{w}^*$, para separar duas distribuições unimodais, pode ser definido como

$$y_L = \frac{2P_2\mu_{1y}/\sigma_{1y} + 2P_1\mu_{2y}/\sigma_{2y}}{1/\sigma_{1y} + 1/\sigma_{2y}} \tag{19}$$

Quanto maior a probabilidade de uma dada classe ocorrer, mais será puxado o limiar para perto da média da outra classe (projetada sobre $\underline{w}^*$). Quanto menor o espalhamento de uma dada classe, mais perto da média desta (projetada sobre $\underline{w}^*$) ficará o limiar. Similarmente ao já demonstrado para o caso de matrizes de covariância iguais, tem-se que $\mu_{2y}<y_L<\mu_{1y}$. Deve-se observar que a expressão (19) quando as variâncias projetadas são iguais se reduz à expressão (16).

Sintetizando o que foi visto, temos:

⇒ de posse do conhecimento de $\underline{\mu}_1$, $\underline{\mu}_2$ e Σ (= $\Sigma_1=\Sigma_2$) ou Γ ($=P_1\Sigma_1+P_2\Sigma_2$), obtemos a direção ótima para projetar um vetor de atributos;

⇒ dado um vetor $\underline{x}_o$ cuja classificação é desejada, calcula-se $y_o = \underline{w}^{*T}\underline{x}_o$ e

escolhe-se a classe ω_1 se $y_o \geq y_L$, com y_L dado por (16) ou (19), e a classe ω_2 em caso contrário.

Discriminante de Fisher para o caso amostral

Na prática, raramente se conhecem $\underline{\mu}_1$, $\underline{\mu}_2$, Σ_1 e Σ_2 e portanto deve-se estimá-los a partir de um conjunto de amostras provenientes de ambas as classes.

Suponha que se dispõe de n_1 amostras $\underline{x}_{11}$, $\underline{x}_{12}$,... $\underline{x}_{1n_1}$ da classe ω_1 e de n_2 amostras $\underline{x}_{21}$, $\underline{x}_{22}$... $\underline{x}_{2n_2}$ da classe ω_2 , com $n_1+n_2=N$. Com isto podemos estimar o que necessitamos

$$\overline{\underline{x}}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} \underline{x}_{1i} \tag{20}$$

$$\overline{\underline{x}}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} \underline{x}_{2i} \tag{21}$$

$$\underset{(dxd)}{S_1} = \frac{1}{n_1-1} \sum_{i=1}^{n_1} (\underline{x}_{1i}-\overline{\underline{x}}_1)(\underline{x}_{1i}-\overline{\underline{x}}_1)^T \tag{22}$$

$$\underset{(dxd)}{S_2} = \frac{1}{n_2-1} \sum_{i=1}^{n_2} (\underline{x}_{2i}-\overline{\underline{x}}_2)(\underline{x}_{2i}-\overline{\underline{x}}_2)^T \tag{23}$$

onde (20) e (21) estimam $\underline{\mu}_1$ e $\underline{\mu}_2$, respectivamente, e (22) e (23) estimam Σ_1 e Σ_2, respectivamente. Todos estes estimadores são não-viciados. No caso de supormos que as duas classes têm a mesma matriz de covariância podemos obter um estimador global para $\Sigma = \Sigma_1 = \Sigma_2$:

$$S_{fusão} = \frac{\sum_{i=1}^{n_1} (\underline{x}_{1i} - \overline{\underline{x}}_1)(\underline{x}_{1i} - \overline{\underline{x}}_1)^T + \sum_{i=1}^{n_2} (\underline{x}_{2i} - \overline{\underline{x}}_2)(\underline{x}_{2i} - \overline{\underline{x}}_2)^T}{n_1+n_2-2} \tag{24}$$

ou

$$S_{fusão} = \frac{(n_1-1)S_1 + (n_2-1)S_2}{n_1+n_2-2} \qquad (25)$$

onde (24) ou (25) são estimadores não viciados de $\sum$ se as amostras de ω_1 e ω_2 forem todas independentes. Note que, mesmo supondo $\Sigma_1 \neq \Sigma_2$, $S_{fusão}$ é útil, pois é um estimador de $\Gamma = P_1 \Sigma_1 + P_2 \Sigma_2$. Se desejar, pode-se usar, neste caso, $S = \frac{n_1 S_1}{N} + \frac{n_2 S_2}{N}$ ao invés de $S_{fusão}$.

Neste contexto amostral a função discriminante linear de Fisher é $y = \underline{w}^{*T}\underline{x}$, com $\underline{w}^*$ obtido da maximização de

$$J(\underline{w}) = \frac{(\underline{w}^T(\bar{\underline{x}}_1 - \bar{\underline{x}}_2))^2}{\underline{\underline{w}}^T S_{fusão} \underline{\underline{w}}} = \frac{(\bar{y}_1 - \bar{y}_2)^2}{s^2_{iy}} \qquad (26)$$

onde s^2_{iy} é uma variância intra-classe amostral de y.

O mesmo Lema 3 é utilizado para maximizar (26), obtendo-se

$$\underline{w}^* = S^{-1}_{fusão} (\bar{\underline{x}}_1 - \bar{\underline{x}}_2) \qquad (27)$$

e a função discriminante linear de Fisher resulta.

$$y = (\bar{\underline{x}}_1 - \bar{\underline{x}}_2)^T S^{-1}_{fusão} \underline{x} \qquad (28)$$

Deve-se ter $n_1+n_2-2 > d$, pois senão $S_{fusão}$ é singular (prove como exercício). A média para a classe ω_1 projetada sobre a reta discriminante é $\bar{y}_1 = \underline{w}^{*T} \cdot \bar{\underline{x}}_1$. O ponto de limiar y_L tomado como $\frac{n_2}{N} \bar{y}_1 + \frac{n_1}{N} \bar{y}_2$, para o caso em que se possa supor $\Sigma_1 = \Sigma_2$ fica

$$y_L = (\bar{\underline{x}}_1 - \bar{\underline{x}}_2)^T S^{-1}_{fusão} \left(\frac{n_2}{N} \bar{\underline{x}}_1 + \frac{n_1}{N} \bar{\underline{x}}_2 \right) \qquad (29a)$$

Caso não se possa supor $\Sigma_1 = \Sigma_2$, pode-se adotar

$$y_L = \frac{2\frac{n_2}{N}\sqrt{\underline{w}^{*T} S_2 \underline{w}^*}\ \underline{w}^{*T} \bar{\underline{x}}_1 + 2\frac{n_1}{N}\sqrt{\underline{w}^{*T} S_1 \underline{w}^*}\ \underline{w}^{*T} \bar{\underline{x}}_2}{\sqrt{\underline{w}^{*T} S_1 \underline{w}^*} + \sqrt{\underline{w}^{*T} S_2 \underline{w}^*}} \qquad (29b)$$

que é uma expressão análoga à apresentada em (19) para o caso populacional.

Dada uma amostra $\underline{x}_o$ por classificar, escolhemos a classificação em ω_1 se

$$y_o = (\overline{\underline{x}}_1 - \overline{\underline{x}}_2)^T S_{fusão}^{-1} \underline{x}_o \geq y_L \qquad (30)$$

ou em ω_2 em caso contrário.

Se tomarmos $\underline{v}_o \triangleq S_{fusão}^{-1}(\overline{\underline{x}}_1 - \overline{\underline{x}}_2)$ e $v_{d+1} = -y_L$ vemos que a regra de classificação fica $d(\underline{x}) \geq 0$ para ω_1 e $d(\underline{x}) < 0$ para ω_2 onde $d(\underline{x})$ é uma função de decisão linear $d(\underline{x}) = \underline{v}_o^T \underline{x} + v_{d+1}$. Deve-se, entretanto, enfatizar que, em geral, v_{d+1} é determinado heuristicamente, ao passo que $\underline{v}_o$ obedece a um critério de otimalidade.

Formulação por matrizes de somas de quadrados e produtos

Em alguns contextos dentro da área de reconhecimento de padrões, é bastante vantajoso trabalhar com matrizes de somas de quadrados e produtos ("matrizes SQP"). Três matrizes, de dimensão dxd, são de importância, que serão apresentadas para número arbitrário de classes c:

i a matriz de SQP total:

$$T = \sum_{k=1}^{c} \sum_{i=1}^{n_k} (\underline{x}_{ki} - \overline{\underline{x}})(\underline{x}_{ki} - \overline{\underline{x}})^T \qquad (31)$$

onde c é o número de classes, que no presente contexto seria igual a 2, e $\underline{x}_{ki}$, i=1,2,...,n_k são os vetores da classe ω_k. Adicionalmente temos:

$\overline{\underline{x}}_k = \frac{1}{n_k} \sum_{j=1}^{n_k} \underline{x}_{kj}$ (média da classe ω_k)

$\overline{\underline{x}} = \frac{1}{c} \sum_{k=1}^{c} \overline{\underline{x}}_k$ (média global)

$N = \sum_{k=1}^{c} n_k$ (número total de amostras ou padrões, com n_k o número de padrões da classe ω_k)

ii a matriz de SQP intra-classes ("within groups")

$$W = \sum_{k=1}^{c} \sum_{i=1}^{n_k} (\underline{x}_{ki} - \underline{\bar{x}}_k)(\underline{x}_{ki} - \underline{\bar{x}}_k)^T \qquad (32)$$

iii a matriz de SQP entre-classes ("between groups")

$$B = \sum_{k=1}^{c} \sum_{i=1}^{n_k} (\underline{\bar{x}}_k - \underline{\bar{x}})(\underline{\bar{x}}_k - \underline{\bar{x}})^T = \sum_{k=1}^{c} n_k (\underline{\bar{x}}_k - \underline{\bar{x}})(\underline{\bar{x}}_k - \underline{\bar{x}})^T \qquad (33)$$

Em certas aplicações, pode ser interessante utilizar uma versão modificada de **W**, onde se divide o valor da segunda somatória por n_k, desta forma deixando de dar maior importância às classes com mais amostras. Com esta modificação, teríamos em **W** uma estimativa de uma matriz de covariância intra-classes global (equivalente) ou média.

Teorema : Vale a relação

$$T = B + W \qquad (34)$$

Demonstração : Basta tomar

$$T = \sum_{k=1}^{c} \sum_{i=1}^{n_k} \left[(\underline{\bar{x}}_k - \underline{\bar{x}}) + (\underline{x}_{ki} - \underline{\bar{x}}_k) \right] \left[(\underline{\bar{x}}_k - \underline{\bar{x}}) + (\underline{x}_{ki} - \underline{\bar{x}}_k) \right]^T$$

e notar que $\sum_{k=1}^{c} \sum_{i=1}^{n_k} (\underline{\bar{x}}_k - \underline{\bar{x}})(\underline{x}_{ki} - \underline{\bar{x}}_k)^T = 0$ e

$$\sum_{k=1}^{c} \sum_{i=1}^{n_k} (\underline{x}_{ki} - \underline{\bar{x}}_k)(\underline{\bar{x}}_k - \underline{\bar{x}})^T = 0 ,$$

para, de (32) e (33), concluir que (34) é verdadeira. Notar que o teorema foi provado para o caso geral de c classes.

Teorema : Para 2 classes (c=2) vale a relação

$$B = \frac{n_1 n_2}{N} (\underline{\bar{x}}_1 - \underline{\bar{x}}_2)(\underline{\bar{x}}_1 - \underline{\bar{x}}_2)^T = \frac{n_1 n_2}{N} \underline{\bar{\delta}} \cdot \underline{\bar{\delta}}^T \qquad (35)$$

onde $\underline{\bar{\delta}} \triangleq \underline{\bar{x}}_1 - \underline{\bar{x}}_2$

Demonstração :

$$B = n_1 \overline{\underline{x}}_1 \overline{\underline{x}}_1^T - N \overline{\underline{x}}\ \overline{\underline{x}}^T + n_2 \overline{\underline{x}}_2 \overline{\underline{x}}_2^T$$

mas $\overline{\underline{x}} = \dfrac{n_1 \overline{\underline{x}}_1 + n_2 \overline{\underline{x}}_2}{N}$ e portanto

$$B = n_1 \overline{\underline{x}}_1 \overline{\underline{x}}_1^T - N \frac{(n_1 \overline{\underline{x}}_1 + n_2 \overline{\underline{x}}_2)}{N} \cdot \frac{(n_1 \overline{\underline{x}}_1^T + n_2 \overline{\underline{x}}_2^T)}{N} + n_2 \overline{\underline{x}}_2 \overline{\underline{x}}_2^T$$

$$\therefore \quad B = \frac{n_1 N \overline{\underline{x}}_1 \overline{\underline{x}}_1^T - n_1^2 \overline{\underline{x}}_1 \overline{\underline{x}}_1^T - n_1 n_2 \overline{\underline{x}}_1 \overline{\underline{x}}_2^T - n_1 n_2 \overline{\underline{x}}_2 \overline{\underline{x}}_1^T - n_2^2 \overline{\underline{x}}_2 \overline{\underline{x}}_2 + N n_2^T \overline{\underline{x}}_2 \overline{\underline{x}}_2^T}{N}$$

mas $n_1 N \overline{\underline{x}}_1 \overline{\underline{x}}_1^T = n_1(n_1+n_2) \overline{\underline{x}}_1 \overline{\underline{x}}_1^T = n_1^2 \overline{\underline{x}}_1 \overline{\underline{x}}_1^T + n_1 n_2 \overline{\underline{x}}_1 \overline{\underline{x}}_1^T$ e

$$\therefore B = \frac{n_1 n_2}{N} (\overline{\underline{x}}_1 \overline{\underline{x}}_1^T - \overline{\underline{x}}_1 \overline{\underline{x}}_2^T - \overline{\underline{x}}_2 \overline{\underline{x}}_1^T + \overline{\underline{x}}_2 \overline{\underline{x}}_2^T) = \frac{n_1 n_2}{N} (\overline{\underline{x}}_1 - \overline{\underline{x}}_2)(\overline{\underline{x}}_1 - \overline{\underline{x}}_2)^T$$

c.q.d.

Deve-se notar que B sendo da forma (35), igual a $\underline{\alpha} \cdot \underline{\alpha}^T$, ele terá característica (posto) igual a 1 (tomando $\underline{\alpha} \cdot \underline{\alpha}^T$, basta ver que se dividirmos a i-ésima linha por α_i e multiplicarmos por α_j obtemos a j-ésima linha).

Teorema: A expressão para J ($\underline{w}$) em (26), válida apenas para 2 classes, pode ser expressa na forma equivalente

$$J(\underline{w}) = \gamma \frac{\underline{w}^T B \underline{w}}{\underline{w}^T W \underline{w}} \qquad (36)$$

onde $\gamma \in \mathbb{R}^+$ e B pode ser expresso como em (35).

Demonstração

O numerador em (26) é $\underline{w}^T(\overline{\underline{x}}_1 - \overline{\underline{x}}_2)(\overline{\underline{x}}_1 - \overline{\underline{x}}_2)^T \underline{w}$, e de (35) temos que o numerador de (26) é $\dfrac{N}{n_1 n_2} \underline{w}^T B \underline{w}$. O denominador em (26) é $\underline{w}^T S_{fusao} \underline{w}$, e, portanto, de (22), (23), (25) e (32) temos que $S_{fusao} = W/(N-2)$. Conclui-se que (36) é verdadeira com $\gamma = N(N-2)/n_1 n_2$.

É evidente que a maximização de (36), com qualquer valor de γ, resulta no mesmo $\underline{w}^*$ que a maximização de (26). Não obstante o paralelismo de (36) com (26), como as matrizes B e W são definidas para número arbitrário de classes, passaremos a pensar no caso mais geral de c classes.

Teorema : O máximo de $\underline{w}^T B\underline{w}/\underline{w}^T W\underline{w}$ é obtido quando $\underline{w}$ é o autovetor de $W^{-1}B$ correspondente ao maior autovalor de $W^{-1}B$.

Demonstração:

Como $\underline{w}^T B\underline{w}/\underline{w}^T W\underline{w}$ não depende da escala de $\underline{w}$ mas só da direção de $\underline{w}$, podemos transformar o problema na busca do $\underline{w}=\underline{w}^*$ que maximize $\underline{w}^T B\ \underline{w}$ sujeito a $\underline{w}^T W\underline{w} = 1$. Aplicando o método de multiplicador de Lagrange queremos

$$\max_{\underline{w}} \left(\underline{w}^T B\underline{w} - \lambda(\underline{w}^T W\underline{w}-1)\right)$$

e portanto $$\frac{\partial(\underline{w}^T B\underline{w})}{\partial\ \underline{w}} - \lambda \frac{\partial(\underline{w}^T W\underline{w})}{\partial\ \underline{w}} = \underline{0}$$

$$\therefore \quad B\underline{w}^* - \lambda W\underline{w}^* = \underline{0} \text{ , de onde}$$

$$(B - \lambda W)\underline{w}^* = \underline{0} \tag{37}$$

e supondo W não singular

$$(W^{-1}B - \lambda I)\underline{w}^* = \underline{0} \tag{38}$$

e portanto $\underline{w}^*$ é autovetor de $W^{-1}B$ para λ autovalor de $W^{-1}B$. No caso de 2 classes há só 1 autovalor não nulo mas no caso de c classes fica a dúvida : qual dos autovetores de $W^{-1}B$ é para ser adotado ? De (38) temos

$$W^{-1}B\underline{w}^* = \lambda\ \underline{w}^*$$

ou seja

$$B\underline{w}^* = \lambda\ W\underline{w}^*$$

e pré-multiplicando por $\underline{w}^{*T}$ obtém-se

$$\underline{w}^{*T} B\underline{w}^* = \lambda\ \underline{w}^{*T} W\ \underline{w}^*$$

e como o que se deseja é maximizar $\underline{w}^T B \underline{w}$ sujeito a $\underline{w}^T W \underline{w} = 1$, então conclui-se que λ deve ser o maior autovalor do conjunto de autovalores de $W^{-1}B$.

Deve-se notar que o teorema é geral, válido para qualquer número de classes. No caso de 2 classes, B tem característica 1, e, portanto $W^{-1}B$ tem um único autovalor diferente de zero. Este autovalor é dado por

$$tr(W^{-1}B)=tr(W^{-1}\underline{\delta}\cdot\underline{\delta}^T \frac{n_1 n_2}{N})= \frac{n_1 n_2}{N} \quad \underline{\delta}^T W^{-1} \underline{\delta}$$

e o correspondente autovetor é

$$\underline{e}=W^{-1}\underline{\delta} = W^{-1}(\underline{\bar{x}}_1 - \underline{\bar{x}}_2)$$

pois satisfaz a igualdade $W^{-1}\underline{\delta}\cdot\underline{\delta}^T\underline{e} = \underline{\delta}^T W^{-1}\underline{\delta}\ \underline{e}$. Conforme esperado o autovetor aqui obtido, para o caso de duas classes, é o mesmo que o vetor $\underline{w}^*$ da expressão (27).

Convém ressaltar que, apesar de W^{-1} e B serem matrizes simétricas (dxd), $W^{-1}B$ (dxd) não é, em geral, simétrica ($(W^{-1}B)^T = BW^{-1} \neq W^{-1}B$), e portanto seus autovetores não formam uma base ortogonal.

Para a determinação dos autovalores e autovetores pode ser melhor usar (37) pois não requer a inversão da matriz W.

Uma técnica para a determinação de direções adicionais pode ser encontrada em Sammon (1970) e Foley e Sammon (1975).

CASO DE MÚLTIPLAS CLASSES OU POPULAÇÕES

Partiremos diretamente para a análise amostral visto ser a que interessa na prática. Faremos uso das matrizes de SQP que já foram definidas em (31), (32) e (33) para número arbitrário de classes.

Novamente, se o objetivo é obter uma direção ótima, podemos adotar a mesma função critério já empregada para o caso de duas classes

$$J(\underline{w}) = \frac{\underline{w}^T B \underline{w}}{\underline{w}^T W \underline{w}}$$

e, como foi demonstrado anteriormente, o autovetor $\underline{e}_1$ associado ao maior autovalor λ_1 de $W^{-1}B$ é a direção ótima $\underline{w}_1^*$. Em alguns casos pode ser que a projeção apenas na direção ótima conduza a bons resultados. Entretanto, em outros casos, direções adicionais serão necessárias. Para isto, pode-se adotar diferentes abordagens, duas sendo mencionadas a seguir:

i após determinada a direção ótima dada pelo autovetor $\underline{e}_1$, procurar a segunda direção ótima no sub-espaço ortogonal a $\underline{e}_1$, obtendo-se $\underline{v}_2$, com $\underline{v}_2^T \cdot \underline{e}_1 = 0$. Para se obter a j-ésima direção, basta otimizar a função critério no sub-espaço ortogonal ao sub-espaço das direções ótimas já determinadas ($\underline{e}_1$, $\underline{v}_2$,....,$\underline{v}_{j-1}$). Para detalhes, vide Okada & Tomita (1985) e Duchene & Leclercq (1988).

ii partir diretamente de uma função critério que possa fornecer várias direções ótimas de uma só vez, ou seja, deseja-se determinar a matriz A (mxd), com m<d, tal que para a transformação linear $\underline{y}=A^T\underline{x}$ tenhamos um máximo da função critério (por exemplo) $tr(W_y^{-1}B_y)$, onde as matrizes W_y e B_y são matrizes de espalhamento no espaço de vetores $\underline{y}$. Fukunaga (1990) demonstra que as direções ótimas são aquelas dos autovetores $\underline{e}_i$ da matriz $W^{-1}B$, ressaltando-se que os autovetores não são ortogonais pois $W^{-1}B$ não é simétrica.

Lembrando que matriz B é da forma $\underline{\alpha}_1 \cdot \underline{\alpha}_1^T + \underline{\alpha}_2 \cdot \underline{\alpha}_2^T + \ldots + \underline{\alpha}_c \cdot \underline{\alpha}_c^T$, onde $\sum_{i=1}^{c} \underline{\alpha}_i = \underline{0}$, e como cada matriz $\underline{\alpha}_i \cdot \underline{\alpha}_i^T$ tem característica 1 e só há c-1 matrizes $\underline{\alpha}_i \cdot \underline{\alpha}_i^T$ linearmente independentes, então a característica de B é $\theta = \min [\gamma, d]$, onde em geral, $\gamma = c-1$ (em casos particulares pode resultar menor que c-1). Portanto, a característica de $W^{-1}B$ também é θ pois W^{-1} é não singular. Decorre dis to que há θ autovalores não nulos para $W^{-1}B$ e portanto esta técnica pode fornecer no máximo θ direções ótimas. Se $m=\theta$, basta tomar todos os autovetores obtidos; se $m<\theta$, tomar os m autovetores associados aos m maiores autovalores (lembrando que o valor da função critério, $tr(W_y^{-1}B)$, é igual à soma dos autovalores no espaço de dimensão m) e se $m>\theta$ ficam faltando $m-\theta$ direções que devem ser determinadas com alguma outra abordagem. Uma observação é que apesar dos autovetores não serem ortogonais, eles apresentam uma espécie de ortogonalidade, conforme visto no teorema a seguir.

Teorema: Para autovalores distintos tem-se $\underline{e}_i^T W \underline{e}_j = 0$, $i \neq j$.

Demonstração:

Sabemos que $B\underline{w}^* = \lambda W \underline{w}^*$ e portanto

$$\underline{w}_j^{*T} B \underline{w}_i^* = \underline{w}_j^{*T} \lambda_i W \underline{w}_i^*$$

$$\underline{w}_i^{*T} B \underline{w}_j^* = \underline{w}_i^{*T} \lambda_j W \underline{w}_j^*$$

e transpondo a última equação :

$$\underline{w}_j^{*T} B \underline{w}_i^* = \lambda_j \underline{w}_j^{*T} W \underline{w}_i^*$$

Subtraindo esta última equação da primeira:

$$(\lambda_i - \lambda_j) \underline{w}_j^{*T} W \underline{w}_i^* = 0 \qquad \therefore \quad p/\ \lambda_i \neq \lambda_j$$

$$\underline{w}_j^{*T} W \underline{w}_i^* = 0 \ ; \quad i \neq j$$

Há portanto um tipo de ortogonalidade normalizada entre as diversas direções ótimas.

Em termos práticos, espera-se que um número pequeno de funções discriminantes $\underline{e}_i^T\underline{x}$ seja suficiente para prover uma boa separabilidade entre as c classes de forma a haver uma boa redução na dimensionalidade do problema de classificação.

Suponhamos que se tomem $r \leq \theta$ funções discriminantes. Uma possível regra de classificação é:

* associar $\underline{x}$ à classe ω_k se

$$\sum_{j=1}^{r}\left(\underline{w}_j^{*T}(\underline{x}-\overline{\underline{x}}_k)\right)^2 \leq \sum_{j=1}^{r}\left(\underline{w}_j^{*T}(\underline{x}-\overline{\underline{x}}_i)\right)^2, \text{ para todo } i\neq k$$

Isto pode ser expresso de forma mais compacta se tomarmos

$$\underline{y} \triangleq \left[\underline{w}_1^{*T}\cdot\underline{x} \quad \underline{w}_2^{*T}\cdot\underline{x}\ldots \quad \underline{w}_r^{*T}\cdot\underline{x}\right]^T \quad \text{e} \quad \overline{\underline{y}}_i = \left[\underline{w}_1^{*T}\cdot\overline{\underline{x}}_i \quad \underline{w}_2^{*T}\cdot\overline{\underline{x}}_i \ \ldots\ \underline{w}_r^{*T}\ \overline{\underline{x}}_i\right]^T$$

sendo, respectivamente, a projeção de $\underline{x}$ nas r direções $\underline{w}_j^*$ e a projeção do i-ésimo vetor médio $\overline{\underline{x}}_i$ nas mesmas r direções. Tomando a norma ou distância Euclideana temos a regra de classificação ;

* associar $\underline{x}$ à classe ω_k se

$$\| \underline{y} - \overline{\underline{y}}_k \|_2^2 \leq \| \underline{y} - \overline{\underline{y}}_i \|_2^2 \qquad \text{para todo} \quad i\neq k$$

A abordagem apresentada não é a única possível para múltiplas classes. Uma visão alternativa do problema é tomar as classes aos pares, calculando c(c-1)/2 funções discriminantes associadas a todos os pares i, j de classes, o que equivale ao Caso 2 do capítulo de Funções de Decisão :

$$y_{ij} = \underline{w}_{ij}^{*T}\,\underline{x} \quad \text{ou} \quad y_{ij} = (\overline{\underline{x}}_i - \overline{\underline{x}}_j)^T\, S_{ij}^{-1}\,\underline{x} \qquad \text{(semelhante a (28))}$$

onde

$$S_{ij} \triangleq \frac{(n_i-1)\,S_i + (n_j-1)\,S_j}{n_i+n_j-2} \qquad \text{(semelhante a (25))}$$

$$S_i \triangleq \frac{1}{n_i-1}\sum_{k=1}^{n_i}(\underline{x}_{ik}-\overline{\underline{x}}_i)(\underline{x}_{ik}-\overline{\underline{x}}_i)^T \qquad \text{(semelhante a (22))}$$

e $\quad \bar{\underline{x}}_i = \frac{1}{n_i} \sum_{k=1}^{n_i} \underline{x}_{ik}$ (semelhante a (20))

Para cada par de classes i,j pode-se definir um limiar de decisão seguindo as expressões (29a) ou (29b), dependendo da igualdade ou não das matrizes de covariância envolvidas. Tomando para simplicidade notacional o caso (29a), tem-se o limiar

$$\underline{w}_{ij}^{*T} \cdot \bar{\underline{x}}_{ij} \triangleq \underline{w}_{ij}^{*T} (P_j \bar{\underline{x}}_i + P_i \bar{\underline{x}}_j)/(P_i+P_j) = (\bar{\underline{x}}_i - \bar{\underline{x}}_j)^T S_{ij}^{-1} (P_j \bar{\underline{x}}_i + P_i \bar{\underline{x}}_j)/(P_i+P_j)$$

e podemos então definir as funções de decisão :

$$d_{ij}(\underline{x}) = \underline{w}_{ij}^{*T} (\underline{x} - \bar{\underline{x}}_{ij}) \qquad i \; ; \; j = 1, 2, \ldots c \; ; \; i \neq j$$

ou

$$d_{ij}(\underline{x}) = \underline{w}_{ij}^{*T} \underline{x} - \underline{w}_{ij}^{*T} \bar{\underline{x}}_{ij}$$

que pode ser escrito na forma $d(\underline{x}) = \underline{v}_o^T \underline{x} + v_{d+1}$, como feito no capítulo sobre Funções de Decisão, bastando fazer $\underline{v}_o = \underline{w}_{ij}^*$ e $v_{d+j} = - \underline{w}_{ij}^{*T} \bar{\underline{x}}_{ij}$

Com isto, a regra de decisão fica:

* classificar $\underline{x}$ em ω_k se

$$d_{kj}(\underline{x}) > 0 \quad \text{para} \quad \forall \; j \neq k \; ; \; j=1 \; , \ldots, \; c$$

Outras idéias poderiam ser tentadas, como por exemplo:

⊗ Criar uma função de decisão por classe. Por exemplo para a classe i , agruparíamos as classes restantes em 1 única classe e determinaríamos a função discriminante de Fisher para as 2 classes:

$$d_i(\underline{x}) = \underline{w}_i^{*T} \left(\underline{x} - \frac{\bar{\underline{x}}_i + \bar{\underline{x}}_o}{2} \right)$$

onde $\quad \underline{w}_i^* = S_{fusao}^{-1} (\bar{\underline{x}}_i - \bar{\underline{x}}_o)$

e $\bar{\underline{x}}_o$ se refere à média das médias das outras classes que não a i-ésima. Neste caso, o problema se enquadra no Caso 1 visto no capítulo sobre funções de decisão, se a regra de classificação para a classe ω_k for

$$d_k(\underline{x}) > 0 \quad e \quad d_j(\underline{x}) < 0 \qquad \text{para } j \neq k$$

Não nos parece uma tarefa exeqüível avaliar teoricamente,para casos genéricos, o desempenho das diversas abordagens possíveis. Algumas considerações sobre estimação de taxas de erro podem ser encontradas em Lachenbruch (1975).

