AMS Classification: 62H30, 68T10.

1 Introdução

Formalmente, dado um objecto na forma de um vector com características $\mathbf{X} = (X_1, X_2, ..., X_p) \in \chi$, um conjunto de c classes distintas $G_1, G_2, ..., G_c$ e uma amostra de treino $(\mathbf{X}_1, G_i)$, $(\mathbf{X}_2, G_j)$, ..., $(\mathbf{X}_n, G_k)$ com i, j, k =1,2,...,c, a tarefa da classificação consiste em estimar a verdadeira classe depois de observado $\mathbf{X}$.

O classificador divide o espaço das características em regiões de decisão, $D_1, D_2, ..., D_c$, às quais se associam as diversas classes. Estas regiões podem ser obtidas por optimização de um determinado critério. Exemplos de critérios utilizados são a maximização da probabilidade *a posteriori* ou a minimização do custo total esperado de classificação incorrecta. O aspecto crucial desta partição são as hipersuperfícies ou fronteiras de decisão, designadas por superfícies de decisão. As superfícies são expressas em termos de funções de decisão ou funções discriminantes as quais irão ser designadas genericamente por $d(\mathbf{X})$.

Dado um classificador definido pelas regiões de decisão D_1 , D_2 , ... D_c então a regra de decisão consiste em classificar um novo objecto com vector de características $\mathbf{X}$ na classe G_i i=1,2,...,c se $d_i(\mathbf{X}) > d_j(\mathbf{X})$, $\forall j \neq i$, ou ainda, $i = \arg\max d_j(\mathbf{X})$, j=1,2,...,c.

De uma forma mais genérica pode-se dizer que um classificador é uma aplicação: $\chi \rightarrow \{ G_1, G_2, ..., G_c, I, O\}$ onde I representa indecisão e O observações atípicas (*outliers*). Note-se que I e O constituem dúvida podendo, qualquer uma delas, ser estudada na forma de classe de rejeição. Nesta situação as regiões de decisão passarão a denotar-se $D_1^*, D_2^*, ..., D_c^*, D_I , D_O$. Neste momento apenas se irá tratar da rejeição no sentido da indecisão deixando-se para trabalho futuro o problema das observações atípicas.

2 A questão da rejeição

Frequentemente é preferível não classificar, i.é, decidir pela rejeição, do que optar por uma decisão com elevada probabilidade de classificação incorrecta; são geralmente as observações "incertas" (entende-se por observações incertas aquelas para as quais $\exists i,j, i \neq j$: $d_i(\mathbf{X}) \approx d_j(\mathbf{X})$) que mais contribuem para o número de classificações incorrectas (ou má classificação). Os inconvenientes causados pela rejeição devem ser sempre julgados contra as consequências de uma má classificação embora na prática as vantagens ou desvantagens só possam ser avaliadas tomando em consideração a aplicação específica.

Há aplicações para as quais o custo de má classificação (custo de classificar um objecto na classe G_i quando de facto pertence à classe G_j) é muito grande e portanto uma rejeição de observações elevada é aceitável, tornando o número de más classificações o menor possível; exemplos típicos (Leondes, 1998) são aqueles que surgem no campo da medicina, tal como a apreciação de imagens

para detecção de cancro. Noutras aplicações em que por exemplo há controlo humano posterior, é preferível classificar sempre numa classe, mesmo com risco de elevado número de más classificações. Uma escolha adequada da regra que conduz à rejeição permite que o comportamento do classificador se sintonize com a aplicação em causa.

A ideia consiste em utilizar uma constante que actua como um limiar de segurança contra um número excessivo de erros de classificação e que poderá ser especificado pelo utilizador ou escolhido após a experimentação de diversos valores (se possível numa *amostra de teste*) obtendo-se estimativas da probabilidade de classificação incorrecta e da probabilidade de rejeição (taxa de erro e taxa de rejeição, respectivamente).

Embora a opção de rejeição possa aliviar ou remover o problema do elevado número de más classificações, algumas observações correctamente classificadas podem ser convertidas em rejeição. Não esqueçamos contudo que rejeição pode não significar exclusão mas apenas o suster da decisão para tratamento posterior.

Chow foi o primeiro a introduzir a opção de rejeição num problema de classificação (Chow, 1957,1970). No primeiro artigo formulou uma regra de decisão óptima. No segundo determinou uma relação fundamental entre as taxas de erro e rejeição e apresentou as curvas de compromisso erro/rejeição, que podem e devem ser utilizadas para descrever e comparar a *performance* empírica dos métodos de classificação.

3 Classificador do máximo *a posteriori*

Assumindo-se conhecidas as probabilidades *a priori* das classes, π_i i=1,2,...,c , as funções densidade de probabilidade condicionais às classes, $f_i(\mathbf{X})$ (ou o seu rácio) e custos iguais, vem pelo teorema de Bayes,

$$q_i(\mathbf{X}) = P(G_i/\mathbf{X}) = \frac{\pi_i f_i(\mathbf{X})}{\sum_k \pi_k f_k(\mathbf{X})} \qquad (1)$$

tendo-se a seguinte regra de decisão: $\mathbf{X} \in D_i$ se $q_i(\mathbf{X}) > q_j(\mathbf{X})$ $\forall$ j≠i i,j = 1,2, ..., c ou equivalentemente, $\mathbf{X} \in D_i$ se $\pi_i f_i(\mathbf{X}) > \pi_j f_j(\mathbf{X})$ $\forall$ j≠i i,j = 1,2, ..., c