MEDIDAS DE DISTÂNCIA

Este capítulo inicialmente apresenta medidas de distância, ou dissimilaridade ou afastamento, entre vetores de padrões ou de atributos. A seguir, apresentam-se medidas de distância entre um vetor e uma classe, e finalmente, medidas de distância entre duas classes. Os atributos serão supostos variáveis contínuas em $\mathbb{R}$.

DISTÂNCIA ENTRE DOIS VETORES

Definição: Distância entre dois vetores arbitrários $\underline{x}$ e $\underline{y} \in \mathbb{R}^d$ é qualquer função real que satisfaz as 3 propriedades:

i $d(\underline{x},\underline{y}) = d(\underline{y},\underline{x})$

ii $d(\underline{x},\underline{y}) \geq 0$

iii $d(\underline{x},\underline{x}) = 0$

Se além de i a iii também valerem

iv $d(\underline{x},\underline{y}) = 0$ se e só se $\underline{x} = \underline{y}$

v $d(\underline{x},\underline{y}) \leq d(\underline{x},\underline{z}) + d(\underline{z},\underline{y})$ (desigualdade do triângulo)

então $d(\underline{x},\underline{y})$ é uma métrica

São apresentados a seguir alguns exemplos importantes de medidas de distância:

i Distância de Minkowski de ordem s

É uma definição bastante geral, que engloba outras definições de distância mais conhecidas como a Euclideana. É uma métrica.

$$d_s(\underline{x},\underline{y}) = \| \underline{x}-\underline{y} \|_s = \left[\sum_{j=1}^{d} | x_j - y_j |^s \right]^{1/s} \qquad s \in \mathbb{N}$$

ii Distância módulo ("city block")

É a métrica de Minkowsky para s=1:

$$d_1(\underline{x},\underline{y}) = \| \underline{x}-\underline{y} \|_1 = \sum_{j=1}^{d} | x_j - y_j |$$

O nome "city block" vem do fato da semelhança desta medida com a forma de se medir distância ao longo de ruas e quarteirões entre dois pontos em uma cidade.

iii Distância Euclideana

É a definição mais conhecida e mais utilizada, sendo a métrica de Minkowski para s=2:

$$d_2(\underline{x},\underline{y}) = \| \underline{x}-\underline{y} \|_2 = \left[\sum_{j=1}^{d} (x_j-y_j)^2 \right]^{1/2} = \left[(\underline{x}-\underline{y})^T (\underline{x}-\underline{y}) \right]^{1/2}$$

iv Distância de Pearson ou ponderada

$$d_p(\underline{x},\underline{y}) = \left[\sum_{j=1}^{d} \frac{(x_j-y_j)^2}{s_j^2} \right]^{1/2} = \left[\sum_{j=1}^{d} \left(\frac{x_j-y_j}{s_j} \right)^2 \right]^{1/2}$$

onde s_j^2 é a variância da j-ésima variável.

Esta distância é utilizada no lugar da Euclideana quando as variáveis que compõem os vetores $\underline{x}$ e $\underline{y}$ não tem a mesma unidade ou a mesma ordem de grandeza. Por exemplo: para caracterizar um sinal de duração finita poderíamos utilizar os atributos: x_1 = amplitude de pico, x_2 = duração, x_3 = tempo de subida. Fica claro que x_1 não pode ser comparado com x_2 e x_3 por terem unidades diferentes. Mas x_2 e x_3 também não serão comparáveis caso as ordens de grandeza sejam diferentes. Suponhamos como ilustração que a duração pode variar na gama entre 20 e 60 ms e o tempo de subida na faixa de 1 a 3ms (Fig. 1). Se optarmos pela distância Euclideana $\sqrt{(x_2-y_2)^2 + (x_3-y_3)^2}$ teremos incongruências como: um sinal $\underline{z}$, com z_2 = 20 e z_3= 3 (ponto D na Fig. 1) , está à mesma distância do sinal $\underline{x}$, com x_2 = 20 e x_3 = 1 (ponto A na

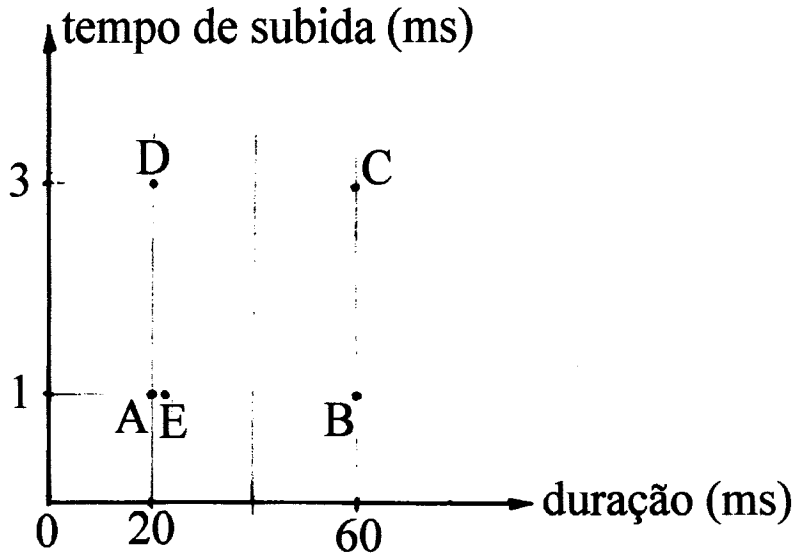


Fig. 1 – Supondo sinais representados por duração e tempo de subida, os dois pares A-D e A-E estariam à mesma distância caso se utilizasse a distância Euclidiana.

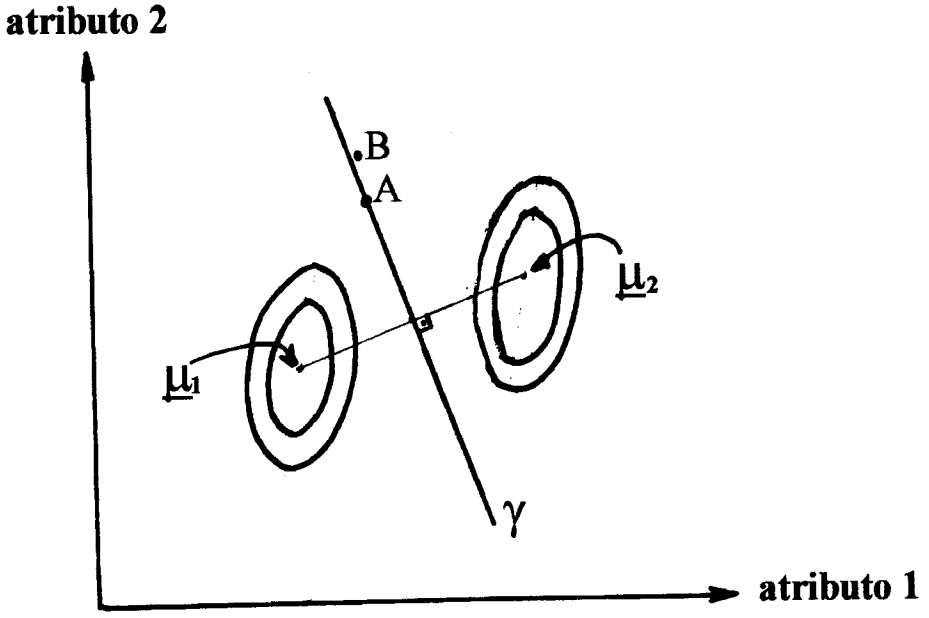


Fig. 2 – Exemplo de duas distribuições com mesma matriz de covariância. A reta γ é a mediatriz do segmento que une os pontos μ_1 e μ_2.

Fig. 1), do que o sinal $\underline{x}$ ao sinal $\underline{y}$, com $y_2 = 22$ e $y_3 = 1$ (ponto E da Fig. 1). Isto não é razoável, para certas aplicações, pois $\underline{y}$ é bem mais similar a $\underline{x}$ do que $\underline{z}$ pois este tem o triplo do tempo de subida que $\underline{x}$ ao passo que $\underline{y}$ só difere de $\underline{p}$ por uma variação de 10% na duração. A distância de Pearson padroniza, ou normaliza, as variáveis ou atributos pelo respectivo desvio padrão. Outras formas de se obter uma padronização é dividir cada variável pela sua gama de variação ou pelo seu máximo, ao invés de dividir pelo desvio padrão.

v Distância de Mahalanobis

É uma distância ponderada conforme visto abaixo:

$$d_M(\underline{x},\underline{y}) = \left[(\underline{x}-\underline{y})^T \sum^{-1} (\underline{x}-\underline{y}) \right]^{1/2}$$

onde $\sum$ é a matriz de covariância do vetor aleatório $\underline{X}$, ou sua estimativa a partir de amostras. $\sum$ é uma matriz simétrica positiva definida.

A distância ponderada é um caso particular da distância de Mahalanobis, bastando fazer $\sum$ = diag $[s_1^2 \quad s_2^2 \ldots\ldots s_d^2]$, que seria a matriz de covariância para o caso de atributos não-correlacionados.

Suponha haver 2 classes, ambas com matriz de covariância $\sum$. No caso geral, as variáveis ou atributos são correlacionados e $\sum$ é não-diagonal (Fig. 2). Se utilizarmos a distância Euclideana, os pontos da reta γ são equidistantes de μ_1 e μ_2. Mas, é fácil ver que isto não significa que qualquer ponto de γ pode tanto ser atribuído à classe ω_1 ou ω_2. O ponto A , por exemplo, está em γ, mas está mais próximo de ω_1. O ponto B está à direita da reta γ o que em termos de distância Euclideana para μ_1 e μ_2 o classificaria em ω_2 , muito embora B esteja mais perto de ω_1. Para podermos usar a distância Euclideana devemos ter as variáveis não correlacionadas o que é conseguido com a transformação de Mahalanobis: dado um vetor $\underline{x}$ nas coordenadas originais utilizar o vetor $\underline{z} = \Sigma^{-1/2}\ \underline{x}$ (caso a matriz de covariância não

seja conhecida, utilizamos um estimador) com métrica Euclideana. Está então justificada a expressão para a distância de Mahalanobis.

Muitas vezes se utiliza o quadrado da distância de Mahalonobis ao invés da distância de Mahalanobis.

vi Distância quadrática

$$d_Q(\underline{x},\underline{y}) = [(\underline{x}-\underline{y})^T A(\underline{x}-\underline{y})]^{1/2}$$

onde A é uma matriz simétrica positiva definida. Esta é uma generalização da distância de Mahalanobis.

vii Distância de Chebyshev ou distância "sup"

$$d_\infty(\underline{x},\underline{y}) = \max_{1 \le i \le d} |x_i - y_i|$$

que é a métrica de Minkowsky para $s \to \infty$

⊙⊙⊙⊙⊙ Exemplo : Em 2 dimensões,representamos na Fig. 3a, o lugar geométrico dos pontos que têm a mesma distância $d(\underline{x},\underline{0})=2$ em relação à origem, utilizando-se d_1, d_2 e d_∞. A Fig. 3b mostra o mesmo para d_p, com $s_1 = 1$ e $s_2 = 2$. Finalmente, a Fig. 3c mostra o lugar geométrico para o caso da distância de Mahalanobis. Em dimensões maiores, a circunferência da Fig. 3a para o caso Euclideano se generaliza para hiper-esfera e para os outros dois casos para hiper-cubo. As Figuras 3b e 3c para dimensões maiores resultariam em uma hiper-elipsóide.

viii Distância não linear

$$d_N(\underline{x},\underline{y}) = \begin{cases} H & \text{se } d(\underline{x},\underline{y}) > L \\ 0 & \text{se } d(\underline{x},\underline{y}) \le L \end{cases} \qquad (H, L \in \mathbb{R}^+)$$

onde H e L são parâmetros a escolher e $d(\underline{x},\underline{y})$ é uma medida de distância também a escolher, podendo ser, por exemplo, qualquer uma das já vistas anteriormente. O parâmetro L define um limiar de tal forma que se dois vetores $\underline{x}$ e $\underline{y}$ têm $d(\underline{x},\underline{y}) \le L$, então isto significaria que o seu afastamento é devido a variabilidades intrínsecas dentro de uma mesma classe ou devido a ruídos. Caso os dois vetores nestas condições pertençam a classes diferentes

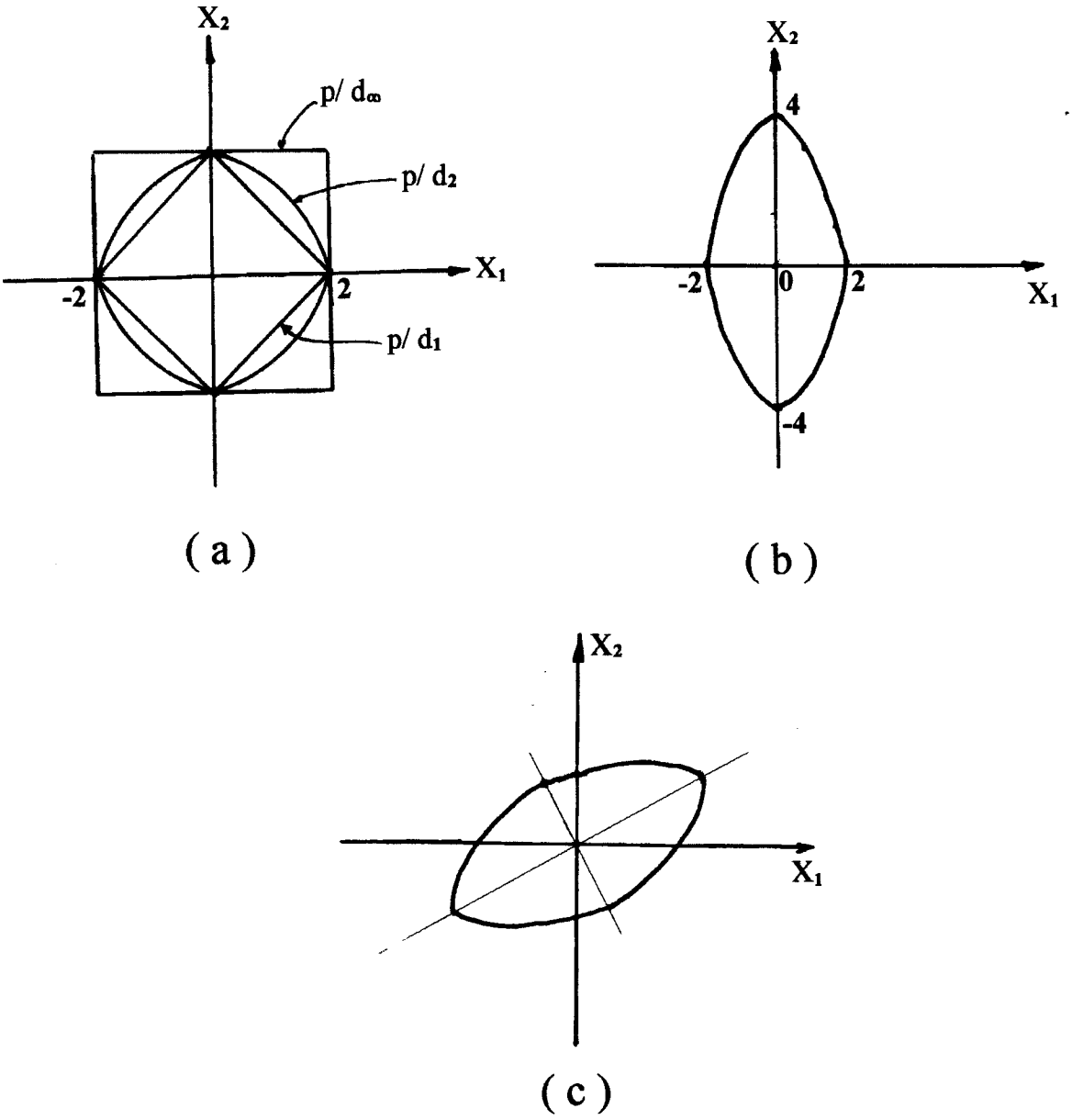


Fig. 3 – (a) Lugares geométricos dos pontos eqüidistantes da origem segundo as distâncias d_1, d_2 e d_∞; (b) idem para a distância d_p (s_1=1, s_2=2); (c) idem para a distância d_M em que a correlação entre x_1 e x_2 é positiva.

então poderá haver erro de classificação e portanto para fins de classificação pode-se tomar a sua distância como nula ($d_N(\underline{x},\underline{y}) = 0$).

A argumentação para a definição desta distância é: caso dois vetores $\underline{x}$ e $\underline{y}$ de classes diferentes estejam a uma distância "razoável", de pelo menos L, então estes vetores não poderão ser classificados erradamente. Do ponto de vista de classificação, se a distância for um pouco, ou for muito maior que L, não faz diferença, pois a probabilidade de erro é nula em ambos os casos. Por esta razão toma-se $d_N(\underline{x},\underline{y}) = H$ para vetores distantes "o suficiente" entre si.

PROXIMIDADE ENTRE DOIS VETORES

Até agora foram apresentados índices de dissimilaridade. Um índice de similaridade ou proximidade por vezes utilizado é a correlação ou co-seno do ângulo entre os vetores $\underline{x}$ e $\underline{y}$:

$$s(\underline{x},\underline{y}) = \frac{\underline{x}^T \cdot \underline{y}}{\|\underline{x}\|_2 \ \|\underline{y}\|_2}$$

Este índice é máximo quando os vetores têm a mesma direção, embora possam ter normas bem diferentes. No caso de $\underline{x}$ e $\underline{y}$ representarem amostras de dois sinais x(t) e y(t), respectivamente, a similaridade seria máxima quando

$$x(t) = \alpha y(t), \quad \alpha \in \mathbb{R}^+,$$

o que significa que a forma dos dois sinais é exatamente a mesma, ao passo que as amplitudes podem ter uma relação arbitrária. É claro que em várias aplicações práticas este índice não é o mais adequado. A Fig. 4 mostra um caso de 2 classes, com amostras ocupando regiões alongadas, aproximadamente na direção de duas retas passando pela origem. Neste caso o ângulo entre os vetores é um bom índice para classificação.

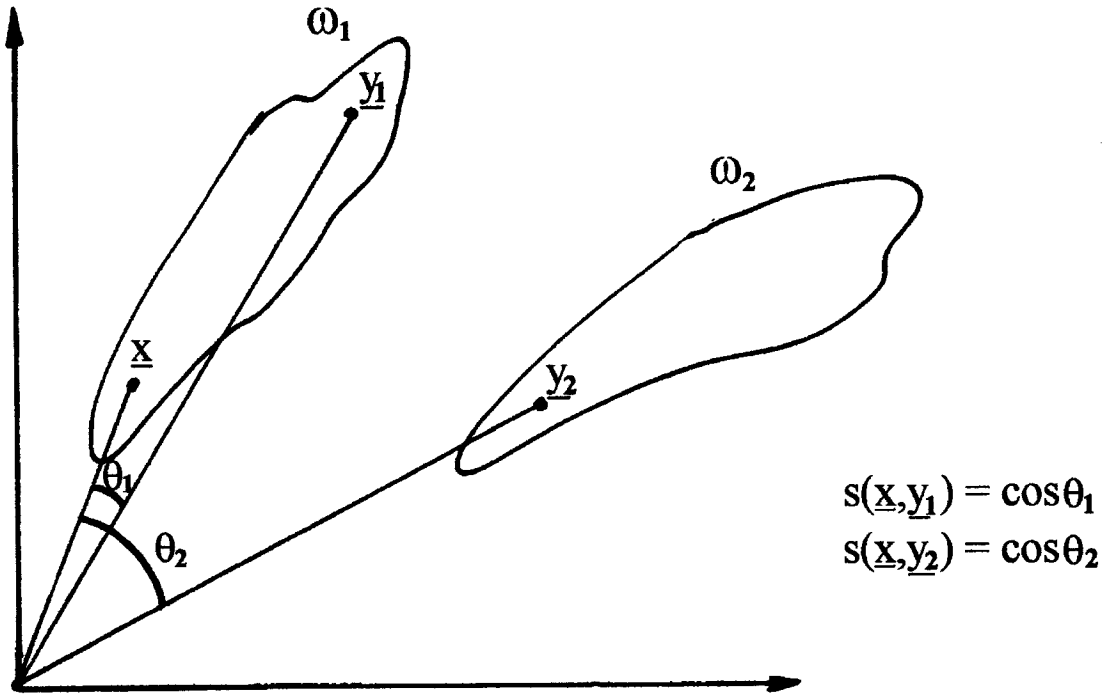


Fig. 4 – Ilustração de um caso em que o conceito de proximidade por ângulo pode ser adequado

DISTÂNCIA ENTRE UM VETOR E UMA CLASSE OU AGRUPAMENTO

A necessidade de se definir distância entre vetor e classe ou agrupamento aparece tanto no classificador por vizinhos mais próximos, já visto, como na análise de agrupamentos em que se tentam agrupar padrões similares entre si. Padrões (com classificação desconhecida) similares entre si são atribuidos a um mesmo agrupamento enquanto padrões dissimilares àqueles, são atribuídos a agrupamentos distintos. Portanto, para estas finalidades, uma classe ou agrupamento será entendida, no que segue, como definida ou representada por um certo número de vetores de atributos disponíveis. Da mesma forma que no caso de medidas de distância entre dois vetores, apresentaremos alguns exemplos importantes na prática.

i Distância ao centróide da classe ou agrupamento

A distância entre um vetor $\underline{x}$ e a j-ésima classe ou agrupamento é definida como a distância entre o vetor $\underline{x}$ e o vetor $\overline{\underline{x}}_j$ que é o centróide da classe ou agrupamento em estudo. Esta distância entre os dois vetores pode ser escolhida como qualquer das distâncias já apresentadas na seção anterior. Convém ressaltar que caso se conheça a matriz de covariância (ou um seu estimador) pode-se, por exemplo, utilizar a distância de Mahalanobis que deverá dar melhores resultados que a distância Euclideana. O mesmo se aplica aos casos seguintes.

ii Distância ao elemento mais próximo da classe ou agrupamento

A distância entre o vetor $\underline{x}$ e a j-ésima classe ou agrupamento é definida como a distância entre o vetor $\underline{x}$ e o vetor mais próximo a $\underline{x}$ que pertence à j-ésima classe ou agrupamento. A Fig. 5a ilustra esta definição.

Esta é a distância utilizada no classificador do vizinho mais próximo

(1-NN) e no método hierárquico de agrupamento de vizinho mais próximo.

iii Distância ao elemento mais afastado da classe ou agrupamento

A distância entre o vetor $\underline{x}$ e a j-ésima classe ou agrupamento é dada pela distância entre o vetor $\underline{x}$ e o vetor mais afastado de $\underline{x}$ que pertence à j-ésima classe ou agrupamento. Na Fig. 5b, como a maioria das amostras da classe ou agrupamento γ_2 está mais distante do vetor $\underline{x}$ do que os vetores de γ_1 da Fig. 5a em relação a $\underline{x}$ (apesar da distância do vizinho mais próximo de γ_1 a $\underline{x}$ ser igual à do vizinho mais próximo de γ_2 a $\underline{x}$), parece razoável dizer que $d(\underline{x},\gamma_1)<d(\underline{x},\gamma_2)$ e, portanto, neste exemplo, a distância ao elemento mais afastado da classe ou agrupamento é relevante.

iv Distância média aos elementos da classe ou agrupamento

A distância entre o vetor $\underline{x}$ e a j-ésima classe ou agrupamento é dada pela média das distâncias de $\underline{x}$ a todos os elementos desta classe ou agrupamento. A Fig. 5c exemplifica um caso em que as distâncias do vizinho mais próximo e do vizinho mais distante a $\underline{x}$ serão iguais aos dos casos das Figs. 5a e 5b. Entretanto, como a maior parte da "massa" de pontos está mais próxima de $\underline{x}$ na Fig. 5c do que na Fig. 5b, fica aparente que possivelmente deveríamos ter $d(\underline{x},\gamma_3)<d(\underline{x},\gamma_2)$, o que é conseguido com a utilização da medida de distância definida neste item. Deve ficar claro para o leitor que a seleção de uma ou outra medida de distância irá depender muito do contexto do problema, e portanto o especialista em reconhecimento de padrões deverá ter uma boa intuição sobre o problema específico sendo tratado, ou obter esta intuição através da parceria com um especialista no problema específico.

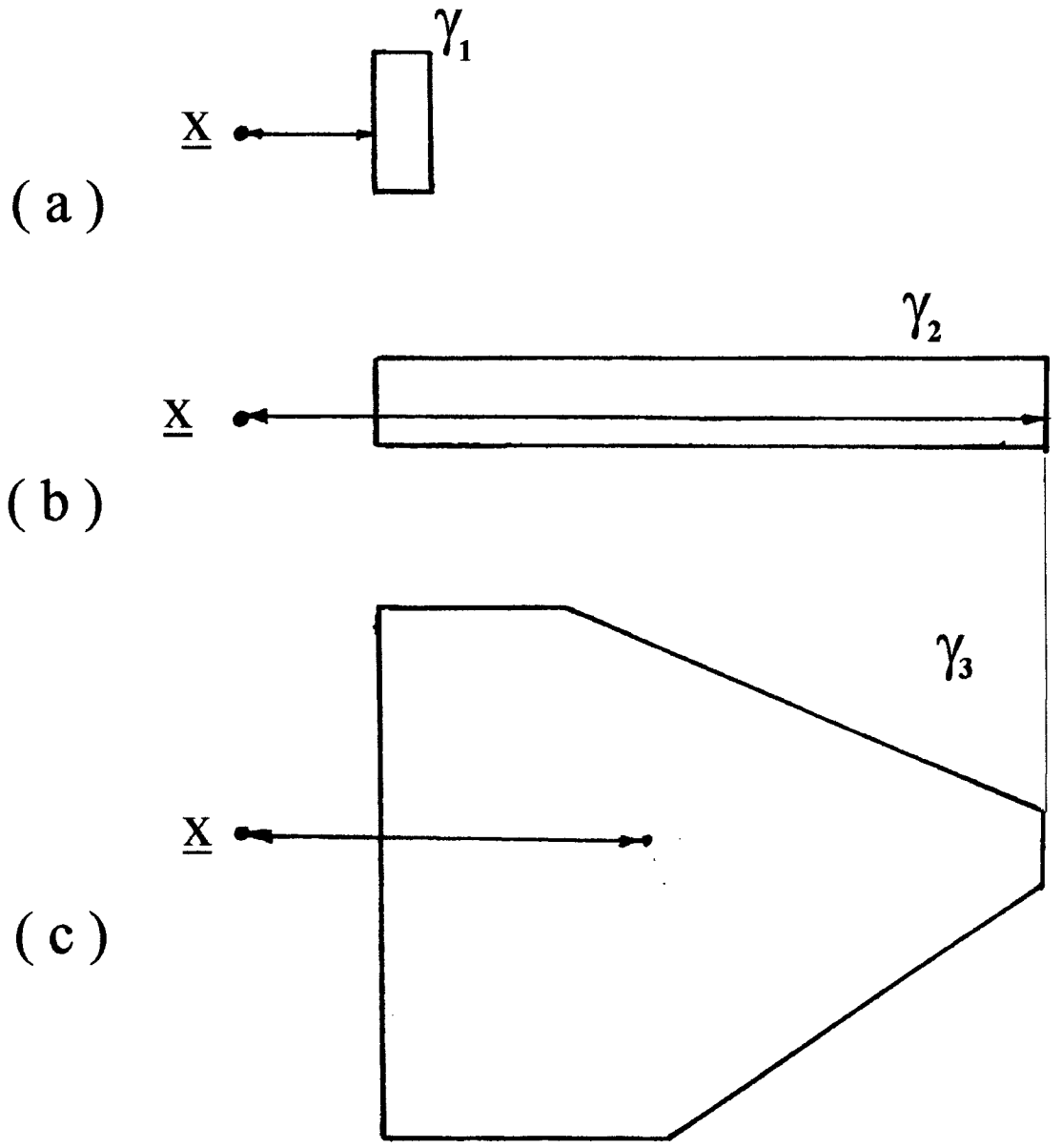


Fig. 5 – Distância entre um vetor e uma classe ou agrupamento.

DISTÂNCIA ENTRE DUAS CLASSES OU AGRUPAMENTOS

Este conceito tem utilidade em várias sub-áreas de Reconhecimento de Padrões, como: análise discriminante, seleção e extração de atributos, análise de agrupamentos.

Dois casos distintos serão apresentados: I no caso de classes ou agrupamentos representados por amostras (vetoriais) conhecidas, a definição de distância se baseia diretamente nas amostras disponíveis; II no caso de classes para as quais se conhecem as distribuições de probabilidade, ou em que estas são estimadas a partir de amostras com classificação conhecida, a definição de distância ou separabilidade utiliza todo o conhecimento probabilístico disponível ou estimado. As primeiras cinco definições de distância pertencem ao caso I ao passo que as restantes ao caso II . Onde relevante, supomos a i-ésima classe ou agrupamento representados por n_i elementos indicados por $\underline{x}_i(k)$, $k = 1,\ldots n_i$.

CASO I: Distâncias baseadas em amostras

I-i Distância entre centróides

A distância entre o i-ésimo e o j-ésimo agrupamento ou classe é definida como a distância entre os respectivos centróides $\overline{\underline{x}}_i$ e $\overline{\underline{x}}_j$:

$$d_{ij} = d\left[\overline{\underline{x}}_i , \overline{\underline{x}}_j \right]$$

I-ii Distância entre vizinhos mais próximos

A distância entre o i-ésimo e o j-ésimo agrupamentos ou classes é definida como a menor distância possível entre um elemento do j-ésimo e um elemento do i-ésimo agrupamento ou classe:

$$d_{ij} = \min d\left[\underline{x}_{ik} , \underline{x}_{jm} \right] , \quad k=1 , \ldots, n_i ; \quad m = 1,\ldots n_j$$

I-iii Distância entre vizinhos mais afastados

A distância entre duas classes ou agrupamentos é definida como a maior distância existente entre dois elementos das duas classes ou agrupamentos, tomando 1 elemento de cada:

$$d_{ij} = \max d(\underline{x}_{ik}, \underline{x}_{jm}), \quad k = 1,\ldots n_i ; \quad m = 1, \ldots n_j$$

I-iv Distância média

A distância entre duas classes ou agrupamentos é definida como a média das distâncias de todos os pares de elementos, tomados um de cada classe ou agrupamento:

$$d_{ij} = \frac{1}{n_i n_j} \sum_{k=1}^{n_i} \sum_{m=1}^{n_j} d\left[\underline{x}_{ik}, \underline{x}_{jm}\right]$$

I-v Distâncias baseadas nas matrizes de espalhamento T, B e W.

As medidas de distância ou separabilidade apresentadas a seguir têm sua inspiração na análise discriminante. A primeira relaciona o traço da matriz de espalhamento entre-classes, ou entre-agrupamentos, com o traço da matriz de espalhamento intra-classe, ou intra-agrupamento.

$$d_{ij} = \frac{tr\ B}{tr\ W}$$

onde

$$B = \frac{n_i n_j}{n_i+n_j} (\overline{\underline{x}}_i - \overline{\underline{x}}_j)(\overline{\underline{x}}_i - \overline{\underline{x}}_j)^T \qquad \text{e}$$

$$W = \sum_{n=1}^{n_i} (\underline{x}_{in} - \overline{\underline{x}}_i)(\underline{x}_{in} - \overline{\underline{x}}_i)^T + \sum_{n=1}^{n_j} (\underline{x}_{jn} - \overline{\underline{x}}_j)(\underline{x}_{jn} - \overline{\underline{x}}_j)^T$$

e portanto $tr\ B = \frac{n_i n_j}{n_i+n_j} d_2^2(\overline{\underline{x}}_i, \overline{\underline{x}}_j)$, ou seja, é proporcional à distância Euclideana entre os centróides.

Essa medida de distância refina a medida de distância entre centróides pois leva em consideração a compactação média intra-classes ou intra-agrupamentos.

Uma outra medida semelhante é

$$d_{ij} = \text{tr } W^{-1}B$$

Esta medida pode ser escrita também como

$$d_{ij} = \Sigma \text{ autovalores de } W^{-1}B = \text{único autovalor não nulo de } W^{-1}B$$

ou ainda como

$$\frac{n_i\, n_j}{n_i+n_j}\ (\overline{\underline{x}}_i - \overline{\underline{x}}_j)^T\ W^{-1}(\overline{\underline{x}}_i - \overline{\underline{x}}_j) \qquad \text{(da propriedade } \text{tr}(\underline{y}\cdot\underline{x}^T)=\underline{x}^T\cdot\underline{y})$$

que é proporcional a uma distância do tipo de Mahalanobis entre $\overline{\underline{x}}_i$ e $\overline{\underline{x}}_j$, tomando W como uma matriz de espalhamento comum às duas classes ou agrupamentos. Deve-se comparar este resultado com o caso anterior. Aqui se tem a distância de Mahalanobis, que leva em conta a correlação entre as variáveis, ao passo que em tr B/ tr W não se leva em conta a correlação entre as variáveis pois se toma o traço de W.

Uma terceira medida utiliza o determinante de matrizes de espalhamento ao invés de traço:

$$d_{ij} = \frac{|\ T\ |}{|\ W\ |}$$

onde T = W + B é a matriz de espalhamento global ou total (vide capítulo sobre Discriminante e Classificador de Fisher). Esta definição pode parecer menos razoável do ponto de vista intuitivo, mas mostraremos abaixo que esta medida está relacionada com a anterior ($\text{tr}(W^{-1}B)$).

Pré e pós multiplicando a equação T=B+W, por $W^{-1/2}$ tem-se:

$$W^{-1/2}\ T\ W^{-1/2} = W^{-1/2}\ BW^{-1/2} + I$$

e tomando o determinante em ambos os membros, tem-se

$$d_{ij} = \frac{|\ T\ |}{|\ W\ |} = \det\left(W^{-1/2}\ B\ W^{-1/2} + I\right)$$

Multiplicando por $W^{1/2}$ à direita e por $W^{-1/2}$ à esquerda não muda o determinante, e portanto

$$d_{ij} = \det(W^{-1}\ B + I)$$

Denotando como ψ a matriz não singular que diagonaliza $W^{-1}B$ através da transformação $\psi^{-1}(W^{-1}B)\psi$, tem-se

$$\det(W^{-1}B+I)=\det[\psi^{-1}(W^{-1}B+I)\psi]=\det(\psi^{-1}(W^{-1}B)\psi + I)$$

onde a matriz diagonal $\psi^{-1}(W^{-1}B)\psi$ só tem 1 autovalor não nulo na sua diagonal e portanto

$$d_{ij} = 1 + \frac{n_i\, n_j}{n_i+n_j}(\bar{\underline{x}}_i - \bar{\underline{x}}_j)^T W^{-1}(\bar{\underline{x}}_i - \bar{\underline{x}}_j) =$$

$$= 1 + \text{único autovalor não nulo de } W^{-1}B \ .$$

Conclui-se que a distância entre <u>duas</u> classes, ou agrupamentos, difere apenas de uma constante aditiva quando se utilizam as definições $tr(W^{-1}B)$ e $|T|/|W|$. Entretanto, para o caso de distância entre <u>várias</u> classes, ou agrupamentos, essas duas definições serão bastante diferentes, pois em $|T|/|W|$ aparecem termos em produto de autovalores de $W^{-1}B$. Deve-se observar que para distância entre <u>duas</u> classes não se pode usar nenhuma definição de distância que inclua $|B|$, pois este será nulo.

CASO II: <u>Distâncias baseadas nas funções densidade de probabilidade</u>

Nesta abordagem, uma distância é definida utilizando-se toda a informação probabilística que caracteriza as duas classes, ou agrupamentos, e com isto tem-se potencialmente uma forma mais poderosa ou completa de se conceituar o que seja uma distância ou separação entre <u>2 classes</u> (ou agrupamentos).

Há várias definições propostas na literatura sendo que a expressão geral é :

$$d_{ij} = \int g\left[p(\underline{x}|\omega_i), p(\underline{x}|\omega_j), P_i, P_j\right] d\underline{x}$$

que deve obedecer às 3 condições:

<u>i</u> $d_{ij} \geq 0$

<u>ii</u> d_{ij} é máximo se $p(\underline{x}|\omega_i) = 0$ quando $p(\underline{x}|\omega_j) \neq 0$ para qualquer $\underline{x}$, ou seja quando as classes são disjuntas

iii $d_{ij} = 0$ se $p(\underline{x}|\omega_j) = p(\underline{x}|\omega_j)$, $\forall$ $\underline{x}$

No que segue, o texto referir-se-á apenas ao caso de classes, ficando subentendido que no caso de agrupamentos, sem informação a priori, seria necessário estimar tanto as funções densidade condicionais como as probabilidades de ocorrência de uma amostra de cada agrupamento, para então calcular numericamente a distância. Caso a estimação das funções densidade seja paramétrica, basta empregar a expressão teórica da distância (por exemplo de Bhattacharyya, que será vista a seguir).

Serão apresentadas algumas definições de separabilidade entre duas classes, conforme propostas na literatura.

II-i Distância de Bhattacharyya

$$d_{ij,B} = - ln \int_{\Omega} \left[p(\underline{x}|\omega_i)\ p(\underline{x}|\omega_j) \right]^{1/2} d\underline{x}$$

Esta definição satisfaz as 3 condições acima, notando-se que se as classes são disjuntas $d_{ij,B} = +\infty$. Define-se o coeficiente de Bhattacharyya como

$$\rho_B = e^{-d_{ij,B}} = \int_{\Omega} \left[p(x|\omega_i) p(\underline{x}|\omega_j) \right]^{1/2} d\underline{x}$$

que pode ser escrito como

$$\rho_B = \int_{\Omega} \left[\frac{p(\underline{x}|\omega_i)}{p(\underline{x}|\omega_j)} \right]^{1/2} p(\underline{x}|\omega_j) d\underline{x} = E\left[\left(\frac{p(\underline{X}|\omega_i)}{p(\underline{X}|\omega_j)} \right)^{1/2} |\omega_j \right]$$

onde $f(\underline{x}) \triangleq \left(\frac{p(\underline{x}|\omega_i)}{p(\underline{x}|\omega_j)} \right)^{1/2}$ é a raiz quadrada da razão de verossimilhança, de

onde $\rho_B = E\left[f(\underline{X}) | \omega_j \right]$. Além do mais, tem-se

$$d_{ij,B} = - ln(\rho_B)$$

A Fig. 6 mostra um exemplo de três funções densidade Gaussianas unidimensionais, juntamente com as respectivas funções $f(\underline{x})$ para os pares 1-2

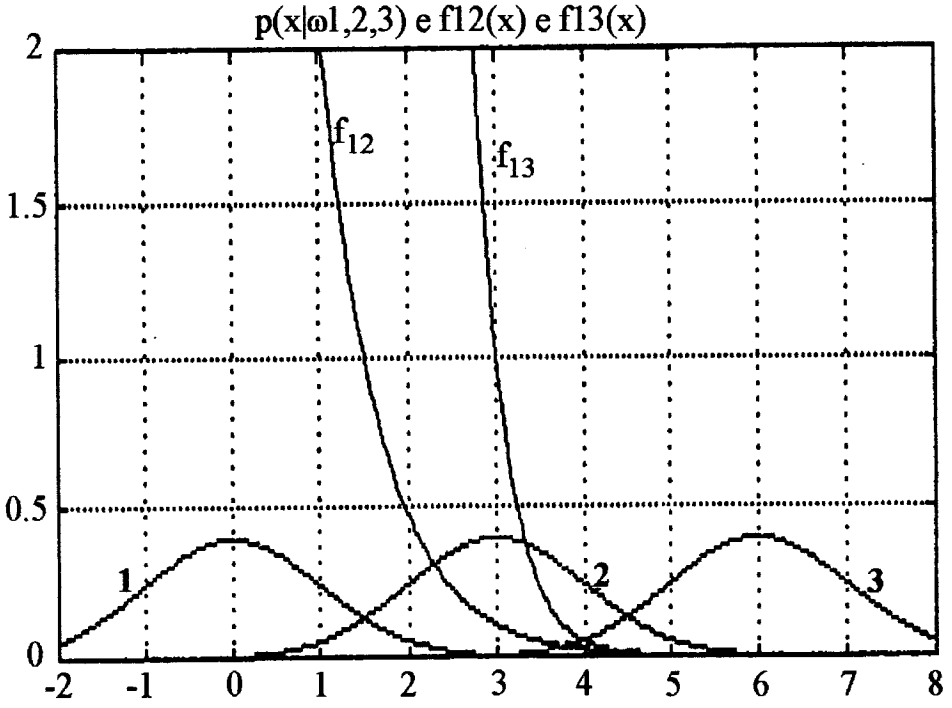


Fig. 6 – Três funções densidade de probabilidade, $p(x|\omega_1)$, $p(x|\omega_2)$ e $p(x|\omega_3)$ Gaussianas juntamente com duas funções $f_{12}(x) = \sqrt{\frac{p(x|\omega_1)}{p(x|\omega_2)}}$ e $f_{13}(x)$.

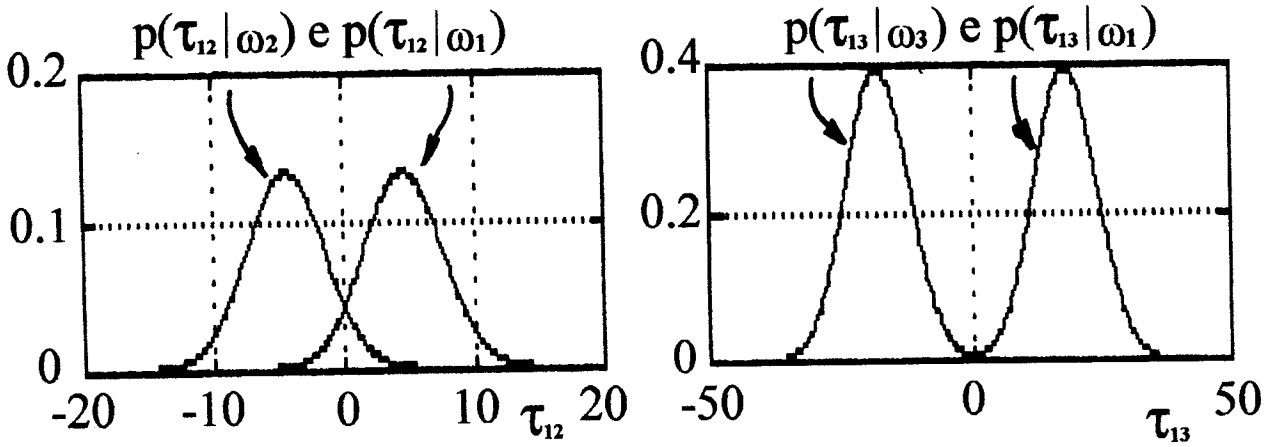


Fig. 7 – Funções densidade da variável aleatória $\tau_{ij}(x)$ para as densidades $f(x|\omega_i)$ utilizadas na Fig. 6.

e 1-3. A comparação das integrais de termos do tipo $f(\underline{x})\cdot p(\underline{x}|\omega_i)$ permite afirmações quanto à menor ou maior distância entre duas funções densidade (segundo a definição de Bhattacharyya).

É fácil notar que se as densidades $p(\underline{x}|\omega_i)$ e $p(\underline{x}|\omega_j)$ estão bem separadas entre si, então $f(\underline{x})$ assume valores pequenos na região em que $p(\underline{x}|\omega_j)$ é grande e vice-versa, fazendo com que $\rho_B = E\left[f(\underline{X})|\omega_j \right]$ seja pequeno e portanto $d_{ij,B}$ seja grande. Se as densidades estão muito juntas então ρ_B é grande e $d_{ij,B}$ é pequena.

Esta medida de distância é invariante a transformações lineares não singulares pois para $\underline{Y} = A\,\underline{X}$ tem-se $p_y(\gamma) = |A^{-1}|\, p_x(\gamma)$ onde os índices y e x indicam para quais vetores são as funções densidade.

II-ii Divergência

$$d_{ij,D} = \int_{\Omega} [p(\underline{x}|\omega_i) - p(\underline{x}|\omega_j)]\, ln \frac{p(\underline{x}|\omega_i)}{p(\underline{x}|\omega_j)}\, d\underline{x}$$

O leitor interessado pode demonstrar que a divergência satisfaz as 3 condições i a iii citadas anteriormente além de ser invariante a transformações não singulares em $\underline{x}$. Uma expressão alternativa para a divergência em termos do logaritmo da razão de verossimilhança é

$$d_{ij,D} = E\left[ln\left(\frac{p(\underline{X}|\omega_i)}{p(\underline{X}|\omega_j)} \right) \middle| \omega_i \right] - E\left[ln\left(\frac{p(\underline{X}|\omega_i)}{p(\underline{X}|\omega_j)} \right) \middle| \omega_j \right]$$

Definindo $ln\left[p(\underline{X}|\omega_i)/p(\underline{X}|\omega_j)\right]$ como $\tau_{ij}(\underline{X})$, obtém-se uma interpretação para a divergência como sendo a diferença entre os valores esperados da variável $\tau_{ij}(\underline{X})$ condicionados a ω_i e ω_j .A Fig. 7 mostra duas funções densidade de probabilidade da variável τ_{ij}, a primeira condicionada a ω_j e a segunda a ω_i , supondo as duas densidades condicionais $p(\underline{x}|\omega_i)$ e $p(\underline{x}|\omega_j)$ como esboçadas na Fig. 6. Deve-se atentar para o fato de que τ_{ij} assume valores positivos com grande probabilidade quando $\underline{x}$ pertence à classe ω_i e valores negativos com grande probabilidade quando $\underline{x}$ pertence à classe ω_j .

Há uma série de outras medidas de distância ou separabilidade, e o leitor pode encontrar uma amostra bastante farta em DeVijver e Kittler (1982).

Várias das medidas de distância podem ser relacionadas com a taxa de erro de Bayes E^*, como por exemplo para o caso da distância de Bhattacharyya:

$$E^* \leq \frac{1}{2} e^{-\rho_B} = \frac{1}{2} e - \left(e^{-d_{ij,B}}\right)$$

As dificuldades computacionais envolvidas nos cálculos destas medidas de distâncias probabilísticas são grandes o que torna seu uso não muito adequado na prática.

Entretanto, no caso Gaussiano as expressões para as distâncias tornam-se mais interessantes do ponto de vista prático:

$$d_{ij,B} = \frac{1}{8}\left(\underline{\mu}_j - \underline{\mu}_i\right)^T \left(\frac{\Sigma_i + \Sigma_j}{2}\right)^{-1}\left(\underline{\mu}_j - \underline{\mu}_i\right) + \frac{1}{2} \ln \frac{0.5\det\left(\Sigma_i + \Sigma_j\right)}{\left[\,|\Sigma_i| \cdot |\Sigma_j|\,\right]^{1/2}}$$

$$d_{ij,D} = \frac{1}{2}\left(\underline{\mu}_j - \underline{\mu}_i\right)^T \left(\Sigma_i^{-1} + \Sigma_j^{-1}\right)\left(\underline{\mu}_j - \underline{\mu}_i\right) + \frac{1}{2} \operatorname{tr}\left\{\Sigma_i^{-1}\Sigma_j + \Sigma_j^{-1}\Sigma_i - 2I\right\}$$

e para $\Sigma_i = \Sigma_j = \Sigma$ tem-se

$$d_{ij,M}(\underline{u}_j, \underline{u}_i) = d_{ij,D} = 8\, d_{ij,B} = \left(\underline{\mu}_j - \underline{\mu}_i\right)^T \Sigma^{-1}\left(\underline{\mu}_j - \underline{\mu}_i\right)$$

onde $d_{ij,M}$ é a distância de Mahalanobis.

SELEÇÃO DE ATRIBUTOS

INTRODUÇÃO

Um padrão pode ser associado a um vetor $\underline{z}$ contendo η ($\eta \in \mathbb{Z}^+$) medidas, observações, variáveis ou atributos z_i. Denotemos por Z o conjunto de atributos $\{ z_1 , z_2 ,\ldots z_n \}$.

⊙⊙⊙⊙⊙ Exemplo Dentro de um problema de análise de sinais, o vetor pode conter todas as η amostras de um dado sinal (tendo-se então η medidas, observações ou variáveis), como por exemplo 10000 amostras de um sinal de eletrocardiograma. O vetor pode, por outro lado, ser uma representação mais compacta do padrão, contendo atributos que juntos são supostos poder sintetizar o que o padrão tem de similar a outros padrões de mesma classe e de dissimilar em relação a padrões de classes diferentes. No exemplo do eletrocardiograma os atributos poderiam ser os intervalos entre batimentos cardíacos e as amplitudes dos picos.

⊙⊙⊙⊙⊙

O problema de seleção de atributos é reduzir a dimensionalidade dos vetores $\underline{z}$ associados aos padrões, de η para d ($d<\eta$), através da seleção de um sub-conjunto dos η atributos ou variáveis originais. No exemplo de potenciais de unidades motoras poder-se-ía sugerir vários atributos como: amplitude pico-a-pico, duração total, duração entre o pico máximo e o mínimo, área do módulo, energia, comprimento (número fractal em Kohn, 1989), assimetria espectral (Kohn, 1989). Poderíamos estar interessados, por exemplo, em verificar quais 3 dos 7 atributos seriam os melhores em termos de manter a separabilidade entre classes. Neste mesmo exemplo de potenciais de ação poder-se-ía partir das próprias amostras de cada potencial de ação e verifi-

car quais são as melhores em termos de separabilidade entre classes (vide por exemplo, Salganicoff et al, 1988).

Há pelo menos três razões pelas quais é desejável efetuar uma redução do número de atributos: i diminuir o custo (instrumentação para efetuar as medidas, tempo de computação, etc) da etapa de medição de variáveis de cada padrão; ii diminuir o custo (equipamento de propósito específico, tempo e memória de computação, etc) da etapa de classificação; iii diminuir redundâncias existentes nas variáveis iniciais que podem prejudicar o desempenho de um classificador dada a necessidade de se obter estimativas a partir de um número finito de padrões de treinamento. Como um exemplo simples deste último caso poderíamos tomar um conjunto de treinamento com 100 padrões. Estes padrões provavelmente seriam adequados para estimar uma função densidade de probabilidade para uma única variável ao passo que se aumentássemos o número de atributos para 4, por exemplo, obteríamos estimativas de função densidade (a 4 variáveis) de qualidade pior. Portanto erros de estimação causam deterioração do desempenho do classificador quando este é aplicado a padrões novos (que não os do conjunto de treinamento) de onde se conclui que um número excessivo de atributos pode prejudicar o desempenho de um classificador que seja projetado a partir de um conjunto de treinamento de tamanho finito.

No capítulo seguinte será abordado o problema de extração de atributos que consiste em determinar um mapeamento das η variáveis para d variáveis ($\mathbb{R}^{\eta} \rightarrow \mathbb{R}^{d}$) que preserve a separabilidade entre classes. Na prática, a extração de atributos exige que todas as η variáveis sejam medidas ao passo que a seleção de atributos tem o mérito de reduzir o número de variáveis a serem medidas, além de reduzir a dimensionalidade do classificador.

Idealmente, tanto a seleção quanto a extração de atributos devem ser realizadas juntamente com a seleção do classificador, de forma a evitar

incompatibilidades entre os atributos finais utilizados e o método de classificação empregado. Na prática, em geral, por motivos de simplicidade, a seleção e a extração de atributos são feitas separadamente do projeto do classificador. A seleção do classificador e a seleção e extração dos atributos, são atividades de projeto, normalmente baseadas em um conjunto de amostras ou padrões com classificação conhecida (conjunto de treinamento ou de aprendizado). Em alguns casos, pode-se fazer a seleção de variáveis adaptativamente aos novos dados que vão chegando para serem classificados.

A seleção de atributos consiste basicamente em uma busca de um subconjunto ótimo de atributos. Para caracterizar a otimalidade, utiliza-se alguma função critério J(.) conveniente, que deve ser otimizada (minimizada ou maximizada). Como a busca exaustiva direta, em geral, é impraticável, emprega-se algum algoritmo mais rápido que ainda forneça resultados adequados.

CRITÉRIOS DE OTIMALIDADE

Há vários critérios propostos na literatura, sendo que apresentaremos duas categorias importantes.

i Mínima Probabilidade de Erro

A minimização da probabilidade de erro de classificação é um objetivo bastante atraente também em seleção de atributos. A sua realização prática é que, novamente, é problemática.

No caso extremamente raro em que se conhece a priori toda a descrição probabilística da geração dos padrões, a taxa de erro de um classificador arbitrário é dada pela expressão (28) do capítulo sobre Teoria de Decisão Bayesiana:

$$E = \sum_{i=1}^{c} \int_{\Omega_i} [1 - P(\omega_i | \underline{x})] p(\underline{x}) d\underline{x} \qquad (1)$$

Dado um classificador e dado um sub-conjunto de m atributos ($m \leq \eta$) de $Z = \{z_1, \ldots, z_\eta\}$, determinam-se as regiões de aceitação Ω_i, e calcula-se E. Esta frase de aparência simples esconde, entretanto, as grandes dificuldades envolvidas na determinação dos Ω_i usados nas integrais nas regiões Ω_i . Se o que se deseja é determinar o melhor sub-conjunto de m elementos de Z então seleciona-se aquele que minimiza E .

Na maioria dos casos práticos não se conhece a descrição probabilística dos padrões e caso se deseje utilizar um classificador que exija este conhecimento (por exemplo o de Bayes) devem-se estimar as funções densidade de probabilidade condicionais e as probabilidades das classes a partir de um conjunto de treinamento. Pode-se então tentar estimar (1) através de métodos numéricos para cálculo de integrais utilizando as estimativas de $P(\omega_i | \underline{x})$ e $p(\underline{x})$ obtidas do conjunto de treinamento. Entretanto, as estimativas de erro obtidas em testes experimentais tendem a ser menores que a taxa de erro verdadeira. Outra abordagem pode ser a utilização de índices como a divergência e a distância de Bhattacharyya que se relacionam com a taxa de erro através de desigualdades. Finalmente outra abordagem consiste em estimar a probabilidade de erro efetuando contagens de classificações erradas. Nesta última técnica deve-se entretanto tomar o cuidado de não utilizar o conjunto de treinamento (utilizado para projetar o classificador) para estimar a probabilidade de erro pois o classificador estará "sintonizado" especificamente para o conjunto de treinamento e a estimativa da probabilidade de erro será muito otimista. Para maiores detalhes vide o capítulo Avaliação do Desempenho de Classificadores.

Caso se utilize algum método que estime a taxa de erro para com isto selecionar o melhor conjunto de m atributos, pode-se obter um gráfico da

taxa de erro em função do número m de atributos. Este gráfico pode ser de grande utilidade para a escolha do número final de atributos d a ser empregado no problema em estudo.

ii Máxima distância entre classes (ou mínima sobreposição de classes)

Caso se disponha de uma medida de distância ou separação entre classes então seleciona-se, para cada m, o conjunto de m atributos que resulta na máxima separação entre as classes. A medida de separação tem que ser global, isto é, para as c classes tomadas simultaneamente. Há um número muito grande de possíveis definições de separabilidade entre classes.

Pode-se optar por uma abordagem em que se utilizam as distâncias entre pares de classes (p.ex., a média das distâncias entre todos os pares de classes, a mínima entre todas as distâncias entre pares, etc). A distância entre duas classes pode, por sua vez, ser escolhida entre várias opções, algumas das quais podem ser vistas no capítulo sobre Medidas de Distância.

Em outra abordagem, utilizam-se as matrizes de espalhamento intra e entre classes (W e B) apresentadas no capítulo sobre Medidas de Distância. Pode-se então definir separabilidade entre classes como $tr(B)/tr(W)$, $tr(W^{-1}B)$, $|W+B|/|W|$, propostas estas que são uma extensão da distância entre 2 classes (vide capítulo sobre Medidas de Distância).

Ainda uma outra abordagem é o emprego de medidas probabilísticas de distância entre 2 classes como por exemplo a distância de Bhattacharyya (vide capítulo sobre Medidas de Distância), sendo necessário optar por alguma generalização para c classes. Uma possível generalização é a média das distâncias entre todos os pares de classes, média esta que pode ser ponderada pelo produto das probabilidades (ou frequências relativas) de cada par de classes.

Uma visão probabilística um pouco diferente nos sugere o emprego de uma medida de separabilidade baseada em entropia. Isto advém das seguintes

considerações:

1 se $P(\omega_i|\underline{x}) = 1/c$ para $i = 1, 2, \ldots, c$, tem-se a maior incerteza possível na escolha da classe a ser associada a $\underline{x}$, o que significa classes com grande sobreposição (pouca separabilidade). Neste caso resulta a máxima probabilidade de erro possível, igual a (c-1)/c.

2 se $P(\omega_k|\underline{x}) = 1$ para um dado k, não há incerteza alguma quanto à classificação a ser dada ao $\underline{x}$ e portanto a probabilidade de erro é nula. A separabilidade entre as classes neste ponto do espaço é total.

3 dos ítens 1 e 2 conclui-se que, fixado um $\underline{x}$, a situação mais interessante é quando as probabilidades a posteriori das classes são bem díspares, favorecendo em particular uma classe. No contexto da teoria da informação, para mínima sobreposição de classes (máxima separabilidade), a entropia

$$H(\underline{x}) = -\sum_{i=1}^{c} P(\omega_i|\underline{x}) \log_2 P(\omega_i|\underline{x}) \qquad (2)$$

deve ser a mínima possível (idealmente deve ser nula). Caso nos fixássemos apenas em um único padrão fornecido, o melhor conjunto de atributos seria aquele que minimizaria $H(\underline{x})$. Como se deseja selecionar atributos não para um único padrão dado, deve-se utilizar como quantificador de grau de sobreposição entre classes a média de $H(\underline{x})$:

$$H = \int_{\Omega} H(\underline{x})\, p(\underline{x})d\underline{x} = -\int_{\Omega} \sum_{i=1}^{c} P(\omega_i|\underline{x}) \log_2 P(\omega_i|\underline{x}) p(\underline{x})d\underline{x} \qquad (3)$$

Portanto, dentro do contexto em discussão, o objetivo de um algoritmo de seleção de atributos seria o de minimizar o quantificador H apresentado em (3).

Quando não se conhecem as formas das funções densidade de probabilidade condicionadas a cada classe deve-se utilizar como critério para seleção de atributos medidas que não exijam a estimação das funções densidade. Isto descarta as medidas probabilísticas de distância ou sobreposição e no caso de se optar pela minimização da taxa de erro é mais recomendável utilizar

uma estimativa baseada em testes empíricos do classificador ao invés de baseada na expressão (1) para a taxa de erro. Alguns trabalhos na literatura têm utilizado as expressões de distância probabilística derivadas para o caso Gaussiano mesmo quando nada se sabe sobre as distribuições.

Quando as distribuições têm forma conhecida, a escolha de um critério para seleção de atributos deve recair em um que utilize todo o conhecimento probabilístico, pois com isto deve-se obter melhores resultados. Ben-Bassat (1982) alerta para o uso da probabilidade de erro pois este critério, por vezes, apresenta baixa sensibilidade.

ALGORITMOS DE BUSCA: UMA VISÃO GERAL

O problema de efetuar a seleção dos d melhores atributos utilizando uma busca exaustiva dentre um conjunto de η atributos ($d<\eta$) normalmente não é factível em termos de tempo de computação uma vez que o número de combinações possíveis $\binom{\eta}{d}$ atinge facilmente valores grandes. Por exemplo se $\eta = 50$ e $d = 10$ deve-se testar aproximadamente 10^{10} conjuntos de 10 atributos. Em um problema menor, como da classificação dos potenciais de unidades motoras, poderíamos ter, por exemplo, $\eta = 15$ e $d = 5$, resultando em 3003 possíveis conjuntos de 5 atributos. Como a quantidade de cálculos para cada subconjunto de atributos a ser testado é enorme, raramente será viável efetuar uma busca exaustiva (quando d (ou η-d) é muito pequeno esta pode ser mais eficiente). Há entretanto um algoritmo, denominado "branch and bound" (Narendra e Fukunaga, 1977), que implicitamente analisa todos os possíveis conjuntos de d atributos sem efetuar a busca exaustiva. Excetuando-se este algoritmo ótimo de busca, os restantes são sub-ótimos, não garantindo portanto a otimalidade da solução, embora seus atrativos em termos de tempo de

computação os torne importantes em aplicações práticas.

Como os algoritmos de busca são independentes do critério de otimização adotado, estes serão descritos adotando o critério genérico da maximização de uma função J(.), a princípio arbitrária. Os algoritmos de busca podem ser do tipo "bottom-up" em que se parte de 1 atributo e novos atributos são adicionados até se atingir o número desejado destes (ou até satisfazer algum critério de parada), ou do tipo "top-down" em que se parte de todos os atributos e vão-se descartando atributos até se chegar ao número desejado destes (ou até se satisfazer um critério de parada).

ALGORITMO ÓTIMO DE BUSCA

Descreveremos brevemente o algoritmo "branch and bound" de Narendra e Fukunaga, (1977).

Parte-se do conjunto de η atributos $\{z_1, z_2, \ldots z_\eta\}$, supostos numéricos:

$$Z = \left\{ z_i \mid i=1, 2,\ldots, \eta \right\} \qquad \eta \in \mathbb{Z}^+$$

Os atributos são ordenados arbitrariamente e serão daqui para diante referidos pela sua ordem $\gamma_j = 1, 2,\ldots, \eta$. Por exemplo, Z poderia ser um conjunto contendo os atributos: duração, amplitude de pico, área, tempo de subida, com os coeficientes γ_j tomando valores 1, 2, 3 ou 4.

Denota-se por $\Gamma_{\bar{k}}$ um conjunto de $\bar{k} = \eta - k$ inteiros $\gamma_1, \ldots, \gamma_{\bar{k}}$, onde cada γ_j pode assumir um valor de 1 a η . $\Gamma_{\bar{k}}$ indica o conjunto dos atributos que são descartados para a obtenção do sub-conjunto de k atributos de interesse. Novamente, a ordem de escolha dos atributos é indiferente e para simplificar pode-se ter

$$\gamma_1 < \gamma_2 < \ldots < \gamma_{\bar{k}} \tag{4}$$

A função critério é denotada $J_{\bar{k}}(\gamma_1, \gamma_2, \ldots, \gamma_{\bar{k}})$, e deve ser maximizada para se obter o sub-conjunto ótimo de k atributos, descartanto-se os $\bar{k} = \eta - k$ atributos $\gamma_1^*, \gamma_2^*, \ldots \gamma_{\bar{k}}^*$

$$J_{\bar{k}}\left(\gamma_1^*,\ldots,\gamma_{\bar{k}}^*\right) = \max_{\gamma_1,\ldots\gamma_{\bar{k}}} J_{\bar{k}}\left(\gamma_1,\ldots\gamma_{\bar{k}}\right)$$

Uma restrição de monotonicidade é imposta à função critério (que é satisfeita por várias funções critério) no sentido que a mesma não deve diminuir de valor (em geral irá aumentar) quando se elimina um elemento de um dado conjunto de atributos descartados:

$$J_1(\gamma_1') \geq J_2(\gamma_1', \gamma_2') \geq \ldots \geq J_{\bar{k}}(\gamma_1', \ldots, \gamma_{\bar{k}}') \qquad (5)$$

onde a linha em γ_i' indica um particular valor fixo. Em outras palavras, esta condição significa que um dado conjunto de atributos não pode ser melhor que um conjunto maior que o contém. Deve-se acrescentar que embora várias funções critério satisfazem analiticamente esta condição, na prática pode acontecer que devido a erros de estimação (por exemplo de funções densidade de probabilidade) a desigualdade (5) não valha na prática. Esta não monotonicidade fará com que o algoritmo "branch and bound" deixe de ser ótimo.

⊙⊙⊙⊙⊙ Exemplo ilustrativo do algoritmo de busca. Suponhamos que escolheram-se 5 atributos para representar cada potencial de ação de unidade motora: pico positivo máximo (z_1), pico negativo mínimo (máximo em módulo) (z_2), duração (z_3), área do módulo (z_4) e energia (z_5). Suponhamos que deseja-se selecionar o melhor sub-conjunto de k = 2 atributos do conjunto de η = 5 atributos fornecido seguindo a condição (4) .

Uma árvore pode ser construída (Fig.1) em que o primeiro nó (no topo) corresponde a nenhum atributo descartado. Deste nó se geram outros nós correspondendo a 1 atributo descartado. Neste nível, os nós (candidatos a descartamento de 1 único atributo) começam do número 1 e crescem, para satisfazer (4). Foi necessário ir até o nó com o atributo número 3 para se poder

cobrir o caso de descartar a tripla de atributos z_3,z_4 e z_5. Para o próximo nível, geram-se nós sucessores de forma a poder chegar ao último (3^o) nível com todas as combinações possíveis. A construção desta árvore é feita mais facilmente da direita (região menos densa) para a esquerda (região mais densa). Inicia-se portanto com o ramo 345. Do nó 2 no nível γ_1 geram-se 245 e 23 . Do nó 3 no nível γ_2 definem-se os ramos 235 e 234 . Do nó 1 no nível γ_1 geram-se 145, 13 e 12. Do nó 3 no nível γ_2 definem-se os ramos 134 e 135 . Finalmente, do nó 2 no nível γ_2 definem-se os ramos 123, 124 e 125. Aqui foi explicada a construção completa da árvore, o que equivale a uma busca exaustiva, o que certamente não nos interessa. Na realidade a construção da árvore tem que ser feita baseada na função critério J(.). Parte-se do nó de origem (nível γ_o), gera-se o ramo mais à direita (345) e calcula-se $\alpha \triangleq J_3(3, 4, 5)$. Volta-se por este ramo ao nó mais próximo em que há ramificação e gera-se o próximo ramo mais à direita (2). Calcula-se $J_1(2)$, e se $J_1(2) \leq \alpha$ percebe-se de (5) que qualquer sub-conjunto de atributos em que se descarta z_2 resultará em um valor de J(.) menor que (ou, no melhor caso, igual a) o valor α do ramo 345. Neste caso toda a seção da árvore originada do nó 2 no nível γ_1 não necessita ser testada por se garantir que o sub-conjunto de descartamento ótimo não está ali situado. Volta-se então para o nó mais próximo em que há ramificação e gera-se o próximo ramo mais à direita (1) (ainda não analisado) calculando-se $J_1(1)$. Se $J_1(1) < \alpha$ então pode-se abandonar a busca em toda a seção da árvore gerada a partir do nó 1 no nível γ_1 e, neste exemplo, a busca estaria encerrada com a seleção dos atributos z_1 e z_2, uma vez que os atributos eliminados teriam sido z_3 , z_4 e z_5 .

Retornando a análise ao nó 2 do nível γ_1 , caso $J_1(2) > \alpha$, gera-se o ramo mais à direita (245) e calcula-se $J_3(245)$. Se $\alpha < J_3(245)$ então o novo α será $J_3(245)$ e o sub-conjunto de atributos de melhor desempenho até o momento passa a ser $\{z_1,z_3\}$ ao invés do que era de início $\{z_1,z_2\}$. Continua-se

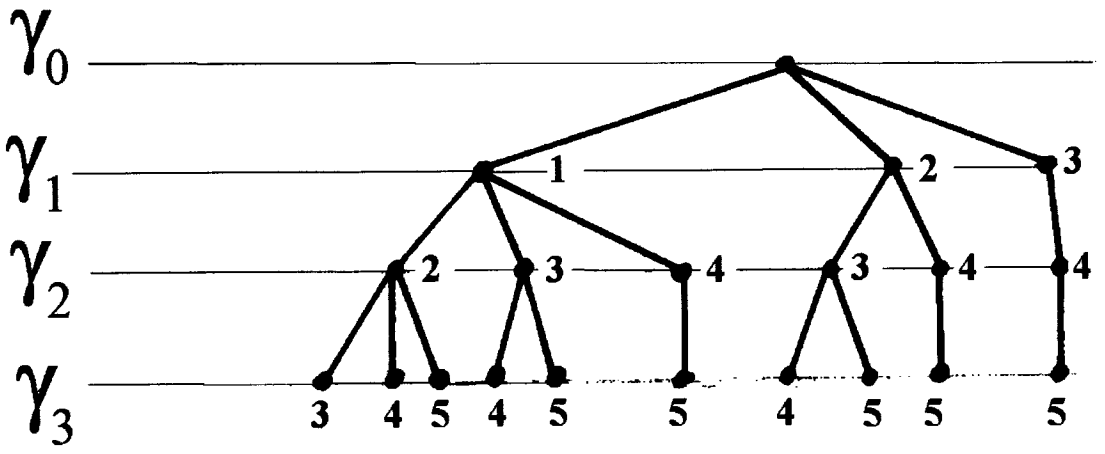


Fig. 1 – Exemplo de árvore para o algoritmo ótimo de busca.

a busca voltando ao nó mais próximo com ramificação e gerando o próximo ramo mais à direita. Se $\alpha > J_3(245)$, volta-se do nó terminal do ramo 245 até o nó mais próximo em que há ramificação e repete-se o procedimento.

ooooo

Por que se utiliza um algoritmo com os conjuntos de atributos a serem descartados e não com os conjuntos de atributos selecionados? A razão se deve ao fato de normalmente se desejar uma redução razoável no número de atributos, com $d < \eta - d$. Com isto, uma árvore como da Fig.1 é mais interessante do que outra em que os ramos são de sub-conjuntos de atributos selecionados pois a eliminação de uma seção de árvore a partir de um dado nó irá resultar em um número bem maior de combinações que não precisam mais ser testadas.

A idéia básica por detrás do algoritmo "branch and bound" foi apresentada através de um exemplo ilustrativo. Maiores detalhes podem ser encontrados em Narendra e Fukunaga (1977) e DeVijver e Kittler (1982). Na primeira destas duas referências também se apresenta um método de cálculo recursivo para funções critério quadráticas o que é muito útil pois diminui muito o volume dos cálculos necessários ao se descartar um novo atributo.

ALGORITMOS SUB-ÓTIMOS DE BUSCA

Há vários algoritmos sub-ótimos de busca, baseados em idéias intuitivas, conforme veremos a seguir. Apesar de terem um desempenho de busca geralmente bastante inferior ao algoritmo ótimo, os métodos sub-ótimos são populares devido à sua eficiência quanto ao tempo de computação.

i Algoritmo de busca utilizando os melhores atributos individuais

Parte-se medindo o desempenho individual de cada atributo quanto à

discriminação entre classes. Selecionam-se os d melhores atributos. Fica claro que o adjetivo "melhor" está relacionado a alguma função critério J(.). O desempenho deste algoritmo é normalmente bastante ruim pois, em geral, o desempenho individual dos atributos não espelha a capacidade de discriminação de conjuntos de atributos.

ii Seleção sequencial direta (ou "para frente")

É um algoritmo "bottom-up" em que se acrescenta 1 atributo por vez, sequencialmente. Isto é, seleciona-se o melhor atributo z_j dentre os η iniciais. Em seguida, verifica-se qual o atributo z_k dentre os η-1 restantes que resulta em melhor discriminação entre classes quando tomado junto com o z_i obtido anteriormente. O algoritmo funciona portanto formando conjuntos de atributos aninhados ("nested"). Em geral o desempenho deste método será superior ao do anterior pois, exceto para o primeiro atributo, ele leva em conta o desempenho de sub-conjuntos de atributos utilizados simultaneamente. O desempenho deste algoritmo é limitado pelo fato de que uma vez selecionado certo atributo ele não pode ser retirado posteriormente, quando eventualmente sua importância deixou de ser grande devido à inclusão de outros atributos. Isto é mais sentido nos primeiros passos do algoritmo, podendo a seleção do 1° atributo desviar totalmente a busca para um sub-conjunto de atributos de qualidade sofrível.

Uma generalização deste algoritmo é obtida selecionando-se θ atributos por vez (note que os conjuntos de atributos formados a cada passo continuam aninhados). O desempenho irá melhorar consideravelmente, às custas de um tempo de computação consideravelmente maior. Uma abordagem diferente poderia ser a de iniciar a seleção com θ atributos por vez reduzindo para a etapa seguinte a seleção para θ-1 atributos por vez e assim por diante (obviamente $\theta-1\geq 1$). Uma variante seria a utilização de outra regra de decréscimo do número de atributos selecionados simultaneamente para as eta-

pas seguintes. Poder-se-ia ainda selecionar θ atributos simultaneamente para o 1^{o} passo e passar a adicionar 1 por vez para os passos seguintes.

iii Seleção sequencial reversa (ou "para trás")

É um algoritmo "top-down" em que se inicia com os η atributos e vai se descartando 1 atributo por vez até se chegar ao número d de atributos desejados. Um determinado atributo é descartado quando a sua ausência otimiza a função critério em relação ao descartamento dos outros atributos descartáveis, isto é, descarta-se o atributo que causa menor queda no valor da função critério. Novamente os conjuntos de atributos selecionados em cada passo serão aninhados. Esse algoritmo deve, via de regra, prover melhor escolha do conjunto de atributos do que o algoritmo direto pois parte-se da totalidade dos atributos analisados conjuntamente, podendo-se desta forma, já desde o primeiro passo, trabalhar com uma análise coletiva dos atributos. Uma característica interessante deste algoritmo é que obtém-se a evolução do valor ótimo da função critério ao se passar de η para d atributos. Isto pode inclusive ser útil quando não se deseja escolher a priori o número final d de atributos pois basta se estabelecer um critério de parada como por exemplo a 80% do valor da função critério para todos os η atributos. Ainda quando comparado com o algoritmo direto, o algoritmo em discussão exige maior tempo de computação uma vez que neste trabalha-se em espaços de dimensões de d até η enquanto naquele de 1 até d.

Novamente pode-se generalizar o algoritmo descartando-se μ atributos por vez com a mesma vantagem e desvantagem já mencionadas no algoritmo direto.

iv Seleção direta-reversa

Esta técnica visa evitar o aninhamento dos conjuntos de atributos gerados ao longo do procedimento. Em um dado passo do algoritmo acrescentam-se θ atributos, um por vez, de acordo com o algoritmo de seleção sequencial

direta. Do conjunto de atributos resultante descartam-se μ atributos, um por vez, de acordo com o algoritmo de seleção sequencial reversa. Caso se deseje um algoritmo "bottom-up" parte-se com nenhum atributo e utiliza-se $\theta>\mu$. Para um algoritmo "top-down" parte-se dos η atributos, toma-se $\theta<\mu$ e aplica-se primeiro o algoritmo reverso (seria um algoritmo "reverso-direto").

Uma generalização importante deste algoritmo consta na utilização dos algoritmos generalizados direto e reverso a cada passo, ou seja, utilizando-se θ ou μ atributos de uma só vez, ao invés de se tomar um atributo por vez.

Em todos os métodos de seleção de atributos apresentados é importante verificar se a função critério pode ser calculada recursivamente de modo que a cada passo não seja necessário efetuar os cálculos a partir da estaca zero mas apenas se façam alguns cálculos para atualização (vide Narendra e Fukunaga, 1977; DeVijver e Kittler, 1982).

EXTRAÇÃO DE ATRIBUTOS

INTRODUÇÃO

As duas técnicas mais conhecidas para redução da dimensionalidade em problemas de estatística multivariada e reconhecimento de padrões são a de componentes principais, também conhecida como técnica da expansão de Karhunen-Loève (para o caso discreto), e a de análise discriminante, criada por Fisher e já estudada em um capítulo anterior. Apresenta-se a seguir uma descrição resumida da análise de componentes principais, podendo o leitor interessado consultar, por exemplo, Fukunaga (1990), Johnson & Wichern (1988), Jolliffe (1986), entre outros, para maiores detalhes.

ANÁLISE DE COMPONENTES PRINCIPAIS

No contexto de reconhecimento de padrões, temos cada padrão sendo descrito por um vetor de η atributos. Procuramos uma transformação linear do espaço vetorial dos η atributos originais para outro espaço, com dimensão d ($d<\eta$), de tal forma que se minimize o erro médio quadrático de representação dos padrões. No caso de sinais aleatórios, o vetor $\underline{x}$ teria como elementos as amostras do sinal x(n) e o objetivo seria conseguir uma representação com menos que η coeficientes, obtendo-se assim uma compressão de dados.

Se tomarmos η vetores linearmente independentes (L.I.), indicados por $\underline{\phi}_i(\eta x1)$, com $i = 1,\ldots,\eta$, formando portanto uma base em $\mathbb{R}^{\eta}$, podemos representar um dado vetor $\underline{x}$, sem erro, como uma combinação linear dos $\underline{\phi}_i$

$$\underline{x} = \sum_{i=1}^{\eta} y_i \underline{\phi}_i = \Phi \underline{y} \qquad (1)$$

onde $\Phi = [\underline{\phi}_1, \underline{\phi}_2 \ldots \underline{\phi}_\eta]$, e $\underline{y} = [y_1, y_2 \ldots y_\eta]^T$, que é o vetor cujas componentes são a representação de $\underline{x}$ na base ϕ.

Para o caso populacional, temos

$$\underline{X} = \Phi \underline{Y} \qquad (2)$$

onde $\underline{X}$ e $\underline{Y}$ são vetores aleatórios e Φ é uma matriz determinística não singular.

Nos interessaremos por uma base Φ ortonormal:

$$\underline{\phi}_i^T \cdot \underline{\phi}_j = \delta_{ij} = \begin{cases} 1 & \text{p/ } i=j \\ 0 & \text{c.c.} \end{cases} \quad \text{ou} \quad \Phi^T\Phi = I = \Phi\,\Phi^T \qquad (3)$$

Neste caso, dado um $\underline{x}$, obtém-se os y_i da expressão (1):

$$y_i = \underline{x}^T \underline{\phi}_i = \underline{\phi}_i^T \underline{x} \ , \ i = 1,\ldots,\eta \qquad (4)$$

ou ainda

$$\underline{y}=\Phi^T\underline{x} \qquad (5)$$

Para o caso populacional temos

$$Y_i = \underline{X}^T\underline{\phi}_i = \underline{\phi}_i^T\underline{X} \quad , \ i=1,\ldots,\eta \qquad (6)$$

ou ainda

$$\underline{Y} = \Phi^T\underline{X} \qquad (7)$$

De (7) conclui-se que $E\left(\underline{Y}\ \underline{Y}^T\right) = R_y = E\left(\ \Phi^T\underline{X}\ \underline{X}^T\Phi\ \right) = \Phi^T\ R_x\ \Phi$.

Deseja-se obter uma boa representação do vetor aleatório $\underline{x}$ (que pode, por exemplo, constar das η amostras de um sinal aleatório x(i), i=1,...,η) através da utilização de um número bem menor de coeficientes. Na formulação por expressão em vetores ou funções ortonormais desejamos utilizar um número de funções ou vetores de base $\underline{\phi}_i$ menor que η.

Suponha que utilizemos $m < \eta$ funções de base, obtendo uma aproximação $\underline{X}'$ para $\underline{X}$:

$$\underline{X}' = \sum_{i=1}^{m} Y_i \underline{\phi}_i \tag{8}$$

O erro de representação é denotado $\Delta\underline{X}$ e definido como $\Delta\underline{X} = \underline{X} - \underline{X}'$, sendo igual a

$$\Delta\underline{X} = \sum_{i=m+1}^{\eta} Y_i \underline{\phi}_i \tag{9}$$

Um erro médio quadrático é

$$\varepsilon^2 = E\left[\|\Delta\underline{X}\|_2^2\right] = E\left[\Delta\underline{X}^T \Delta\underline{X}\right] \tag{10}$$

ou seja

$$\varepsilon^2 = E\left[\sum_{i=m+1}^{\eta} \sum_{j=m+1}^{\eta} Y_i Y_j \underline{\phi}_i^T \underline{\phi}_j\right] = \sum_{i=m+1}^{\eta} E\left(Y_i^2\right) \tag{11}$$

Este erro deve ser minimizado através da escolha de uma "boa" base Φ.

De (6) em (11) temos:

$$\varepsilon^2 = \sum_{i=m+1}^{\eta} \underline{\phi}_i^T R_x \underline{\phi}_i \tag{12}$$

onde R_x é a matriz de auto correlação de $\underline{X}$ ($R = E[\underline{X}\underline{X}^T]$) e os $\underline{\phi}_i$ são ortonormais.

Desejamos pois minimizar (12) em relação a $\underline{\phi}_i$.

Sabemos que R_x é não negativa definida (pois, definindo a variável aleatória $Z = \underline{a}^T \underline{X}$, $E(Z^2) \geq 0 \;\therefore\; E(\underline{a}^T \underline{X}\, \underline{X}^T \underline{a}) \geq 0 \;\therefore\; \underline{a}^T R_x \underline{a} \geq 0$, $\forall\, \underline{a}$), e portanto para minimizar ε^2 devemos minimizar cada termo $\underline{\phi}_i^T R_x \underline{\phi}_i$, lembrando que a otimização deve ser para $\underline{\phi}_i^T \underline{\phi}_i = 1$. Para impor esta restrição usamos a técnica de multiplicador de Lagrange. A restrição $\underline{\phi}_i^T \underline{\phi}_j = 0$ p/ $i \neq j$, já usada para se chegar a (12), decorrerá automaticamente do fato dos autovetores de R_x serem ortogonais. Devemos ter

$$\frac{\partial}{\partial \underline{\phi}_i}\left(\underline{\phi}_i^T R_x \underline{\phi}_i - \lambda_i(\underline{\phi}_i^T \underline{\phi}_i - 1)\right) = \underline{0} \tag{13}$$

o que é satisfeito para $\underline{\phi}_i^*$ tal que

$$2 R_x \underline{\phi}_i^* - 2 \lambda_i \underline{\phi}_i^* = \underline{0}$$

ou seja

$$R_x \underline{\phi}_i^* = \lambda_i \underline{\phi}_i^* \tag{14}$$

e portanto os $\underline{\phi}_i^*$ são os autovetores da matriz de auto correlação R_x. Os autovalores são os λ_i , com $\lambda_i \geq 0$, pois R_x é positiva semi-definida.

O erro médio quadrático ótimo ε^{*2} é

$$\varepsilon^{*2} = \sum_{i=m+1}^{\eta} \lambda_i \tag{15}$$

pois, de (14): $\underline{\phi}_i^T R_x \underline{\phi}_i = \lambda_i$. Deve-se notar que os autovalores λ_i que entram no cômputo de ε^{*2} são os associados aos autovetores que <u>não</u> são utilizados na expansão (8). Se desejamos que ε^{*2} seja o mínimo possível, então os λ_i em (15) devem ser os (η-m) menores autovalores da matriz R_x. Portanto na expansão para $\underline{X}'$ utilizam-se os m autovetores associados aos m maiores autovalores de R_x . Como a matriz R_x é simétrica, seus autovetores serão ortogonais/ortonormais. Os elementos de $\underline{Y}$ são chamadas de componentes principais, embora por vezes as direções $\underline{\phi}_i$, i=1,2,...,m recebam este nome.

Quanto menor for m, maior será o erro ε^{*2} mas também será maior a taxa de compressão η/m. Lembrar que, dado um vetor $\underline{x}$ de dimensão η, ou um sinal x(n) com η amostras, ele será representado por apenas m coeficientes y_i (i=1,...m). Para o caso de sinais, a Fig. 1 mostra um exemplo para o caso de transmissão através de algum canal de comunicações ou de gravação em uma memória.

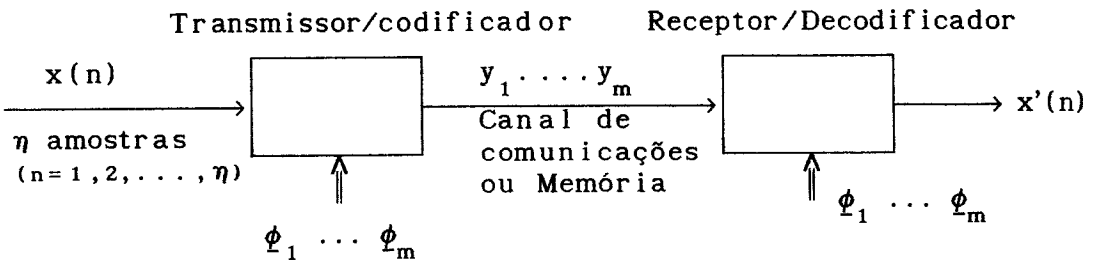


Fig. 1- Exemplo da aplicação da técnica de componentes principais para compressão de sinais aleatórios.

É fácil ver que os coeficientes da expansão têm auto-correlação nula:

$$E(Y_i Y_j) = E\left(\underline{\phi}_i^T \; \underline{X} \, \underline{X}^T \, \underline{\phi}_j\right) = \underline{\phi}_i^T \, R_X \, \underline{\phi}_j = \lambda_j \, \delta_{ij}$$

Para termos um critério para a escolha do <u>número</u> de coeficientes (vetores de base, funções ou sinais de base) trabalharemos com o conceito de energia, uma vez que os vetores ou sinais são de duração finita. A energia média da família de vetores ou sinais aleatórios $\underline{X}$ de η amostras é:

$$\bar{E}_x = E(X_1^2) + E(X_2^2) + \ldots + E(X_\eta^2) = tr[E(\underline{X}\underline{X}^T)] = tr(R_x) = \sum_{j=1}^{\eta} \lambda_j \qquad (16)$$

Para o sinal $\underline{X}'$ temos

$$\bar{E}_{x'} = E\left[\underline{X}'^T \underline{X}'\right] = E\left[\underline{Y}_r^T \, \Phi_I^T \, \Phi_I \, \underline{Y}_r\right] \qquad (17)$$

onde $\underline{Y}_r = \left[Y_1 \ldots Y_m\right]^T$, $\Phi = [\Phi_I \Phi_{II}]$ é a matriz dos η autovetores de R_X, Φ_I é a matriz (ηxm) dos m autovetores de R_X associados aos m maiores autovalores e Φ_{II} é a matriz contendo os demais autovetores. Apenas como complemento temos que:

$$R_y = \Phi^T R_X \Phi = \begin{bmatrix} \Phi_I^T \\ \Phi_{II}^T \end{bmatrix} R_X \left[\Phi_I \; \Phi_{II}\right] = \left[\begin{array}{c|c} \Phi_I^T R_x \Phi_I & \Phi_I^T R_x \Phi_{II} \\ \hline \Phi_{II}^T R_x \Phi_I & \Phi_{II}^T R_x \Phi_{II} \end{array}\right] =$$

$$= \left[\begin{array}{ccc|cc} \lambda_1 & & 0 & & \\ & \lambda_2 & & 0 & \\ 0 & & \lambda_m & & \\ \hline & & & \lambda_{m+1} & 0 \\ 0 & & & 0 & \lambda_\eta \end{array}\right] \qquad (18)$$

onde $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m \geq \lambda_{m+1} \geq \ldots \geq \lambda_\eta$

De (18) conclui-se que $R_y = diag(\lambda_1 \ldots \lambda_\eta)$ e que portanto as componentes tem auto-correlação nula, como já visto acima.

Como $\Phi^T \Phi = I$ (ηxη), temos $\Phi_I^T \Phi_I = I$ (mxm) $= \Phi_I \Phi_I^T$. Retomando da derivação feita até a equação (17) temos:

$$\bar{E}_{x'} = E\left[\underline{Y}_r^T \underline{Y}_r \right] = E\left(\sum_{i=1}^{m} Y_i^2 \right) = \sum_{i=1}^{m} E\left[\underline{\phi}_i^T \underline{X} \underline{X}^T \underline{\phi}_i \right] = \tag{19}$$

$$= \sum_{i=1}^{m} \underline{\phi}_i^T R_x \underline{\phi}_i = \sum_{i=1}^{m} \lambda_i$$

Uma forma alternativa de derivar este resultado é :

$$\underline{Y} = \left[\underline{Y}_r^T \quad \underline{Y}_d^T \right]^T$$

$$E\left(\underline{Y} \underline{Y}^T \right) = R_y = \Phi^T R_x \Phi = \text{diag}\left(\lambda_1 \lambda_2 \ldots \lambda_\eta \right) =$$

$$= E\left\{ \begin{bmatrix} \underline{Y}_r \\ \underline{Y}_d \end{bmatrix} \begin{bmatrix} \underline{Y}_r^T & \underline{Y}_d^T \end{bmatrix} \right\} = \left[\begin{array}{c|c} R_{y_r} & R_{y_r y_d} \\ \hline R_{y_r y_d} & R_{y_d} \end{array} \right]$$

$$\therefore \quad R_{y_r} = \text{diag}\left(\lambda_1 \lambda_2 \ldots \lambda_m \right)$$

e como $E\left(\underline{Y}_r^T \underline{Y}_r \right) = \text{tr}\left(R_{y_r} \right)$, obtém-se (19).

Portanto a relação entre as energias médias de $\underline{X}'$ e $\underline{X}$ é um bom indicador da qualidade de reconstrução:

$$\gamma = \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{j=1}^{\eta} \lambda_j} \qquad (0 \leq \gamma \leq 1)$$

Escolhe-se o número m que forneça um γ desejado, por exemplo 0,85 ou 0,90. Este m passa a ser a dimensão d do espaço de atributos final no contexto de reconhecimento de padrões.

Em certas ocasiões, subtrai-se o vetor médio $\bar{\underline{x}}$ (ou vetor esperado μ_x, no caso populacional) e com isto a expressão (12) terá Σ_x (a matriz de covariância) ao invés da matriz de autocorrelação R_x. Com isto, o interesse passa a ser de representar variações em torno da média conhecida, obtendo-se os Y_i não-correlacionados (covariância mula).

Na prática, estima-se a matriz de auto correlação (ou de covariância) a partir dos dados para a seguir determinar seus autovalores e autovetores.

Do que foi exposto, deve-se notar que em nenhum instante foi mencionado que se trata de um problema de classificação em que há c classes. O fato é que a técnica de componentes principais não leva em conta se há 1 ou mais classes, pois ela toma o conjunto de todos padrões como sendo uma única grande classe. Desta forma, é fácil acontecer que as componentes principais não as mais adequadas para fins de classificação, como é ilustrado na Fig. 2. Neste exemplo bi-dimensional, a primeira componente principal aponta para a direção em que a variância global é maior, embora a componente número 2 seja a mais adequada para fins de classificação. Não se deve concluir deste exemplo, entretanto, que deve-se então selecionar as componentes menos importantes sob o ponto de vista de representação. Em nossos trabalhos com potenciais de unidades motoras, houve casos em que não foi possível encontrar uma combinação de componentes principais que fosse melhor do que as próprias coordenadas originais. Deve-se portanto utilizar esta técnica com cuidado quando a aplicação está relacionada com classificação de padrões.

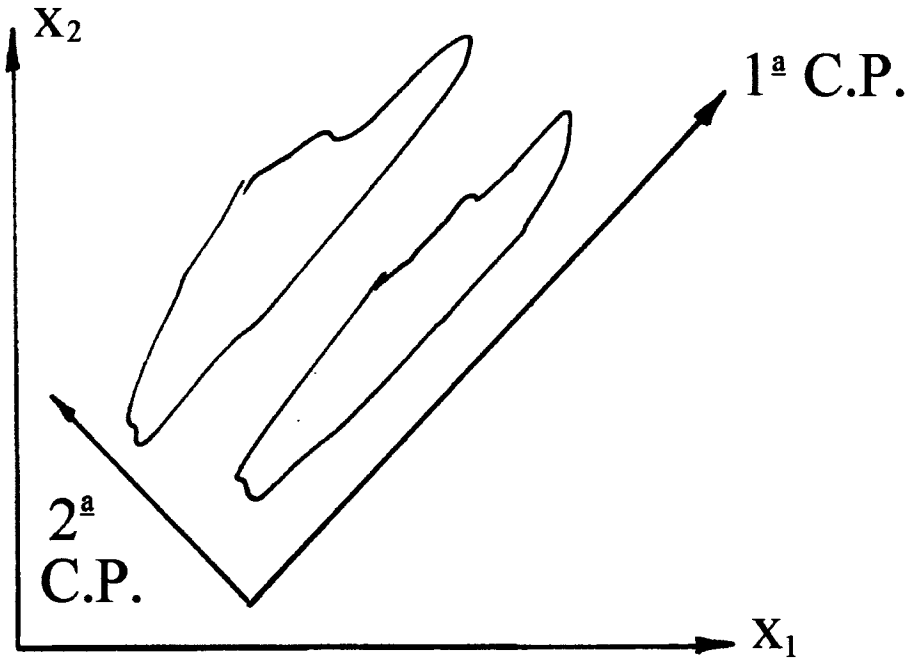


Fig. 2 – Ilustração de um caso em que a primeira componente principal não tem nenhum poder discriminatório.

ANÁLISE DE AGRUPAMENTOS

INTRODUÇÃO

Em várias aplicações práticas dispõe-se de um conjunto de dados ou amostras com classificação desconhecida. As metodologias já abordadas não são adequadas pois elas pressupõe que há um conjunto de amostras de treinamento com classificação conhecida.

Como não se conhece nada a priori, a tarefa é bastante difícil: a partir de amostras de um vetor aleatório $\underline{X}$ deseja-se decompor a sua função densidade de probabilidade misturada $p(\underline{x})$ em c densidades condicionais $p(\underline{x}|\omega_i)$:

$$p(\underline{x}) = \sum_{i=1}^{c} P_i \, p(\underline{x}|\omega_i) \overset{\Delta}{=} \sum_{i=1}^{c} f(\underline{x}|\omega_i) \qquad (1)$$

Pode-se estimar $p(\underline{x})$ a partir das amostras mas isto não é suficiente. Há uma série de métodos de agrupamento que formam uma abordagem indireta ao problema de se separar modas de $p(\underline{x})$. Estes métodos tentam particionar o conjunto de amostras de tal forma que elementos de um mesmo agrupamento apresentem grande similaridade e amostras de agrupamentos diferentes tenham baixa similaridade. Isto está (indiretamente) relacionado com a busca de modas em $p(\underline{x})$ uma vez que amostras próximas (com grande similaridade) provavelmente estarão associadas a uma mesma moda. As técnicas de agrupamento formam uma parte importante da sub-área de reconhecimento de padrões com aprendizado sem supervisão.

Há basicamente duas abordagens na análise de agrupamentos ("cluster analysis"):

i por partição, em que se emprega algum algoritmo iterativo para parti-

cionar o conjunto de amostras de forma a otimizar uma função-critério de agrupamento.

ii hierárquica, em que se procede passo a passo sem iteração, de um primeiro nível até o último, obtendo-se partições encadeadas.

Uma vez obtidos agrupamentos a partir de um conjunto de padrões S_N com classificação a priori desconhecida, qualquer padrão adicional pode ser classificado utilizando-se algum método de classificação baseado no conjunto S_N que agora será de padrões com classificação conhecida (embora nem todos padrões de S_N estarão com a classificação correta).

Antes de aplicar algum método de agrupamento, deve-se verificar se os dados precisam ser padronizados devido a escalas e/ou unidades diferentes. Isto equivale a utilizar, por exemplo, a distância de Pearson, ou ponderada, ao invés da distância Euclideana, muito embora normalmente implementam-se os algoritmos com uma distância pré-escolhida, normalmente a Euclideana, antes efetuando-se a padronização desejada.

MÉTODOS DE AGRUPAMENTO POR PARTIÇÃO

É dado o conjunto $S_N = \left\{ \underline{x}_1 , \underline{x}_2 , \ldots \underline{x}_N \right\}$ de vetores dx1 para serem classificados sem supervisão. Deseja-se obter K agrupamentos G_i, i = 1,..,K, cada um representado pelo seu vetor médio ou centróide $\overline{\underline{x}}_i$. Decidimos chamar de K o número de agrupamentos ao invés de c pois normalmente não sabemos o número de classes. Mesmo em um algoritmo de agrupamento que "tenta descobrir" o número de classes somente em casos bastante favoráveis é que se pode ter K = c . Sendo n_i a cardinalidade de G_i temos

$$\overline{\underline{x}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \underline{x}_{ij} \qquad \text{onde} \quad \underline{x}_{i,j} \in G_i \text{ , com } j = 1, 2, \ldots n_i \quad \text{e} \quad i = 1, 2, \ldots, K$$

e $N = \sum_{i=1}^{c} n_i$

Com a escolha do vetor médio como representativo de um dado agrupamento e utilizando-se distância Euclideana, a distância entre $\underline{x}$ e o agrupamento G_i é $d_2(\underline{x} , \overline{\underline{x}}_i)$. Para $\underline{x} \in G_i$ deve-se ter :

$$d_2(\underline{x} , \overline{\underline{x}}_i) < d_2(\underline{x} , \overline{\underline{x}}_k) \qquad \forall\ k \neq i \tag{3}$$

Indicaremos por $G_{N,K}$ uma possível partição do conjunto S_N em K agrupamentos:

$$G_{N,K} = \left\{ G_j \,\middle|\, G_i \cap G_j = \emptyset,\ \forall\ i \neq j \ \text{ e } \bigcup_{j=1}^{K} G_j = S_N \right\}$$

A cardinalidade do conjunto de todas as partições possíveis de S_N em K agrupamentos é dada pelos números de Stirling do 2º tipo (Feller, 1959, pg. 58)

$$\frac{1}{K!} \sum_{i=1}^{K} (-1)^{K-i} \binom{K}{i} (i)^N \tag{4}$$

e é fácil ver que o número de partições cresce vertiginosamente com N e/ou K. Para N >> K pode-se utilizar a aproximação por $K^N/(K!)$ Por exemplo, para N = 10 e K = 4 o número de partições possíveis é 34105 e para N = 19 e K = 4 o número sobe para 11.259.666.000. Disto se conclui que não é factível uma busca entre todas as partições possíveis, ainda mais nos casos em que não se sabe a priori o número de classes existentes.

Entretanto, quer se faça uma busca ótima ou sub-ótima de agrupamentos, é necessário se adotar um índice de qualidade, às vezes chamado de função critério ou função objetivo, para permitir a avaliação de cada partição investigada. O objetivo na busca é minimizar, ou eventualmente maximizar, o índice adotado. Apresenta-se a seguir alguns índices de qualidade importantes.

i Índice quadrático

Inicialmente definimos a variação (quadrática) ou espalhamento no agrupamento G_i como

$$J_i(G_{N,K}) = \sum_{j=1}^{n_i} d_2^2(\underline{x}_{ij}, \overline{\underline{x}}_i) = \sum_{j=1}^{n_i} (\underline{x}_{ij}-\overline{\underline{x}}_i)^T(\underline{x}_{ij}-\overline{\underline{x}}_i) \; ; \quad i = 1,2,...K \tag{5}$$

Em (5) poder-se-ia utilizar a distância de Mahalanobis ao invés da Euclideana.

A variação intra-agrupamentos total, levando em conta K agrupamentos, é

$$J(G_{N,K}) = \sum_{i=1}^{K} J_i(G_{N,K}) \tag{6}$$

Esta função critério é a soma total dos quadrados das distâncias de cada vetor ou padrão $\underline{x}_i$ ao centróide do agrupamento a que pertence. O centróide $\underline{x}_i$ do agrupamento G_i é visto como o melhor representante de G_i, e $J(G_{N,K})$ mede o erro quadrático de se representar as N amostras de S_N pelas K médias ou centróides $\overline{\underline{x}}_1, \overline{\underline{x}}_2, \ldots \overline{\underline{x}}_K$ dos K agrupamentos.

A solução ótima $G^*_{N,K} = \left\{ G^*_j \mid G^*_i \cap G^*_j = \emptyset, \; \forall \; i \neq j \; \text{ e } \bigcup_{j=1}^{K} G^*_j = S_N \right\}$ é aquela que minimiza $J(G_{N,K})$ para um certo K pré-fixado

$$J(G^*_{N,K}) = \min_{G_{N,K}} J(G_{N,K}) \tag{7}$$

Este critério tende a fornecer agrupamentos compactos, preferencialmente esféricos, o que em vários casos é indesejável. Um agrupamento com poucos elementos pode não ser detectado por este critério pois seus elementos poderão ser associados a outros para formar um agrupamento mais similar aos demais.

Pode-se obter para (5) a expressão alternativa:

$$J_i(G_{N,K}) = \frac{1}{2n_i} \sum_{j=1}^{n_i} \sum_{k=1}^{n_i} d_2^2(\underline{x}_{ij}, \underline{x}_{ik}) \tag{8}$$

Demonstração: Basta tomar a dupla somatória com a expansão da expressão $(\underline{x}_{ij}-\underline{x}_{ik})^T(\underline{x}_{ij}-\underline{x}_{ik})$.

Deve-se notar que (8) exprime J_i agora em função das distâncias entre vetores do i-ésimo agrupamento e não mais em função das distâncias de cada vetor ao centróide do agrupamento.

Pode-se exprimir o espalhamento global como

$$J = \frac{1}{2} \sum_{i=1}^{K} n_i \overline{d_i^2} \tag{9}$$

onde

$$\overline{d_i^2} = \frac{1}{n_i^2} \sum_{j=1}^{n_i} \sum_{k=1}^{n_i} d_2^2(\underline{x}_{ij},\underline{x}_{ik}) \tag{10}$$

Este $\overline{d_i^2}$ é a distância (Euclideana) quadrática média entre vetores do agrupamento G_i . Poder-se-ia propor outros índices de qualidade diferentes de (9), partindo-se de alguma outra definição para $\overline{d_i^2}$, por exemplo empregando uma função de dissimilaridade que não a Euclideana na expressão (10). Um outro exemplo seria selecionar uma medida de distância (Euclideana, etc) e definir $\overline{d_i^2}$ como a máxima distância entre 2 quaisquer vetores do agrupamento G_i .

ii Variabilidade intra-agrupamentos baseado no traço de W

Define-se uma matriz de somas de quadrados e produtos (SQP), ou matriz de espalhamento intra-agrupamentos, semelhante ao feito em análise discriminante:

$$W = \sum_{k=1}^{K} W_k = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (\underline{x}_{ki} - \overline{\underline{x}}_k)(\underline{x}_{ki} - \overline{\underline{x}}_k)^T \tag{11}$$

onde a matriz de SQP W_k é específica do agrupamento G_k .

A solução ótima é aquela que minimiza o traço de W

$$tr(W) = \sum_{k=1}^{K} tr(W_k) = \sum_{k=1}^{K} \sum_{i=1}^{n_k} tr\left[(\underline{x}_{ki} - \overline{\underline{x}}_k)(\underline{x}_{ki} - \overline{\underline{x}}_k)^T\right] =$$

$$= \sum_{k=1}^{K} \sum_{i=1}^{n_k} (\underline{x}_{ki} - \overline{\underline{x}}_k)^T (\underline{x}_{ki} - \overline{\underline{x}}_k) = \sum_{k=1}^{K} \sum_{i=1}^{n_k} d_2^2 (\underline{x}_{ki}, \overline{\underline{x}}_k) \qquad (12)$$

de onde vemos que este índice é equivalente ao quadrático já visto.

iii Variabilidade entre-agrupamentos baseado no traço de B

Define-se uma matriz de SQP B entre agrupamentos semelhantemente ao feito em análise discriminante :

$$B = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (\underline{x}_k - \overline{\underline{x}})(\underline{x}_k - \overline{\underline{x}})^T \qquad (13)$$

onde $\overline{\underline{x}} = \frac{1}{K} \sum_{i=1}^{K} \underline{x}_i$

Como T = B+W (expressão (34) no capítulo sobre o Discriminante Linear de Fisher) e T é independente da particular partição, temos que a maximização de tr(B) equivale à minimização de tr(W) e portanto a otimização de tr(B) e do índice quadrático fornece o mesmo resultado.

iv Variabilidade intra-agrupamentos baseado no determinante de W

A solução ótima é a que minimiza $|W|$ = det(W) que é equivalente à minimização de $|W|/|T|$ que é a chamada estatística lambda de Wilks (a equivalência segue do fato que T é independente de qualquer partição dos dados). Este critério não é adequado quando N-K < d ou se os padrões pertencem a um sub-espaço do espaço de atributos pois então W é singular. Estas duas condições raramente irão acontecer na prática. Caso aconteça W ser singular é necessário reduzir a dimensionalidade quer por seleção ou extração de atributos.

Notar que a minimização de $|W|$ é equivalente à maximização de $|T|/|W|$ que é por vezes utilizado.

Este critério tende a dar bons resultados quando os formatos dos agrupamentos são iguais ou seja ele supõe que as classes tem matrizes de covariância iguais. Vide critério vii .

v Variabilidade entre-agrupamentos baseado no determinante de B

A maximização de $|B|$, ou $|B|/|T|$, ou ainda, a minimização de $|T|/|B|$, é conveniente apenas quando $K \geq d$ pois, em caso contrário, B é singular.

vi Relação de variabilidades entre/intra agrupamentos ($tr(W^{-1}B)$)

A solução ótima é a que maximiza o traço da matriz $W^{-1}B$ (ou BW^{-1}), ou seja, que maximiza a soma dos autovalores da matriz BW^{-1}. Uma variante é o índice tr(B)/tr(W).

vii Variabilidade intra-agrupamentos baseado em $\prod_{i=1}^{K} |W_i|^{n_i}$

A minimização de $\prod_{i=1}^{k} |W_i|^{n_i}$ tem mostrado melhores resultados do que a minimização de $|W|$, quando os agrupamentos têm formas diferentes entre si.

Teorema O índice quadrático, ou equivalentemente o índice tr(W), bem como os índices tr(B) e $|W|$, são invariantes a transformaçoes ortogonais (rotações) no espaço de atributos.

Demonstração : Tomando $\underline{y}_{ki} = A \cdot \underline{x}_{ki}$, temos que a matriz de covariância no novo espaço é $V = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (\underline{y}_{ki} - \overline{\underline{y}}_k) \cdot (\underline{y}_{ki} - \overline{\underline{y}}_k)^T = AWA^T$, onde A é, por hipótese, uma matriz ortogonal. Os autovalores de V são dados por $\det(\lambda I - AWA^T)=0$, e como $\det(\lambda I - AWA^T)=0$ é equivalente a $\det(\lambda A^T A - W)=0$ que é igual a $\det(\lambda I - W)=0$,

tem-se que os autovalores de W e V são os mesmos, e portanto o traço de ambas matrizes também é o mesmo. Segue que também o determinante de W é invariante a transformações ortogonais. Se a matriz A não for ortogonal, sendo apenas não-singular, então os autovalores de V serão diferentes dos de W e em geral a soma e o produto dos autovalores de V serão diferentes da soma e o produto dos autovalores de W. Como $V = AWA^T$ temos $|V| = |A|^2|W|$ e no caso geral $|A| \neq I$. Um caso importante é a matriz usada para padronização $A = \text{diag}\ (a_1, a_2 \ldots a_d)$, e, neste caso, $|A| = \prod_{i=1}^{d} a_i$.

É importante ressaltar que certos índices e critérios de agrupamentos são sensíveis à padronização, frequentemente necessária para compatibilizar os diferentes atributos empregados, que muitas vezes são medidos em escalas e unidades diferentes. Uma padronização pelo desvio padrão, equivalente ao emprego da distância ponderada, equivale a uma multiplicação dos vetores de atributos por uma matriz A não-singular (no caso seria diagonal). Diferentes matrizes A resultarão em diferentes partições ou agrupamentos para alguns dos critérios apresentados, o que é indesejável. Deve-se enfatizar que as matrizes normalmente usadas para padronização não são ortogonais.

Deve-se enfatizar que para um dado índice fornecer uma partição ótima, independentemente de transformações lineares não-singulares, não é necessário que o índice em si seja invariante a transformações lineares. Deve-se, isto sim, investigar se para duas partições diferentes e arbitrárias, com valores de índice maior numa que na outra, se mantém a desigualdade nos novos valores dos índices calculados para os vetores sujeitos a uma transformação linear não-singular arbitrária. Vamos examinar alguns casos:

a Caso dos índices quadrático e tr(W). Temos

$$J = \sum_{i=1}^{K} \left\{ \sum_{j=1}^{n_i} \left[\underline{x}_{ij}^T \ \underline{x}_{ij} \right] - n_i \, \overline{\underline{x}}_i^T \, \overline{\underline{x}}_i \right\}$$

e para os dados transformados segundo a matriz não-singular A :

$$J_{Tr} = \sum_{i=1}^{K} \left\{ \sum_{j=1}^{n_i} \left[\underline{x}_{ij}^T A^T A \, \underline{x}_{ij} \right] - n_i \, \underline{x}_i^T A^T A \, \underline{x}_i \right\}$$

A minimização de J é suposta ser obtida para uma partição G^*. Entretanto nada pode garantir que o mesmo G^* minimiza J_T , e portanto, o critério de minimizar o índice quadrático ou tr(W) é, em geral, sensível à particular escolha de padronização dos dados.

b Caso dos índices $|W|$ e $|B|$. Como para os vetores transformados a matriz de espalhamento intra-classes é $W_T = AWA^T$ temos $|W_T| = |A|^2|W|$ e portanto minimizar $|W|$ é o mesmo que minimizar $|W_T|$.

c Caso do índice $tr(BW^{-1})$. Para o caso de vetores transformados temos o índice $tr(ABA^T(A^T)^{-1}W^{-1}A^{-1} = tr(ABW^{-1} A^{-1})$, onde os autovalores de $ABW^{-1}A^{-1}$ são os mesmos de BW^{-1} e portanto o critério de maximizar $tr(BW^{-1})$ é invariante a transformações não singulares. O critério de maximizar $|BW^{-1}|$ também é invariante, mas não é muito útil, pois em muitos casos $|BW^{-1}|=0$ (quando d < c).

Além do problema da escolha do índice a ser otimizado, tem-se o problema de como efetuar a busca da partição ótima ou sub-ótima. Uma classe importante de algoritmos de busca é a dos algoritmos iterativos em que se parte de uma partição inicial e depois se tenta melhorar o desempenho (otimizando a função critério). Amostras são então transferidas de um agrupamento a outro de forma a melhorar o valor da função critério.

Transferência de um vetor entre dois agrupamentos

Vejamos qual o efeito causado sobre $J(G_{N,K})$ (índice quadrático) ao se transferir um certo vetor $\underline{x}_o$ do agrupamento G_k para o agrupamento G_j. Indicaremos por apóstrofo os valores calculados após transferir $\underline{x}_o$ de G_k para G_j . Os novos agrupamentos G'_k e G'_j terão vetores médios

$$\bar{\underline{x}}'_k = \bar{\underline{x}}_k - \frac{1}{n_k - 1}\left(\underline{x}_o - \bar{\underline{x}}_k\right) = \frac{n_k \bar{\underline{x}}_k - \underline{x}_o}{n_k - 1} \tag{14}$$

$$\bar{\underline{x}}'_j = \bar{\underline{x}}_j + \frac{1}{n_j + 1}\left(\underline{x}_o - \bar{\underline{x}}_j\right) = \frac{n_j \bar{\underline{x}}_j + \underline{x}_o}{n_j + 1} \tag{15}$$

A retirada de $\underline{x}_o$ de G_k faz o espalhamento no grupo k passar de $J_k(G_{N,K})$ para $J'_k(G'_{N,K})$, onde deve-se lembrar que para $J'_k(G'_{N,K})$ utiliza-se $\bar{\underline{x}}'_k$ e não mais $\bar{\underline{x}}_k$. Escreveremos daqui para diante J_k ao invés de $J_k(G_{N,K})$, por simplicidade.

$$J_k = \sum_{j=1}^{n_k} \underline{x}_{kj}^T \underline{x}_{kj} - n_k \bar{\underline{x}}_k^T \bar{\underline{x}}_k \tag{16}$$

$$J'_k = \sum_{j=1}^{n_k - 1} \underline{x}_{kj}^T \underline{x}_{kj} - (n_k - 1)\, \bar{\underline{x}}'^T_k \bar{\underline{x}}'_k \tag{17}$$

onde supusemos que $\underline{x}_o = \underline{x}_{k,n_k}$. Temos que

$$\sum_{j=1}^{n_k - 1} \underline{x}_{kj}^T \underline{x}_{kj} = \sum_{j=1}^{n_k} \underline{x}_{kj}^T \underline{x}_{kj} - \underline{x}_o^T \underline{x}_o \tag{18}$$

De (14) tem-se que :

$$(n_k - 1)^2\, \bar{\underline{x}}'^T_k \cdot \bar{\underline{x}}'_k = n_k^2\, \bar{\underline{x}}_k^T \bar{\underline{x}}_k - 2 n_k\, \underline{x}_o^T \bar{\underline{x}}_k + \underline{x}_o^T \underline{x}_o \tag{19}$$

Por definição

$$d_2^2(\underline{x}_o , \bar{\underline{x}}_k) = \underline{x}_o^T \underline{x}_o - 2\, \underline{x}_o^T \bar{\underline{x}}_k + \bar{\underline{x}}_k^T \bar{\underline{x}}_k \tag{20}$$

Multiplicando (20) por $n_k /(n_k - 1)$ tem-se :

$$\frac{n_k}{n_k - 1}\, d_2^2(\underline{x}_o , \bar{\underline{x}}_k) = \frac{n_k}{n_k - 1}\, \underline{x}_o^T \underline{x}_o - \frac{2n_k}{n_k - 1}\, \underline{x}_o^T \bar{\underline{x}}_k + \frac{n_k}{n_k - 1}\, \bar{\underline{x}}_k^T \bar{\underline{x}}_k \tag{21}$$

de (16), (17), (18) e (19).

$$J_k - J'_k = \underline{x}_o^T \underline{x}_o - n_k \bar{\underline{x}}_k^T \bar{\underline{x}}_k + \frac{n_k^2\, \bar{\underline{x}}_k^T \bar{\underline{x}}_k}{n_k - 1} - \frac{2\, n_k}{n_k - 1}\, \underline{x}_o^T \bar{\underline{x}}_k + \frac{\underline{x}_o^T \underline{x}_o}{n_k - 1} \tag{22}$$

de (22) :

$$J_k - J'_k = \frac{n_k}{n_k - 1} \underline{x}_o^T \underline{x}_o - \frac{2\, n_k}{n_k - 1} \underline{x}_o^T \overline{\underline{x}}_k + \frac{n_k \overline{\underline{x}}_k^T \overline{\underline{x}}_k}{n_k - 1} \qquad (23)$$

de (21) e (23) temos :

$$J'_k = J_k - \frac{n_k}{n_k - 1} d_2^2(\underline{x}_o , \overline{\underline{x}}_k) \qquad (24)$$

O agrupamento G_j ao receber o novo vetor $\underline{x}_o$ tem seu centróide deslocado para $\overline{\underline{x}}'_j$ resultando em novo valor J'_j para a função de espalhamento neste grupo. De forma análoga à derivação anterior obtemos:

$$J'_j = J_j + \frac{n_j}{n_j + 1} d_2^2(\underline{x}_o , \overline{\underline{x}}_j) \qquad (25)$$

Note que em (24) e (25) as expressões utilizam os centróides originais dos dois agrupamentos devido à sua maior utilidade prática (só recalculam-se os centróides quando já for selecionada a nova partição, ou seja o destino final de $\underline{x}_o$).

A função critério global muda de J para J' com

$$J' - J = \frac{n_j}{n_j + 1} d_2^2(\underline{x}_o , \overline{\underline{x}}_j) - \frac{n_k}{n_k - 1} d_2^2(\underline{x}_o , \overline{\underline{x}}_k) \qquad (26)$$

e para que a transferência de $\underline{x}_o$ de G_k para G_j diminua o espalhamento global devemos ter

$$\frac{n_j}{n_j + 1} d_2^2(\underline{x}_o , \overline{\underline{x}}_j) < \frac{n_k}{n_k - 1} d_2^2(\underline{x}_o , \overline{\underline{x}}_k) \qquad (27)$$

ou seja

$$d_2^2(\underline{x}_o , \overline{\underline{x}}_j) < \left[\frac{n_k (n_j + 1)}{(n_k - 1) n_j} \right] d_2^2(\underline{x}_o , \overline{\underline{x}}_k) \qquad (28)$$

Como a expressão entre colchetes em (28) é maior que 1, conclui-se que,

sempre que algum algoritmo de agrupamento optar por transferir um elemento $\underline{x}_o$ de G_k para G_j com

$$d_2^2(\underline{x}_o , \overline{\underline{x}}_j) < d_2^2(\underline{x}_o , \overline{\underline{x}}_k) , \qquad (29)$$

a função critério espalhamento quadrático global diminuirá. A desigualdade (29) significa que a distância de $\underline{x}_o$ ao centróide original de G_j é menor que a distância de $\underline{x}_o$ ao centróide original de G_k (original significando antes de se efetuar a passagem de $\underline{x}_o$ de G_k para G_j). A desigualdade (28) é menos restritiva que a (29), ou seja, pode-se ainda conseguir diminuição em J na transferência de G_k para G_j quando (29) não é satisfeita mas (28) é. Por exemplo, se $d_2^2(\underline{x}_o , \underline{x}_k)= 2000$, $d_2^2(\underline{x}_o , \underline{x}_j) = 2020$, $n_k = 3$, $n_j = 4$, conclui-se que (29) não é satisfeita mas (28) é. De (26) tem-se que J'-J = - 1384, o que indica que vale a pena efetuar a realocação da amostra $\underline{x}_o$. Pode-se concluir, que quando houver pelo menos 1 agrupamento com poucas amostras, é preferível utilizar (28) ao invés de (29).

Supondo que se está analisando a necessidade da transferência de um particular elemento $\underline{x}_o$ pertencente a G_k , deve-se calcular todas as distâncias ao quadrado $d_2^2(\underline{x}_o , \overline{\underline{x}}_j)$, $j \neq k$, selecionar o mínimo, e, se com este valor mínimo a desigualdade (28) (ou a (29) caso se deseje) for satisfeita então efetuar a transferência.

Há um grande número de algoritmos de agrupamento utilizando a filosofia de partição, a maioria requerendo que se especifique o número de agrupamentos desejado. Serão apresentados a seguir alguns algoritmos importantes dentro desta categoria, em que a proximidade é definida em relação a centróides.

i Algoritmo de K-médias de MacQueen ("K-means")

Supomos dados N vetores ou amostras compondo o conjunto S_N e que desejamos K agrupamentos.

1 - Tomar os primeiros K vetores de S_N como vetores de partida. Cada um passa a definir um agrupamento (de 1 único elemento). Caso haja correlação (serial) entre vetores subsequentes de S_N pode-se escolher os K vetores de partida aleatoriamente dentre os N vetores de S_N .

2 - Atribuir sequencialmente cada um dos N-K vetores ao agrupamento com o centróide (vetor médio do agrupamento) mais próximo. Após cada atribuição, recalcular o centróide do agrupamento que foi acrescido.

3 - Terminada a etapa 2, tomar os K centróides resultantes como novos pontos de partida para K agrupamentos. Haverá agora uma segunda passada pelos dados só que mantendo os centróides fixos nos valores fornecidos pela etapa 2. Cada um dos N vetores é reclassificado atribuindo-o ao agrupamento correspondente ao centróide mais próximo. O diagrama de fluxo da Fig. 1 sintetiza o algoritmo.

Com este algoritmo tem-se a seguinte carga computacional:

(*) N-K cálculos de centróides

(*) $K(N-K)+NK = 2NK-K^2$ cálculos de distâncias (euclideanas, se já supomos dados padronizados)

(*) 2N-K buscas de mínimo entre K números

Este é um algoritmo básico, sendo que hoje em dia o nome K-means significa uma família de técnicas.

O algoritmo apresentado tem a característica de ser rápido quando comparado com outros. Os resultados fornecidos por este algoritmo dependem de :

(*) número de agrupamentos escolhido a priori

(*) ordem de apresentação dos vetores, tanto para a inicialização quanto pelo restante do procedimento.

(*) geometria dos agrupamentos .

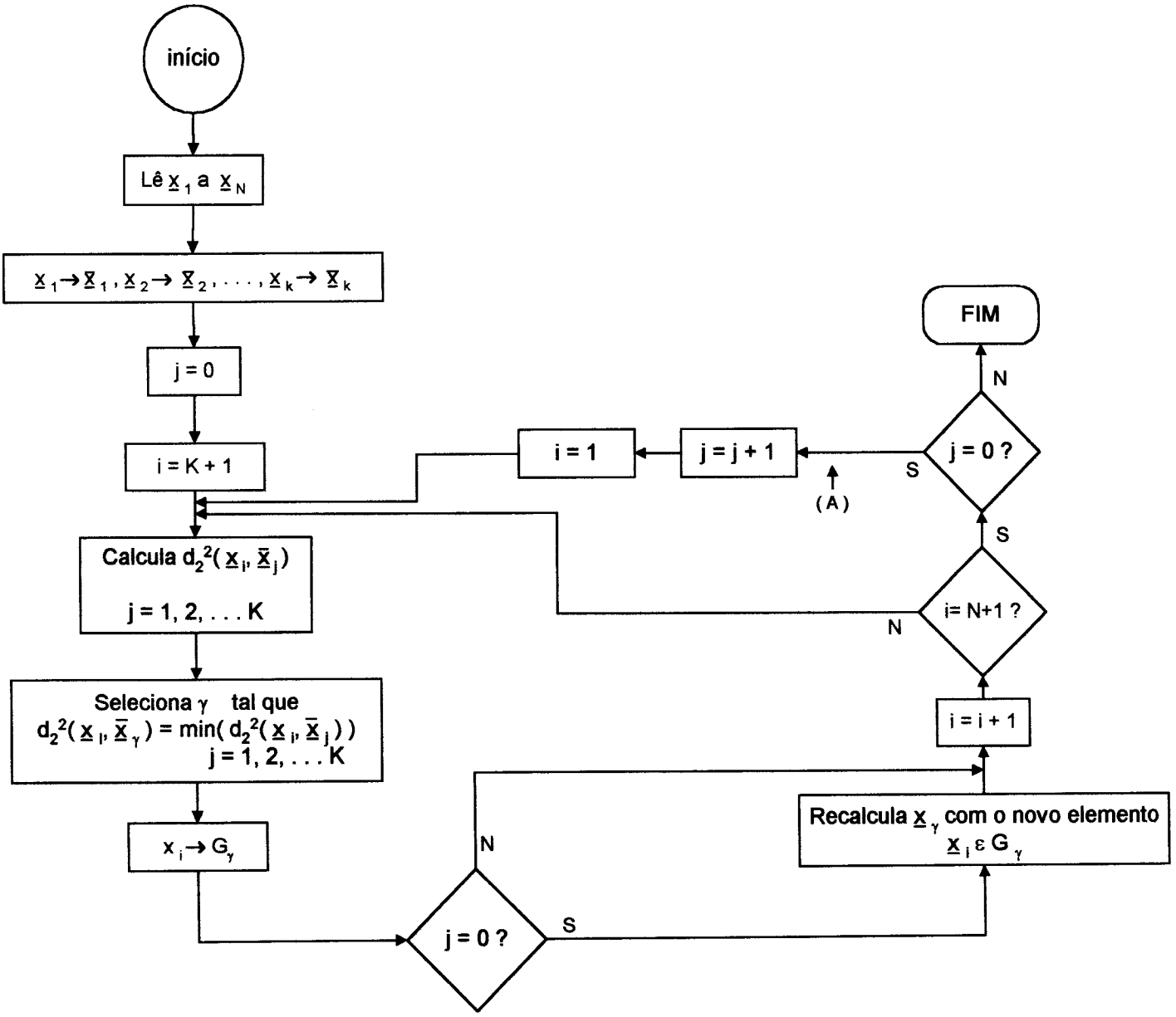


Fig. 1 - Diagrama de fluxo do algoritmo de K-médias

Deve-se notar que o critério de otimização é o quadrático mas, fica claro que devido ao algoritmo de busca muito rudimentar, em geral, a partição obtida não será a ótima podendo inclusive estar bem longe dela. Recomenda-se repetir o algoritmo pelo menos uma vez utilizando-se vetores de partida diferentes e comparar os agrupamentos obtidos nas diferentes vezes em que o algoritmo foi utilizado.

ii Algoritmo de Forgy

É um algoritmo mais simples que o de K-médias de MacQueen sendo na realidade o seu percursor.

Forgy propôs dois modos de inicialização, um em que se escolhe 1 vetor de partida para cada agrupamento (por exemplo os K primeiros vetores do conjunto S_N), e outro, em que já se parte de uma partição $G_{N,K}(0)$ dos dados. O diagrama de fluxo da Fig. 2 explica o algoritmo supondo que se parta de uma inicialização com 1 vetor em cada agrupamento.

O bloco que pergunta se $G_{N,K}(k) = G_{N,K}(k-1)$ significa perguntar se na última passagem (k-ésima) pelo algoritmo houve alteração na atribuição de pelo menos 1 vetor de S_N em relação à partição (k-1) ésima. Uma critério de parada alternativo é verificar se não houve alteração nos valores dos centróides da iteração (k-1) ésima para a k-ésima.

Um aspecto de simplicidade do algoritmo é que ele mantém fixa a escolha inicial dos vetores representativos dos agrupamentos o que é bom computacionalmente falando mas é ruim em termos da qualidade de agrupamento dos dados. Para compensar a grande probabilidade de formar uma primeira partição muito ruim o algoritmo permite um número arbitrário de iterações até que não haja mais alteração na atribuição de cada vetor de S_N . Na prática, em geral há convergência para menos que 10 iterações.

Para mostrar que o algoritmo de Forgy é convergente vamos antes for-

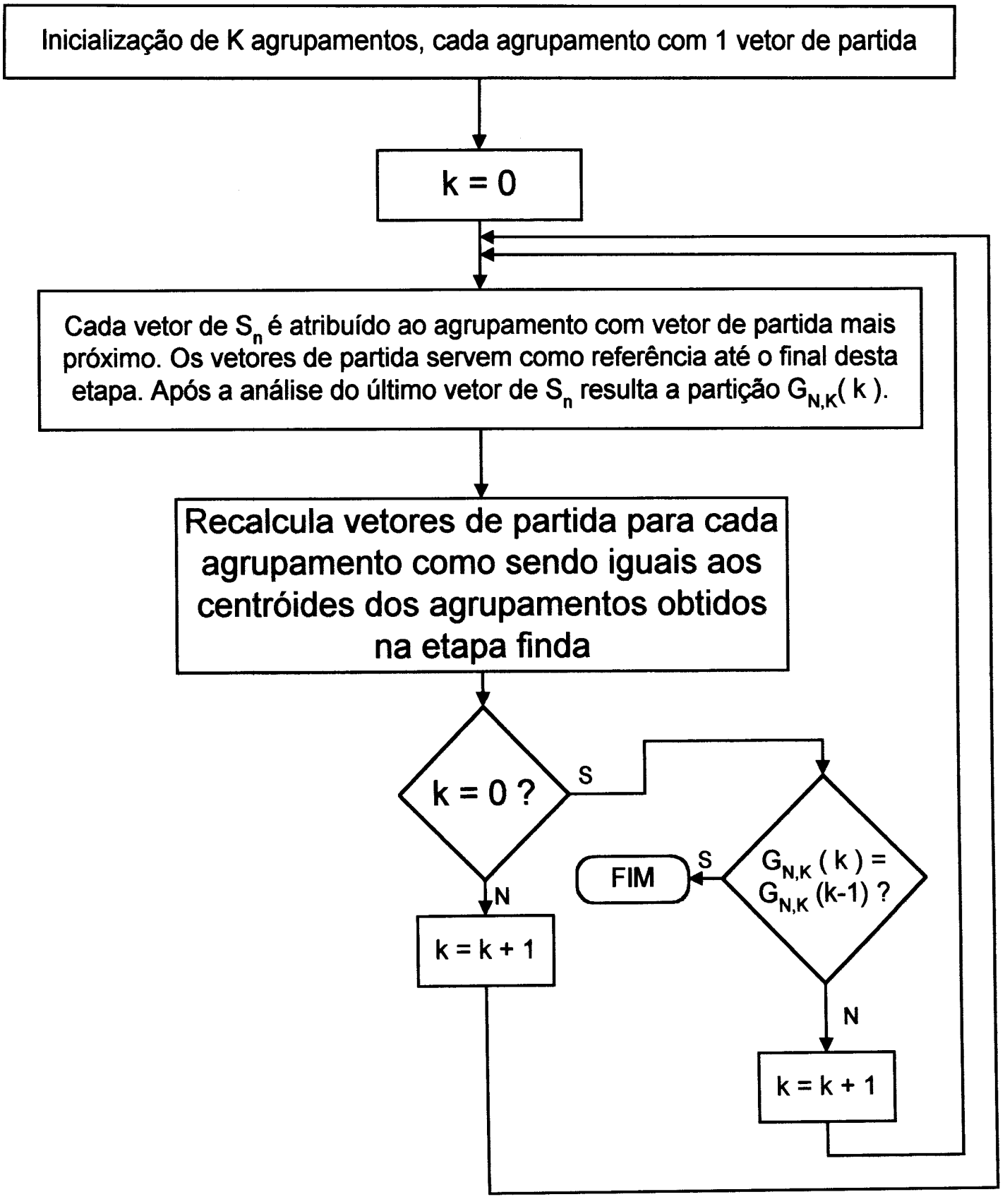


Fig. 2 - Diagrama de fluxo do algoritmo de Forgy

necer um Lema :

Lema: Para um agrupamento de dados G_K , a soma das distâncias Euclideanas ao quadrado, em torno de um ponto de referência arbitrário, tem o seu valor mínimo único quando o ponto de referência coincide com o centróide.

Demonstração:

$$\frac{\partial}{\partial \underline{x}_r} \left[\sum_{j=1}^{n_k} (\underline{x}_{kj} - \underline{x}_r)^T (\underline{x}_{kj} - \underline{x}_r) \right] = \underline{0}$$

$$- \sum_{j=1}^{n_k} \underline{x}_{kj} + 2n_k \underline{x}_r = 0$$

$$\therefore \quad \underline{x}_r = \frac{1}{n_k} \sum_{j=1}^{n_k} \underline{x}_{kj}$$

No algoritmo de Forgy, cada vetor é realocado quando sua distância ao ponto de referência de um dado agrupamento é menor que ao ponto de referência fixo do agrupamento em que estava. Portanto, a soma dos quadrados das distâncias em torno do respectivo ponto de referência diminui em maior grau para o agrupamento em que estava o vetor, do que aumenta para o agrupamento para onde foi, no total resultando um decréscimo. Por ocasião da adoção dos centróides como novos pontos de referência diminui-se ainda mais a soma total dos quadrados das distâncias intra-agrupamentos, concluindo-se então que o algoritmo é convergente.

O algoritmo só é sensível à ordem dos vetores em S_N no que tange uma inicialização em que se tomam, por exemplo, os K primeiros vetores de S_N para vetores de partida. O restante do procedimento é insensível à ordem de utilização dos vetores pois os centróides só são recalculados ao final da atribuição de todos os vetores.

Como a atribuição é baseada na maior proximidade a K centróides fixos durante a atribuição dos N vetores, segue que as fronteiras entre pares de

agrupamentos são hiperplanos (vide o capítulo Classificação de Padrões por Mínima Distância, 1 protótipo por classe).

A utilização da distância Euclideana neste algoritmo significa que o índice está sendo o quadrático, embora, como a busca não é exaustiva, não é possível assegurar ótimo global, mas sómente um ótimo local.

iii Um algoritmo de K-médias iterativo (Wishart, McRae)

É uma mescla das abordagens de Forgy e MacQueen. O diagrama de fluxo da Fig. 3 explica o algoritmo, sendo que para a partição inicial pode-se, por exemplo, empregar uma iteração do algoritmo de MacQueen, ou seja até o ponto (A) no diagrama de fluxo correspondente (Fig. 1).

O desempenho deste algoritmo, a princípio, é melhor que o de Forgy e MacQueen, mas exige um tempo de computação geralmente maior (pelo menos em relação ao de MacQueen). Como o critério empregado é o (29) que é mais restritivo que o (27), conclui-se que o algoritmo é convergente.

Da mesma forma que nos outros dois algoritmos já abordados, caso no presente algoritmo se empregue a distância Euclideana tem-se o critério quadrático. Vale lembrar que este critério tende a fornecer agrupamentos compactos esféricos ou eventualmente hiperelipsoidais.

Outro aspecto a ressaltar é que nenhum dos 3 algoritmos apresentados faz uma re-atribuição ou realocação de um dado vetor $\underline{x}_o$ baseado no critério expresso por (27) ou (28). De fato, no de Forgy os centróides são recalculados somente após terminada a atribuição de todos os vetores. Na segunda etapa do algoritmo de K-médias de MacQueen, em que há realocação, os centróides previamente determinados passam a ser referências fixas. No algoritmo de K-médias iterativo a decisão de realocação é baseada em (29), ou seja, na distância aos centróides. Portanto, pode deixar de haver realocação de uma amostra mesmo que isto redunde em um decréscimo em J. Pode-se então utili-

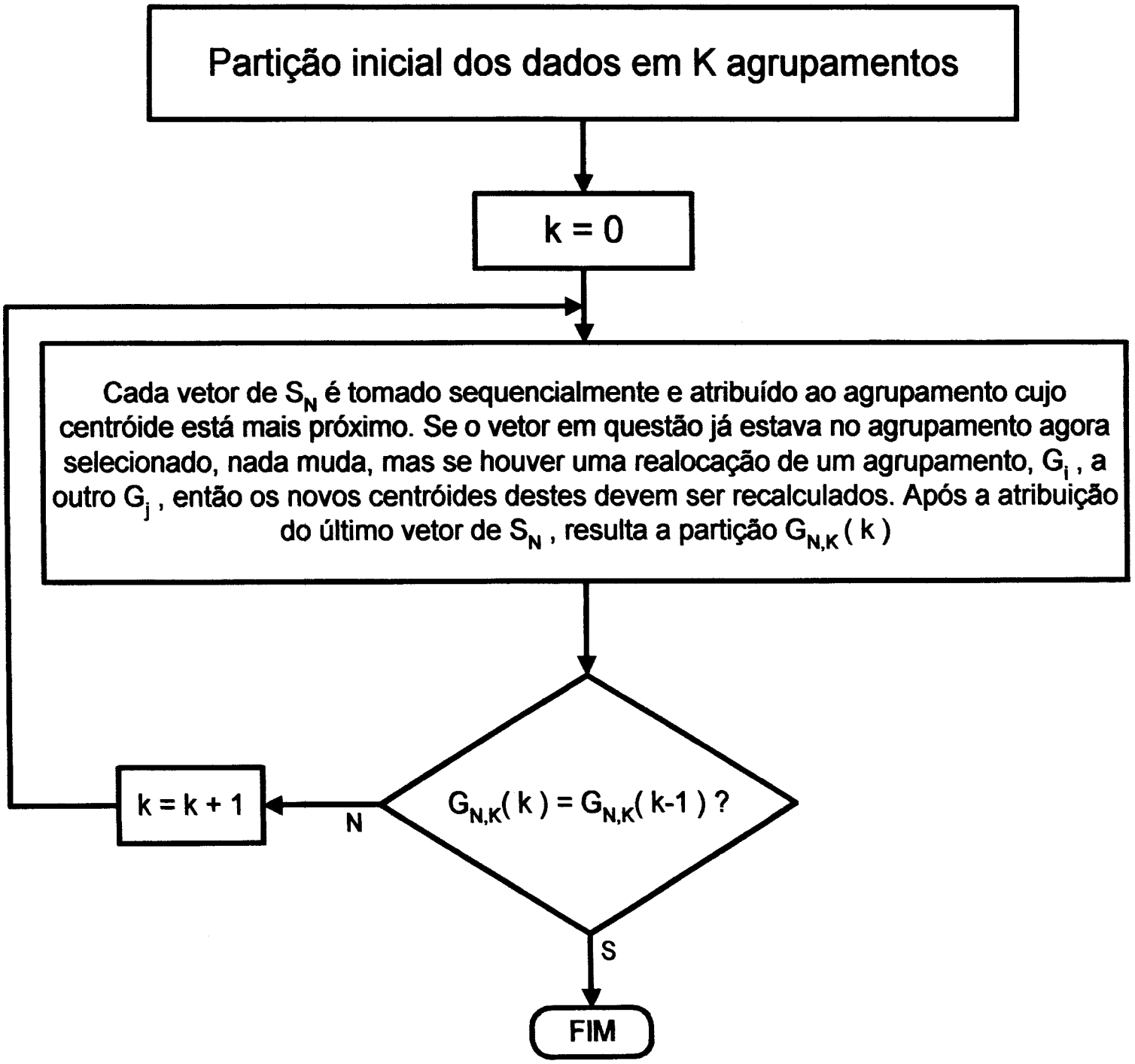


Fig. 3 - Diagrama de fluxo de um algoritmo de K-médias iterativo.

zar, com vantagens, o critério de realocação dado por (27) ou (28) no algoritmo de K-médias iterativo fornecido, conforme enfatizado no item seguinte.

iv Um algoritmo de K-médias de mínimo espalhamento quadrático intra-agrupamento

Neste se utiliza a desigualdade em (27) como critério para realocação de um vetor de um agrupamento a outro. Desta forma converge-se para um mínimo local da função critério quadrático, que mede o espalhamento intra-agrupamentos. A partição inicial indicada no diagrama de fluxo da Fig. 4 pode, por exemplo, ser gerada por uma etapa do algoritmo de K-médias de Mac Queen (ponto (A) no diagrama de fluxo apresentado na Fig. 1).

Em muitas aplicações não há conhecimento sobre o número de agrupamentos existentes no conjunto de dados fornecidos. Nestes casos deve-se utilizar algoritmos em que não é pré-fixado o número K de agrupamentos uma vez que uma escolha errada de K pode levar à criação de agrupamentos totalmente inadequados e inúteis. Em continuidade aos algoritmos já descritos, apresentam- se outros, em que o número de agrupamentos é estabelecido pelo algoritmo a partir dos próprios dados e de certos critérios.

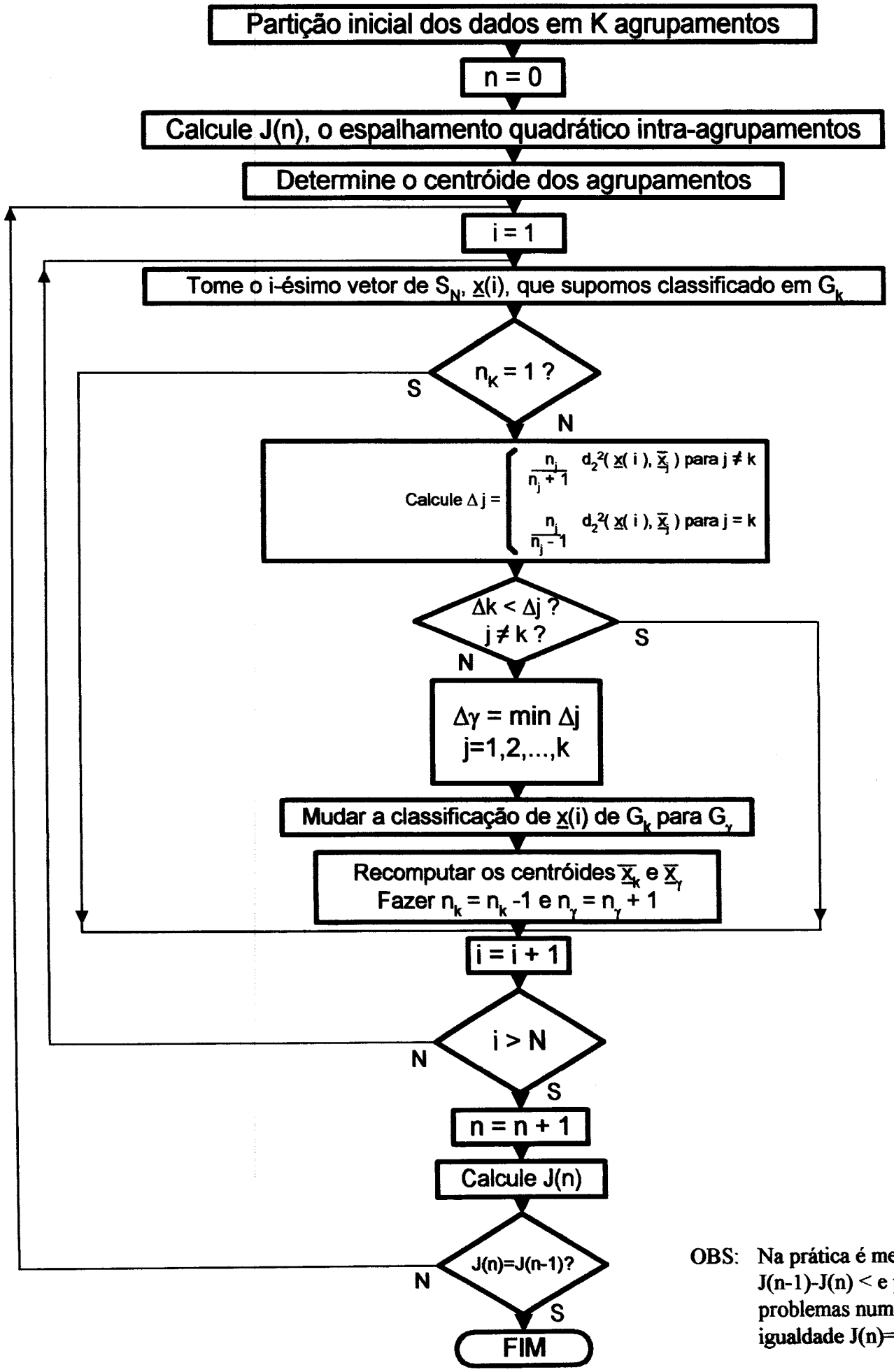
v Algoritmo de K-médias de MacQueen com número variável de agrupamentos

É baseado no algoritmo básico de K-médias já discutido, contando com a adição de dois parâmetros α e θ . As etapas são :

(1) Selecione K , α e θ

(2) Tome os primeiros K vetores de S_N como os agrupamentos de partida, cada agrupamento tendo 1 único vetor que é portanto o próprio centróide.

(3) Calcule as distâncias entre todos os pares de centróides e caso a mínima distância seja menor que α então os dois agrupamentos correspondentes devem ser aglutinados, passando-se a ter K-1 agrupamentos. O novo agru-



OBS: Na prática é melhor utilizar J(n-1)-J(n) < e para se evitar problemas numéricos com a igualdade J(n)=J(n-1).

Fig. 4 - Diagrama de fluxo de um algoritmo de K-médias de mínimo espalhamento quadrático intra-agrupamento

pamento deve ter seu centróide recalculado. As distâncias deste centróide aos outros devem ser calculadas. Repetir a operação de aglutinação caso a mínima distância seja menor que α . Continuar até que todas os centróides dos agrupamentos remanentes estejam a uma distância de pelo menos α . Caso se escolha K muito pequeno e α muito grande poder-se-á finalizar esta etapa com apenas 1 agrupamento (o que raramente é o caso). O parâmetro α controla a aglutinação, podendo grosseiramente ser pensado como uma aproximação a um "diâmetro" do agrupamento de menor volume.

(4) Tome os N-K vetores restantes de S_N sequencialmente, e, para simplicidade, denotemos por $\underline{x}_o$ o vetor em análise no momento. Se a distância de $\underline{x}_o$ ao centróide mais próximo é maior que θ ($\theta>\alpha$) então este $\underline{x}_o$ define um novo agrupamento com o próprio $\underline{x}_o$ como centróide. Em caso contrário, atribua $\underline{x}_o$ ao agrupamento com o centróide mais próximo. A cada atribuição recalcule o centróide do agrupamento escolhido. Calcule a distância deste novo centróide aos outros centróides. Se a menor distância for menor que α então aglutinar os dois agrupamentos correspondentes e calcular o centróide do novo agrupamento. Calcular as distâncias do novo centróide a todos os outros e repetir o procedimento de aglutinação até que todas os centróides estejam a uma distância maior que α entre si.

(5) Finalizada a etapa (4) adotar os centróides resultantes como novos vetores de partida e reclassificar os N vetores utilizando como critério a mínima distância aos vetores de partida.

Da mesma forma que o algoritmo de K-médias com número fixo de agrupamentos, o presente algoritmo tem o mérito de exigir menor tempo de computação que outros mais complexos.

Como na segunda passada pelos dados mantém-se fixos o número de agru-

pamentos e os vetores de referência, não se pode garantir que os centróides dos agrupamentos resultantes ao final da etapa 5 estejam todos afastados de pelo menos α .

A escolha de α e θ não é fácil, podendo-se utilizar alguma consideração baseada nos desvios padrões globais dos atributos estimados a partir dos N dados disponíveis.

vi Algoritmo de agrupamento por vizinho mais próximo (Lu e Fu)

Neste algoritmo, o usuário tem que fornecer um limiar L. O diagrama de fluxo da Fig. 5 sintetiza o algoritmo, que é totalmente heurístico, muito simples e que trata os dados sequencialmente. Por isto a partição obtida depende da ordem em que os vetores se encontram em S_N , podendo-se chegar a resultados bastante ruins. A escolha de L também é um aspecto que irá determinar o sucesso ou o fracasso do algoritmo não havendo critérios quantitativos para esta escolha. Em síntese, o número de agrupamentos e os próprios agrupamentos resultantes dependem de L, da geometria das classes naturalmente existentes e da sequência em que são tomados os vetores de S_N . Há a possibilidade de haver encadeamento de agrupamentos naturais formando um único agrupamento.

Pode-se utilizar outra distância entre vetor e agrupamento que não a distância do vizinho mais próximo como, por exemplo, a distância ao centróide do agrupamento.

Havendo uma escolha "razoável" para L , o algoritmo dará bons resultados no caso de agrupamentos conexos e convexos bem separados entre si. A vantagem do algoritmo é a simplicidade, tratando os dados sequencionalmente, com uma única passada pelos dados.

O número existente de algoritmos de agrupamento baseados em partição é muito grande, devendo o leitor consultar a bibliografia fornecida no final deste texto.

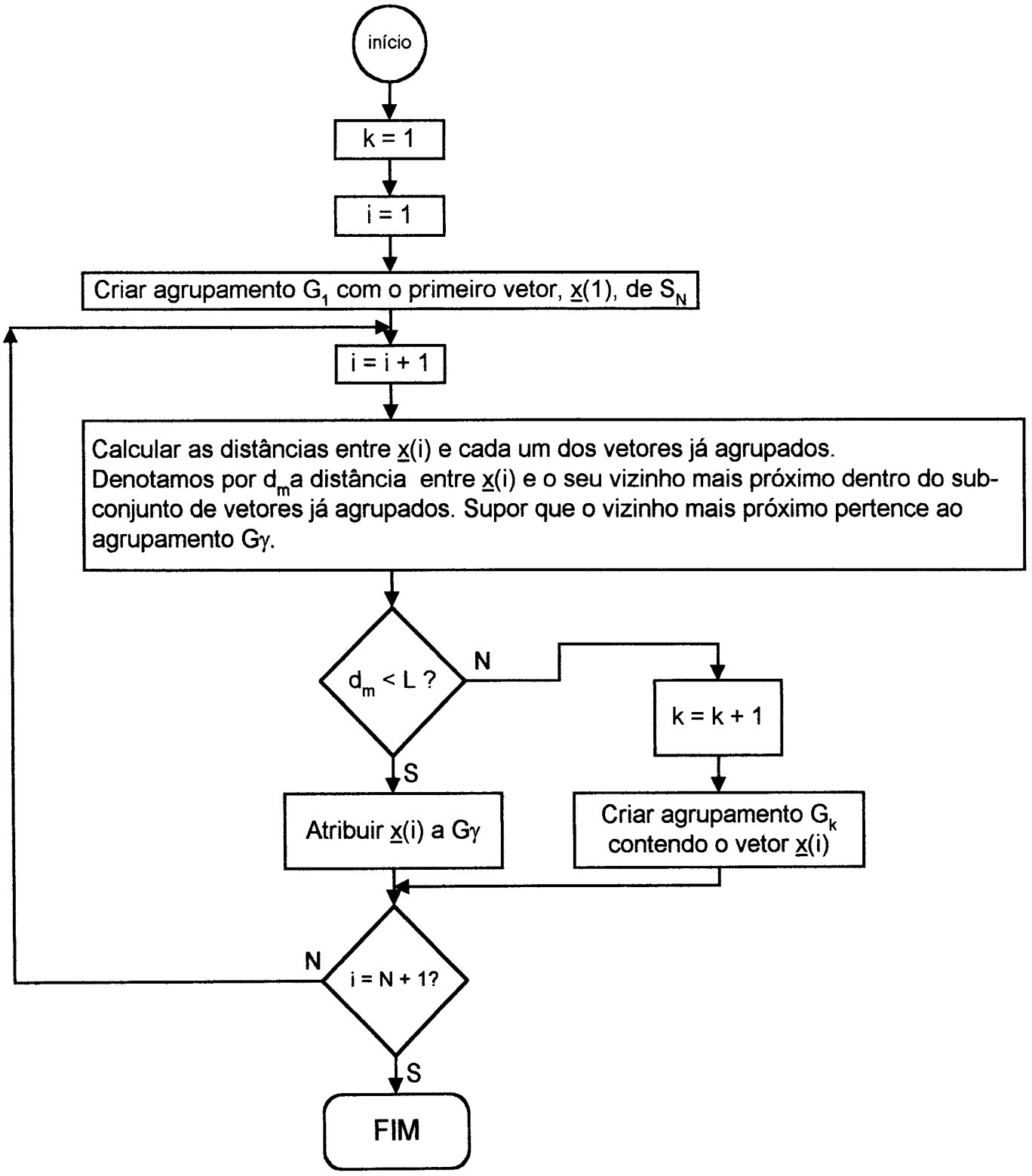


Fig. 5 - Diagrama de fluxo de um algoritmo de partição por vizinho mais próximo.

MÉTODOS DE AGRUPAMENTO HIERÁRQUICOS

São métodos mais úteis para áreas como biologia, psicologia, ciências humanas, etc. Para ciências exatas os métodos particionais são mais relevantes, pois nestas, geralmente, se tem um número grande de elementos para agrupar (centenas, milhares) e se deseja obter uma única partição que de preferência otimize (pelo menos localmente) alguma função critério. Por outro lado, em biologia e outras áreas, o número de amostras, em geral, não é grande, e o pesquisador gosta de visualizar uma árvore que representa as relações de similaridade ou dissimilaridade entre os elementos e entre os agrupamentos, permitindo assim obter intuição quanto a semelhanças, graus de semelhanças e números possíveis de agrupamento.

O nome hierárquico vem do fato do método criar uma sequência de partições $G_N(k)$ onde cada partição tem agrupamentos que são subconjuntos próprios dos agrupamentos da partição seguinte (no caso de algoritmos aglomerativos), ou da partição anterior (no caso de algoritmos divisivos).

Um algoritmo <u>aglomerativo</u> de agrupamento hierárquico tem a seguinte estrutura, explicada construtivamente:

<u>i</u> Partir de N agrupamentos, cada um representado por um dos vetores de S_N. Calcular uma matriz de distâncias D entre todos os pares de agrupamentos, ou seja, entre todos os pares de vetores $\underline{x}(i)$ e $\underline{x}(j)$ de S_N .

<u>ii</u> Procurar em D o par de agrupamentos mais próximo ou similar. Seja este par o agrupamento G_i e G_j com distância d_{ij} .

<u>iii</u> Aglutinar G_i com G_j formando o agrupamento G_{ij} . Obter a nova matriz D eliminando as linhas e colunas i e j e acrescentando uma linha e uma coluna contendo as distâncias do novo agrupamento G_{ij} aos demais N-2

agrupamentos.

iv Repetir ii e iii N-1 vezes até obter 1 único agrupamento. Armazenar as fusões todas que foram ocorrendo juntamente com as distâncias em que ocorreram.

Trataremos aqui somente de algoritmos aglomerativos pela sua maior importância prática. Para uma primeira leitura sobre algoritmos divisivos, que tem uma filosofia inversa à dos aglomerativos, vide Everitt (1981). Portanto, para inicialização, parte-se do conjunto S_N de vetores (dx1) e escolhe-se um índice de similaridade ou dissimilaridade entre os vetores. Geralmente utiliza-se alguma medida de distância e então se calcula uma matriz simétrica D(NxN) de distâncias $d[\underline{x}(i),\underline{x}(j)]$. Deve-se notar que para a etapa iii deve-se também definir o que se entende por distância entre agrupamentos. Três definições serão vistas nos algoritmos que se seguem.

i Algoritmo hierárquico de vizinho mais próximo ("single linkage")

Neste, a distância d_{ij} entre agrupamentos G_i e G_j é definida como a menor distância entre um elemento de G_i (dentre todos) e um elemento de G_j (dentro todos). Em termos de um algoritmo aglomerativo a matriz D da etapa i contém distâncias entre vetores. Na etapa iii é formado o agrupamento G_{ij} e o cálculo de sua distância ao agrupamento arbitrário G_γ é muito fácil:

$$d_{ij,\gamma} = d\left[G_{ij}\ ,\ G_\gamma\right] = \min\left\{\ d\left[G_i\ ,\ G_\gamma\ \right],\ d\left[G_j\ ,\ G_\gamma\ \right]\right\}$$

Muito embora não seja necessário do ponto de vista do algoritmo computacional, (que faz os cálculos de distância sequencialmente), convém repetir que a distância entre 2 agrupamentos G_i e G_j é a distância entre os 2 elementos (1 de cada agrupamento) mais próximos de G_i e G_j .

Normalmente se faz um desenho em forma de árvore, o "dendograma", mostrando as aglutinações nas junções dos ramos, com pares de ramos tendo

tamanho indicativo da distância entre os 2 agrupamentos aglutinados. A partir deste dendograma pode-se decidir qual o número mais "razoável" de agrupamentos, baseando esta decisão tanto nas distâncias entre os agrupamentos quanto em conhecimentos do problema específico. Caso o número de agrupamentos tenha sido pré-fixado, pode-se facilmente fazer o algoritmo parar quando este for atingido.

⊙⊙⊙⊙⊙ Exemplo: N = 6 , matriz de distâncias entre vetores de S_6 :

$$D = \begin{array}{c|cccccc} 1 & 0 & & & & & \\ 2 & 13 & 0 & & & & \\ 3 & 15 & (1) & 0 & & & \\ 4 & 16 & 2 & 9 & 0 & & \\ 5 & 10 & 8 & 8 & 11 & 0 & \\ 6 & 14 & 15 & 19 & 12 & 3 & 0 \\ \hline & 1 & 2 & 3 & 4 & 5 & 6 \end{array}$$

Cumpre ressaltar que os elementos da matriz são números inteiros apenas por razões didáticas. Normalmente as distâncias serão número reais (não negativos).

Aglutina-se G_2 com G_3 , obtendo-se G_{23} . Calculam-se as distâncias:

$d_{23,1} = 13$, $d_{23,4} = 2$, $d_{23,5} = 8$, $d_{23,6} = 15$.

A nova matriz D é :

$$D = \begin{array}{c|ccccc} 1 & 0 & & & & \\ (23) & 13 & 0 & & & \\ 4 & 16 & (2) & 0 & & \\ 5 & 10 & 8 & 11 & 0 & \\ 6 & 14 & 15 & 12 & 3 & 0 \\ \hline & 1 & (23) & 4 & 5 & 6 \end{array}$$

Aglutina-se G_{23} com G_4 , obtendo-se G_{234} . Calculam-se as distâncias:

$d_{234,1} = 13$, $d_{234,5} = 8$, $d_{234,6} = 12$. A nova matriz D é:

$$D = \begin{array}{c|cccc} 1 & 0 & & & \\ 234 & 13 & 0 & & \\ 5 & 10 & 8 & 0 & \\ 6 & 14 & 12 & (3) & 0 \\ \hline & 1 & 234 & 5 & 6 \end{array}$$

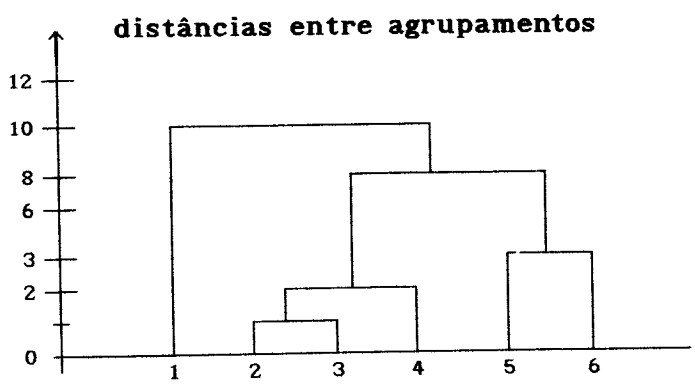
Aglutina-se G_5 com G_6 , obtendo-se G_{56} . Calculam-se as distâncias: $d_{56,1} = 10$ e $d_{56,234} = 8$. A nova matriz D é :

$$D = \begin{array}{c|ccc} 1 & 0 & & \\ 234 & 13 & 0 & \\ 56 & 10 & (8) & 0 \\ \hline & 1 & 234 & 56 \end{array}$$

Aglutina-se G_{56} com G_{234} , obtendo-se G_{23456} . Tem-se: $d_{23456,1} = 10$. A nova matriz D é :

$$D = \begin{array}{c|cc} 1 & 0 & \\ 23456 & (10) & 0 \end{array}$$

Contruímos abaixo o dendograma a partir dos valores obtidos nas operações que acabamos de realizar.



◎◎◎◎◎

A visualização do dendograma é bem mais útil para a obtenção de intuição do que a mera análise da matriz de distâncias inicial. Em muitos casos, fica aparente em que nível da árvore se encontram agrupamentos "naturais".

No exemplo acima, a aglutinação dos agrupamentos G_{234} com G_{56} se dá a um nível (distância entre agrupamento) muito grande (8) quando comparado com os níveis em que houverem aglutinações anteriormente (p.ex. G_{23} com G_4 , ou, G_5 com G_6). Portanto, o dendograma sugere que há 3 agrupamentos : G_1 , G_{56} e G_{234} . Quando restarem agrupamentos com poucos elementos (1 ou 2 elementos, por exemplo) deve-se investigar se são devidos a amostras anômalas ("outliers") causadas por erros de medição, ruídos, etc. O pesquisador sempre deve utilizar todo seu conhecimento sobre o problema específico, bem como seu bom senso, para julgar a adequação de uma ou outra configuração de agrupamentos indicada pelo dendograma.

O algoritmo hierárquico de vizinho mais próximo tem como trunfo poder criar agrupamentos não-elipsoidais e tem como desvantagem a possibilidade de encadear agrupamentos mal separados. A Fig. 6 mostra 2 nuvens de pontos que deveriam definir dois agrupamentos, mas devido ao efeito de encadeamento ("chaining") intrínseco aos métodos hierárquicos de vizinho mais próximo haverá tendência em fundí-los em um único agrupamento relativamente no início da construção da árvore.

Caso a matriz D inicial seja armazenada na memória principal do computador, fica claro que não se pode abordar problemas de porte. Para N vetores a serem agrupados, deve-se armazenar $(N^2-N)/2$ distâncias que definem a matriz D. Por exemplo, para N=300 teremos que armazenar 44.850 distâncias, geralmente em ponto flutuante, o que para certos computadores pode ser proibitivo.

ii Algoritmo hierárquico de vizinho mais distante ("complete linkage")

A distância entre dois agrupamentos G_i e G_j é definida de forma oposta ao do algoritmo anterior: ela é a distância entre os 2 elementos mais distantes de G_i e G_j , tomando um em cada agrupamento. Em uma implementação

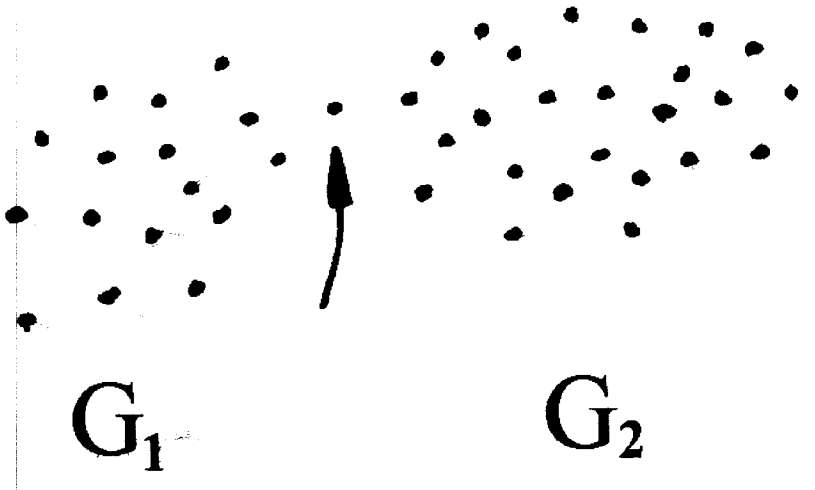


Fig. 6 – Dois agrupamentos que poderiam vir a ser indevidamente unidos muito de imediato devido à amostra marcada com uma seta.

por algoritmo aglomerativo deve-se atentar apenas para a etapa iii em que se torna necessário o cálculo das distâncias do agrupamento aglutinado G_{ij} aos demais agrupamentos. Tomamos como exemplo a distância entre G_{ij} e um agrupamento arbitrário G_γ :

$$d_{ij,\gamma} = d\left[G_{ij} , G_\gamma\right] = \max \left\{ d\left[G_i , G_\gamma \right] , d\left[G_j , G_\gamma \right] \right\}$$

De resto, o algoritmo prossegue como no caso anterior, conforme exemplificado a seguir, onde se utiliza a mesma matriz de distâncias inicial D empregada no exemplo da seção anterior.

⊙⊙⊙⊙⊙ Exemplo: N = 6 e matriz de distâncias entre vetores de S_6 :

$$D = \begin{array}{c|cccccc} 1 & 0 & & & & & \\ 2 & 13 & 0 & & & & \\ 3 & 15 & (1) & 0 & & & \\ 4 & 16 & 2 & 9 & 0 & & \\ 5 & 10 & 8 & 8 & 11 & 0 & \\ 6 & 14 & 15 & 19 & 12 & 3 & 0 \\ \hline & 1 & 2 & 3 & 4 & 5 & 6 \end{array}$$

Aglutina-se G_2 com G_3 pois são os agrupamentos mais próximos ou similares. Calculam-se as distâncias :

$$d_{23,1} = \max \left[d_{2,1} , d_{3,1} \right] = \max \left[13, 15 \right] = 15$$

$d_{23,4} = 9$, $d_{23,5} = 8$, $d_{23,6} = 19$. A nova matriz D é :

$$D = \begin{array}{c|ccccc} 1 & 0 & & & & \\ 23 & 15 & 0 & & & \\ 4 & 16 & 9 & 0 & & \\ 5 & 10 & 8 & 11 & 0 & \\ 6 & 14 & 19 & 12 & (3) & 0 \\ \hline & 1 & 23 & 4 & 5 & 6 \end{array}$$

Os agrupamentos mais similares são G_5 e G_6 e portanto estes são aglutinados fornecendo o agrupamento G_{56} . Calculam-se as distâncias

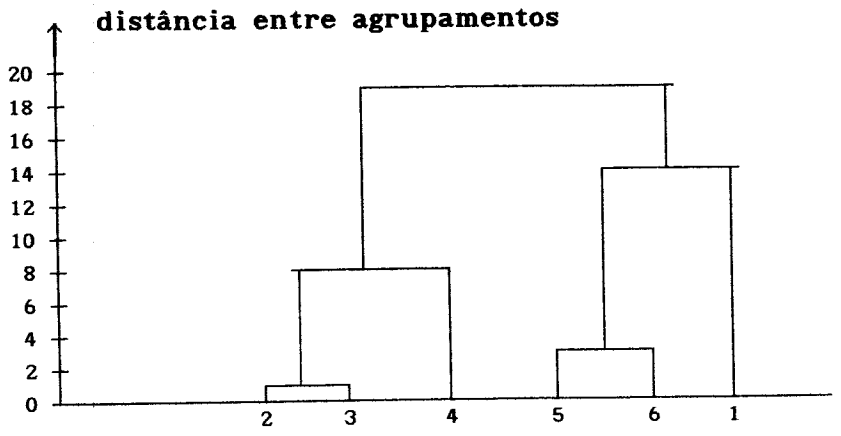
$d_{56,1} = 14$, $d_{56,23} = 19$, $d_{56,4} = 12$. A nova matriz D é :

$$D = \begin{array}{c} \\ 1 \\ 23 \\ 4 \\ 56 \\ \\ \end{array} \begin{array}{c} \begin{array}{cccc} & & & \\ \end{array} \\ \left[\begin{array}{cccc} 0 & & & \\ 15 & 0 & & \\ 16 & (9) & 0 & \\ 14 & 19 & 12 & 0 \end{array}\right] \\ \begin{array}{cccc} 1 & 23 & 4 & 56 \end{array} \end{array}$$

Aglutinam-se G_{23} e G_4 pois são os mais similares, obtendo-se G_{234} . Calculam-se as distâncias $d_{234,1} = 16$ e $d_{234,56} = 19$. A nova matriz D fica :

$$D = \begin{array}{c} \\ 1 \\ 234 \\ 56 \\ \\ \end{array} \begin{array}{c} \begin{array}{ccc} & & \\ \end{array} \\ \left[\begin{array}{ccc} 0 & & \\ 16 & 0 & \\ (14) & 19 & 0 \end{array}\right] \\ \begin{array}{ccc} 1 & 234 & 56 \end{array} \end{array}$$

Aglutina-se G_1 com G_{56} formando o agrupamento G_{156}. Calcula-se a distância $d_{156,234} = 19$, que é a distância entre os dois últimos a agrupamentos. O dendograma é visto abaixo.



ooooo

Uma observação é que nem sempre se consegue desenhar o dendograma sem mudar a ordem dos elementos no eixo horizontal, como aconteceu no dendograma anterior com o 1° elemento.

A comparação dos dendogramas acima indica que os agrupamentos 234 e 56 são resultados comuns aos dois algoritmos (vizinho mais próximo e vizinho

mais distante). A partir daí, há diferenças nos agrupamentos formados devido às diferentes medidas de distância entre agrupamentos adotadas pelos dois algoritmos.

Neste algoritmo de agrupamento por vizinho mais distante, quando agrupamentos G_i e G_j são aglutinados por terem uma distância mútua d_{ij} a menor dentre todos os outros pares de agrupamentos, garante-se que cada elemento do novo agrupamento G_{ij} não está mais distante que o valor d_{ij} de qualquer outro elemento de G_{ij} . É portanto possível desenhar uma hiper-esfera de diâmetro d_{ij} contendo todos os elementos de G_{ij} .

Este algoritmo não apresenta o inconveniente, existente no algoritmo hierárquico de vizinho mais próximo, de causar encadeamento de agrupamentos quando há um par de elementos muito próximos. Por exemplo, na Fig. 6 os dois agrupamentos G_1 e G_2 provavelmente teriam a sua identidade mantida pelo algoritmo hierárquico de vizinho mais distante. Entretanto, o algoritmo não é conveniente para representar agrupamentos alongados pois tende a impor estruturas compactas aproximadamente hiper-elipsoidais.

iii Algoritmo hierárquico de distância média ("average linkage")

A distância entre dois agrupamentos é definida neste algoritmo como sendo a média das distâncias de todos os pares de elementos, um elemento de um par tomado de cada agrupamento. No algorítmo aglomerativo, descrito no início da seção III, o início é semelhante, devendo-se atentar para a etapa iii que é diferente. Após aglutinar os agrupamentos G_α e G_β deve-se a seguir calcular todas as distâncias de $G_{\alpha\beta}$ a todos os demais agrupamentos. Para calcular a distância entre $G_{\alpha\beta}$ e G_γ procede-se como segue :

$$d_{\alpha\beta,\gamma} = \frac{\sum_{i}^{n_i} \sum_{k}^{n_k} d\left[\underline{x}(i), \underline{x}(k)\right]}{n_i \, n_k}$$

onde,

$\underline{x}(i) \in G_{\alpha\beta}$

$\underline{x}(k) \in G_{\gamma}$

n_i = número de elementos de $G_{\alpha\beta}$

n_k = número de elementos de G_{γ}

A definição de distância neste algoritmo é uma espécie de compromisso entre as definições nos dois algoritmos anteriores: não é nem a distância entre vizinhos mais próximos de ambos, nem a distancia entre vizinhos mais distantes de ambos; é uma distancia média calculada para todos os pares de elementos, tomando um de cada agrupamento.

ooooo Exemplo: É dada a matriz D empregada nos exemplos das seções anteriores:

D =

	1	2	3	4	5	6
1	0					
2	13	0				
3	15	(1)	0			
4	16	2	9	0		
5	10	8	8	11	0	
6	14	15	19	12	3	0

Aglutina-se G_2 com G_3 . Calculam-se as distâncias :

$$d_{23,1} = \frac{13+15}{2} = 14 \qquad d_{23,4} = \frac{2+9}{2} = 5{,}5$$

$$d_{23,5} = \frac{8+8}{2} = 8 \qquad d_{23,6} = \frac{15+19}{2} = 17$$

e obtém-se a nova matriz D :

$$D = \begin{array}{c} \\ 1 \\ 23 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{|c|c|c|c|c|} \hline 0 & & & & \\ \hline 14 & 0 & & & \\ \hline 16 & 5,5 & 0 & & \\ \hline 10 & 8 & 11 & 0 & \\ \hline 14 & 17 & 12 & (3) & 0 \\ \hline \end{array} \\ \begin{array}{ccccc} 1 & 23 & 4 & 5 & 6 \end{array} \end{array}$$

Aglutina-se G_5 com G_6 . Calculam-se as distâncias :

$$d_{56,1} = \frac{10+14}{2} = 12 \qquad d_{56,23} = \frac{8+8+15+19}{4} = 12,5$$

$$d_{56,4} = \frac{11+12}{2} = 11,5$$

A nova matriz D fica :

$$D = \begin{array}{c} \\ 1 \\ 23 \\ 4 \\ 56 \end{array} \begin{array}{c} \begin{array}{|c|c|c|c|} \hline 0 & & & \\ \hline 14 & 0 & & \\ \hline 16 & 5,5 & 0 & \\ \hline 12 & 12,5 & 11,5 & 0 \\ \hline \end{array} \\ \begin{array}{cccc} 1 & 23 & 4 & 56 \end{array} \end{array}$$

Aglutina-se agora G_{23} com G_4 obtendo-se G_{234}. As distâncias são:

$$d_{234,1} = \frac{13+15+16}{3} = 14,67$$

$$d_{234,56} = \frac{8+15+8+19+11+12}{3x2} = 12,17$$

A nova matriz D fica :

$$D = \begin{array}{c} \\ 1 \\ 234 \\ 56 \end{array} \begin{array}{c} \begin{array}{|c|c|c|} \hline 0 & & \\ \hline 14,67 & 0 & \\ \hline (12) & 12,17 & 0 \\ \hline \end{array} \\ \begin{array}{ccc} 1 & 234 & 56 \end{array} \end{array}$$

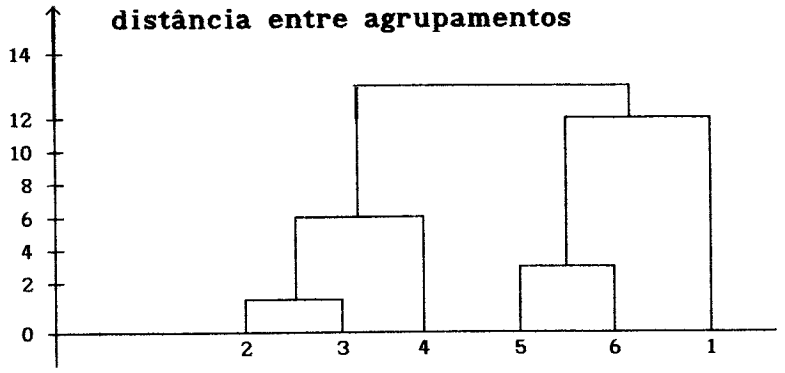
Aglutina-se G_1 com G_{56} obtendo-se G_{156} . Calcula-se

$$d_{156,234} = \frac{13+15+16+8+8+11+15+19+12}{3x3} = 13$$

A nova matriz D fica :

$$D = \begin{array}{c} 156 \\ 234 \end{array} \begin{bmatrix} 0 & \\ (13) & 0 \end{bmatrix}$$

Abaixo vemos o correspondente dendograma.



O dendograma resultante é semelhante ao obtido pelo algoritmo de vizinho mais distante embora os níveis de ordenada (distâncias entre agrupamentos) sejam diferentes. Esta semelhança entre esses dois algoritmos parece ser um achado frequente na prática. No caso presente, a indicação de haver 3 agrupamentos "naturais" é mais marcante que no algoritmo de vizinho mais distante.

ooooo

O algoritmo em discussão, além de exigir mais computações de distância que os dois anteriores, parece exigir o armazenamento da matriz D inicial até o final do algoritmo. Mas, isto não é verdade, pois basta tomar as distâncias médias da última matriz D determinada e multiplicá-las pelo produto dos números de elementos dos agrupamentos respectivos para achar a soma das distâncias, conforme visto no exemplo a seguir.

ooooo Exemplo: Repetir os cálculos de distâncias para o exemplo anterior :
(só serão feitos os cálculos que na resolução anterior exigiram voltar à matriz D original)

$$d_{56,23} = \frac{2 \times d_{5,23} + 2 \times d_{6,23}}{4} = \frac{2 \times 8 + 2 \times 17}{4} = 12,5$$

$$d_{234,1} = \frac{2 \times d_{23,1} + d_{4,1}}{3} = \frac{2 \times 14 + 16}{3} = 14,67$$

$$d_{234,56} = \frac{4 \times d_{23,56} + 2 \times d_{4,56}}{6} = \frac{4 \times 12,5 + 2 \times 11,5}{6} = 12,17$$

$$d_{156,234} = \frac{3 \times d_{1,234} + 6 \times d_{56,234}}{9} = \frac{3 \times 14,67 + 6 \times 12,17}{9} = 13$$

ooooo

Com isto encerra-se a seção sobre métodos hierárquicos neste texto, devendo-se ressaltar que há uma série de outras abordagens além daquelas dos vizinhos mais próximo e mais distante e da distância média, que são os mais populares. Os interessados em algoritmos adicionais devem consultar Jain e Dubes (1988), Everitt (1981), Anderberg (1973), dentre outros.

AGRUPAMENTO DE ATRIBUTOS OU VARIÁVEIS

Até agora só se abordou o problema de agrupar elementos mas pode-se também agrupar atributos. Uma finalidade pode ser a seleção de atributos pois atributos similares podem ser redundantes em termos de classificação ou análise de agrupamentos. Uma medida de distância que tem sido empregada para agrupar variáveis é o quadrado do coeficiente de correlação. O coeficiente de correlação entre as variáveis ou atributos i-ésimo e j-ésimo é :

$$\rho_{ij} = \frac{\sum_{n=1}^{N} (x_{in} - \overline{x_i})(x_{jn} - \overline{x_j})}{\sqrt{\sum_{n=1}^{N} (x_{in} - \overline{x_i})^2 \sum_{n=1}^{N} (x_{jn} - \overline{x_j})^2}}$$

onde

x_{in} representa a n-ésima amostra na i-ésima variável,

N é o número total de amostras

$\overline{x}_i$ é a média da i-ésima variável nas N amostras

Deve-se notar que as variáveis podem estar com unidades e escalas (ordens de grandeza, gama de variações) diferentes pois o coeficiente de correlação já efetua a divisão pelos desvios padrões. Coeficientes de correlação ao quadrado perto de 1 indicam grande similaridade ou proximidade e perto de 0 grande dissimilaridade.

Os três algoritmos hierárquicos apresentados se prestam muito bem para a finalidade de analisar agrupamentos de variáveis. Ao invés de trabalhar com uma matriz de distâncias D, utiliza-se uma matriz de similaridade S contendo os coeficientes de correlação ao quadrado entre os pares de variáveis. Pode-se analisar as correlações ao quadrado utilizando as amostras coletivamente, mas os resultados não serão muito úteis pois a estrutura das classes (no caso de se dispor de um conjunto de treinamento) ou agrupamentos de elementos não é levada em conta. A Fig. 7 serve como uma ilustração de casos em que a medida de correlação em um par de variáveis utilizando a totalidade das amostras, indistintamente, não traz bons resultados. No caso da Fig. 7a, a correlação entre x_1 e x_2, calculada para todas as amostras, é próxima de 1 muito embora na classe II a correlação entre x_1e x_2 é mais próxima de -1 . No caso da Fig. 7b, a correlação entre x_1 e x_2 vai resultar perto de zero muito embora dentro de cada classe as variáveis tem correlação próxima a 1 . Na Fig. 7c, as coordenadas x_1 e x_2 é que têm a maior correlação.(e caso se elimine x_2 fica-se com uma má separabilidade entre classes.

Do acima exposto, pode-se concluir que deve-se fazer análise de agrupamento de variáveis dentro de cada agrupamento de amostras e não na

coletividade das amostras. Caso um dado par de variáveis, por exemplo x_1 e x_2 na Fig. 7c, em todos os agrupamentos de amostras seja agrupado a um nivel alto (correlação ao quadrado próxima a 1) então em certos casos poder-se-ia aglutinar as duas variáveis do par. Pode-se tanto descartar uma das duas variáveis quanto formar uma outra como sendo a média das duas, muito embora neste caso não há seleção de atributos mas sim extração de atributos. No caso de descartar uma variável, não é fácil optar por uma delas, pois uma delas (x_2 na Fig. 7c) pode ser mais útil que a outra em termos de separabilidade entre classes. Mas a decisão de aglutinar duas variáveis apenas baseando-se na correlação ao quadrado próxima de 1 é perigosa. No caso da Fig. 7b, há mais agrupamentos que variáveis, e neste caso a eliminação de x_1 ou x_2 , ou sua combinação, traria péssimos resultados. Em certos casos, como da Fig. 7b, parece mais razoável efetuar uma regressão linear dentro de cada agrupamento e comparar os termos constantes das regressões (b em y = ax+b). Se $\rho_{ij}^2 \simeq 1$ e se os parâmetros a e b para todos os agrupamentos nas regressões entre x_i e x_j forem aproximadamente iguais então x_i e x_j poderiam ser aglutinados. Desta breve discussão baseada em exemplos simples, fica clara a necessidade de se empregar métodos melhores para seleção e extração de atributos. Alguns métodos já foram apresentados em capítulos anteriores.

VALIDAÇÃO DE AGRUPAMENTOS

Pode-se utilizar tanto critérios externos quanto internos para a etapa de validação de agrupamentos. Os critérios externos comparam a partição obtida com informação conhecida a priori sobre o problema. No caso mais completo, a atribuição de cada elemento (dada por um algoritmo de agrupamento) seria comparada com a sua classificação conhecida a priori. O intuito

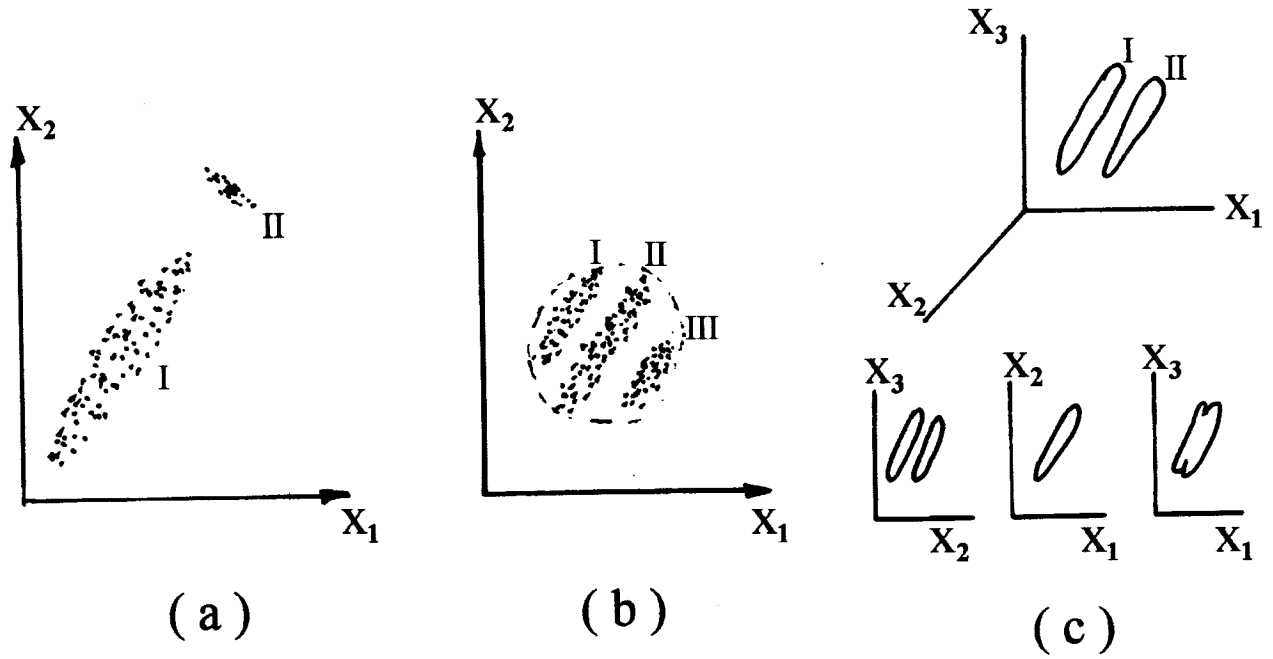


Fig. 7 –Agrupamento de variáveis por correlação.

neste caso seria testar o desempenho de diferentes métodos e algoritmos de agrupamento. Este tipo de avaliação de algoritmos de agrupamento está sendo realizado no Laboratório de Engenharia Biomédica da EPUSP voltado ao problema de classificação automática de potenciais unidades motoras. Inicialmente, um pesquisador utiliza um sistema de classificação manual de potenciais (Régis e Kohn, 1990). Em seguida diferentes algoritmos de agrupamento são empregados e para cada um deles estima-se, por exemplo, a taxa de erro global e as taxas de erro em cada classe.

Os critérios internos se baseiam apenas nos dados e testam, por exemplo, se a partição obtida é adequada face à estrututra das amostras, ou, qual de m partições melhor se adapta às amostras.

Há vários índices utilizados na literatura para testar a validade de uma partição. As dificuldades em conhecer a distribuição de um dado índice para poder quantificar o grau de adequação de uma partição muitas vezes levam à utilização de métodos de Monte Carlo para estimar distribuições e parâmetros.

Pode-se utilizar um índice igual a algum dos índices de qualidade já apresentados neste capítulo. Sua utilidadade maior é na análise relativa de diferentes partições. Por exemplo, pode-se agrupar um conjunto de dados empregando p diferentes algoritmos de agrupamento e compará-los através do índice $|\mathbf{W}|$ ou $\mathrm{tr}(\mathbf{B}\mathbf{W}^{-1})$. No caso do índice $|\mathbf{W}|$, a partição que resultar no menor índice pode ser a mais adequada. Entretanto, conforme já mencionado, cada índice tende a favorecer certas geometrias e é, portanto, perigoso empregá-los cegamente. Vale ressaltar que não se podem empregar testes de hipótese, para fins de avaliação de agrupamentos, apoiadas em estatísticas como $|\mathbf{W}|/|\mathbf{T}|$ (a estatística Λ de Wilks) pois:

i para valer a distribuição teórica da estatística sob a hipótese nula, deve-se ter a classificação correta de cada amostra, ou seja, cada

amostra tem que estar atribuída à população verdadeira;

ii a distribuição teórica é geralmente conhecida só para populações Gaussianas ou, no caso de populações arbitrárias, só em termos assintóticos.

iii a distribuição teórica geralmente supõe certas restrições adicionais como variâncias ou matrizes de covariância iguais.

LIVROS SOBRE RECONHECIMENTO DE PADRÕES E ÁREAS CORRELATAS

J. SCHURMANN, *"Pattern Classification"* NY:Wiley, 1996.

R. SCHALKOFF, *"Pattern Recognition"* NY:Wiley, 1992.

G. McLACHLAN, *"Discriminant Analysis and Statistical Pattern Recognition"*. NY : Wiley, 1992.

I.K. SETHI e A.K. JAIN, *"Artificial Neural Networks and Statistical Pattern Recognition: Old and New Connections"*. Amsterdam: Elsevier, 1991.

W.O. BUSSAB, E.S. MIAZAKI e D.F. de ANDRADE, *"Introdução à Análise de Agrupamentos"*. São Paulo: IME-USP, 1990.

K. FUKUNAGA, *"Introduction to Statistical Pattern Recognition"* (Second Edition). San Diego: Academic Press, 1990

Y.H. PAO, *"Adaptive Pattern Recognition and Neural Networks"*. Reading (MA): Addison Wesley, 1989.

C.W. THERRIEN, *"Decision, Estimation and Classification"*. NY: Wiley, 1989

A.L. JAIN e R.C.DUBES, *"Algorithms for Clustering Data"*. Englewood Cliffs: Prentice Hall, 1988

M. MINSKY e S. PAPERT, *"Perceptrons"* (Second Edition). Cambridge(MA): MIT Press, 1988.

P.A. DEVIJVER e J.KITTLER, (Eds) *"Pattern Recognition Theory and Applications"*. NY: Springer Verlag, 1987.

A. COHEN, *"Biomedical Signal Processing, volume II: Compression and Automatic Recognition"*. Boca Raton (FL): CRC Press, 1986.

S.H.C. du TOIT, A.G.W. STEYN e R.H. STUMPF, *"Graphical Exploratory Data Analysis"*. NY: Springer Verlag, 1986.

I.T. JOLLIFFE, *"Principal Component Analysis"*. NY: Springer-Verlag, 1986.

E.A. PATRICK e J.M. FATTU, *"Artificial Intelligence with Statistical*

Pattern Recognition". Englewood Cliffs : Prentice Hall, 1986.

B.W. SILVERMAN, "*Density Estimation for Statistics and Data Analysis*". London : Chapman & Hall, 1986.

S. WATANABE, "*Pattern Recognition : Human and Mechanical*". NY: Wiley, 1985.

S.T. BOW, "*Pattern Recognition* ". NY: Marcel Dekker, 1984.

L.B. BREIMAN, J.H. FRIEDMAN, R.A.OLSHEN e C.J. STONE, "*Classification and Regression Trees*". Belmont (CA): Wadsworth, 1984.

M. JAMBU e M-O. LEBEAUX, "*Cluster Analysis and Data Analysis*". Amsterdam: North Holland, 1983.

J. JANSSEN, J.F. MARCOTORCHINO e J.M. PROTH, "*New Trends in Data Analysis and Applications*". Amsterdam : North-Holland, 1983.

P.A DE VIJVER e J.KITTLER, "*Pattern Recognition. A Statistical Approach*". Englewood Cliffs : Prentice - Hall, 1982.

K.S. FU, "*Syntactic Pattern Recognition and Applications*". Englewood Cliffs : Prentice-Hall, 1982.

P.R. KRISHNAIAH e L.N.KANAL (Editores), "*Handbook of Statistics. 2 - Classification, Pattern Recognition and Reduction of Dimensionality*". Amsterdam : North-Holland, 1982.

J.C. BEZDEK, "*Pattern Recognition with Fuzzy Objective Function Algorithms*". NY : Plenum, 1981.

B. EVERITT, "*Cluster Analysis*". NY : Wiley, 1981

D.J. HAND, "*Discrimination and Classification*". NY : Wiley, 1981.

J. SKLANSKY e G.N. WASSEL, "*Pattern Classifiers and Trainable Machines*". NY : Springer - Verlag, 1981.

K.S. FU (Editor), "*Digital Pattern Recognition*" (Second Edition). Berlin : Springer-Verlag, 1980

Y.T. CHIEN, "*Interactive Pattern Recognition*". NY: Marcel-Dekker, 1978.

R.C. GONZALEZ e M.G. THOMASON, "*Syntactic Pattern Recognition*". Reading

(MA): Addison Wesley, 1978.

A.K. AGRAWALA, (Editor) *"Machine Recognition of Patterns"*. NY :IEEE Press, 1977. Obs.: É uma coletânea de importantes artigos até 1975.

T. PAVLIDIS, *"Structural Pattern Recognition"*. Berlin : Springer- Verlag, 1977.

J.A.HARTIGAN, *"Clustering Algorithms"*. NY : Wiley, 1975.

P.A.LACHENBRUCH, *"Discriminant Analysis"*. NY : Hafner, 1975.

K.S.FU, *"Syntactic Methods in Pattern Recognition"*. NY : Academic Press, 1974.

J.T.TOU e R.C.GONZALEZ, *"Pattern Recognition Principles"*. Reading (MA) : Addison - Wesley, 1974.

T.Y. YOUNG e T.W. CALVERT, *"Classification, Estimation and Pattern Recognition"*. NY : Elsevier, 1974.

M.R.ANDERBERG, *"Cluster Analysis for Applications"*. NY : Academic Press, 1973.

R.O. DUDA e P.E. HART, *"Pattern Classification and Scene Analysis"*. NY : Wiley, 1973.

H.C. ANDREWS *"Introduction to Mathematical Techniques in Pattern Recognition"*. NY : Wiley, 1972.

K.S.FU, *"Sequential Methods in Pattern Recognition and Machine Learning"*. NY : Academic, 1968.

F. ROSENBLATT *"Principles of Neurodynamics"*. Washington : Spartan, 1962.

LIVROS QUE COBREM VETORES ALEATÓRIOS, ESTATÍSTICA E OUTRAS BASES MATEMÁTICAS

P.G. HOEL, S.C. PORT e C.J. STONE, *"Introduction to Probability Theory"*. Boston: Houghton Mifflin, 1971.

A. PAPOULIS, *"Probability, Random Variables and Stochastic Processes"* (Second Edition).NY : McGraw - Hill, 1984.

OBS.: Existe edição internacional. Utiliza vetores linha.

J.L. MELSA e A.P.SAGE, *"An Introduction to Probability and Stochastic Processes"*. Englewood Cliffs : Prentice Hall, 1973.

P.G. HOEL, S.C. PORT e C.J. STONE, *"Introduction to Statistical Theory"*. Boston: Houghton Mifflin, 1971.

G.G. ROUSSAS, *"A First Course in Mathematical Statistics"*. Reading (MA): Addison-Wesley, 1973.

P.J. BICKEL e K.A. DOKSUM, *"Mathematical Statistics"*. San Francisco: Holden-Day, 1977.

E.J. DUDEWICZ e S.N. MISHRA, *"Modern Mathematical Statistics"*. NY : Wiley, 1988.

A.M. MOOD, F.A. GRAYBILL e D.C. BOES, *"Introduction to the Theory of Statistics"* (Third Edition). N.Y.:McGraw-Hill, 1974.

D.F. MORRISON, *"Multivariate Statistical Methods"* (Second Edition). NY: McGraw-Hill, 1976. (OBS.: Existe edição internacional).

R.A. JOHNSON e D.W.WICHERN, *"Applied Multivariate StatisticalAnalysis"* (Second Edition). Englewood Cliffs : Prentice Hall, 1988.

(OBS: Existe edição internacional)

K.V. MARDIA, J.T. KENT e J.M.BIBBY, *"Multivariate Analysis"*. London: Academic Press, 1979. (OBS.: Existe edição com capa mole)

P.E. GREEN e J.D.CARROL, *"Mathematical Tools for Applied Multivariate Analysis"*. NY : Academic Press, 1976.
(OBS.: Existe edição com capa mole)

P.S. MAYBECK, *"Stochastic Models, Estimation, and Control*, vol. 1:Orlando : Academic Press, 1979.

J.S.MEDITCH, *"Stochastic Optimal Linear Estimation and Control"*.NY : McGraw-Hill, 1969.

G. STRANG, *"Linear Algebra and its Applications"*. NY : Academic Press 1980.

B. NOBLE e J.W. DANIEL, *"Applied Linear Algebra"* (Third Edition). Englewood Cliffs: 1988.
(OBS: Existe edição internacional)

S.R. SEARLE, *"Matrix Algebra Useful for Statistics"*. NY : Wiley, 1982

J.SPANIER e K.B.OLDHAM, *"An Atlas of Functions"*. NY : Hemiaphere, 1987.

REVISTAS

IEEE Transactions on Pattern Analysis and Machine Intelligence

Pattern Recognition

Pattern Recognition Letters

Journal of the American Statistical Association

Biometrics

Journal of Classification

Applied Statistics

IEEE Transactions on Systems, Man and Cybernetics

IEEE Transactions on Biomedical Engineering

IEEE Transactions on Information Theory

IEEE Transactions on Computers

Biometrika

Techometrics

Annals of Statistics

Annals of Mathematical Statistics