Chow (1970) considera que uma regra é óptima se, para uma dada taxa de erro, minimiza a taxa de rejeição. A regra t-óptima de Chow, consiste em rejeitar o objecto se o máximo das probabilidades *a posteriori* é inferior a um dado limiar de rejeição t (*rejection threshold*) o que dá origem à seguinte regra de decisão:

⇒ Aceitação do objecto

$$\mathbf{X} \in D_i^* \quad \text{se} \quad \pi_i f_i(\mathbf{X}) > \pi_j f_j(\mathbf{X}) \tag{2}$$
$$\text{e} \quad \pi_i f_i(\mathbf{X}) \geq (1-t)\sum_{j=1}^{c} \pi_j f_j(\mathbf{X}) \quad \forall j \neq i \quad i,j = 1,2,...,c$$

$\Rightarrow$ Rejeição do objecto caso contrário, se $\max_i [\pi_i f_i(\mathbf{X})] < (1-t) \sum_{j=1}^{c} \pi_j f_j(\mathbf{X})$.

3.1 Acerca do limiar de rejeição

O procedimento adoptado não é mais do que o de particionar o espaço das características numa região $A = \bigcup_{i=1}^{c} D_i^*$ de aceitação e numa região D_I de rejeição da seguinte forma: $A = \{ \mathbf{X} : 1 - \max_i \ q_i(\mathbf{X}) \leq t \}$ e $D_I = \{ \mathbf{X} : 1 - \max_i \ q_i(\mathbf{X}) > t \}$, de onde se pode concluir que quanto menor for t, maior é a região de rejeição. Em 1970, Chow foi ainda mais longe ao demonstrar que tanto a taxa de erro, E, como a taxa de rejeição, R, são ambas funções monótonas do limiar t. Segue de imediato que E aumenta e R diminui à medida que t aumenta. Em particular quando t=0, E=0 e quando t=1, R=0. Devijver e Kittler (1982) referem que para c classes $0 \leq 1 - \max_i q_i(\mathbf{X}) \leq \frac{c-1}{c}$ pelo que, para a opção de rejeição ser activada, é necessário que $0 \leq t \leq \frac{c-1}{c}$.

Para a situação mais simples em se que trabalha apenas com duas classes a condição de rejeição, atrás referida, nunca é activada para $t > \frac{1}{2}$. Neste caso $0 \leq t \leq \frac{1}{2}$ tornando-se óbvio constatar que quando $t=\frac{1}{2}$ cai-se no caso usual de inexistência de rejeição e quando t=0 todos os objectos são rejeitados.

4 Regra de decisão para duas classes

Tomando por base a regra t-óptima de Chow anteriormente apresentada (conjunto de fórmulas (2)), a regra de decisão para duas classes e para as discriminantes linear, quadrática e logística pode ser sintetizada da seguinte forma:

$$\begin{array}{lll} \mathbf{X} \in D_1^* & \text{se} & d(\mathbf{X}) \geq t_1 \\ \mathbf{X} \in D_2^* & \text{se} & d(\mathbf{X}) \leq t_2 \\ \mathbf{X} \in D_I & \text{se} & t_2 < d(\mathbf{X}) < t_1 \end{array}$$

onde t_1 , t_2 , e $d(\mathbf{X})$ são para cada uma das discriminantes, definidos pelas expressões que se passam a indicar.

- Discriminante Linear (**LDA**)

$\mathbf{X}/G_i \sim N_p(\mu_i, \Sigma)$, i=1,2

$d(\mathbf{X}) = (\mu_1 - \mu_2)^T \Sigma^{-1} \mathbf{X} - \frac{1}{2}(\mu_1 - \mu_2)^T \Sigma^{-1}(\mu_1 + \mu_2)$

$t_1 = \ln\left[\frac{\pi_2}{\pi_1} \times \frac{1-t}{t}\right] \quad t_2 = \ln\left[\frac{\pi_2}{\pi_1} \times \frac{t}{1-t}\right]$

• Discriminante Quadrática (**QDA**)

$\mathbf{X}/G_i \sim N_p(\mu_i, \Sigma_i)$ i=1,2

$d(\mathbf{X}) = -D_1(\mathbf{X}) + D_2(\mathbf{X})$

onde $D_i(\mathbf{X})$ representa o quadrado da distância de Mahalanobis entre $\mathbf{X}$ e a média da i-ésima classe

$t_1 = 2\ln\left[\frac{\pi_2}{\pi_1} \times \frac{1-t}{t}\right] - \ln\left(\frac{|\Sigma_2|}{|\Sigma_1|}\right) \quad t_2 = 2\ln\left[\frac{\pi_2}{\pi_1} \times \frac{t}{1-t}\right] - \ln\left(\frac{|\Sigma_2|}{|\Sigma_1|}\right)$

• Discriminante Logística

O logaritmo da razão entre as densidades condicionais às classes é linear.

$d(\mathbf{X}) = \alpha_0 + \alpha^T \mathbf{X} = \ln\left(\frac{f_1(\mathbf{X})}{f_2(\mathbf{X})}\right)$

$t_1 = \ln\left[\frac{\pi_2}{\pi_1} \times \frac{1-t}{t}\right] \quad t_2 = \ln\left[\frac{\pi_2}{\pi_1} \times \frac{t}{1-t}\right]$

5 *Performance* do classificador

Descreve-se em seguida os processos utilizados para a obtenção das taxas de erro e rejeição para cada uma das discriminantes.

• Discriminante Linear (LDA)

Nesta situação podem-se obter os valores de E e R invocando um resultado básico da estatística multivariada: Se $\mathbf{X} \in G_1$ então $d(\mathbf{X}) \sim N(\frac{1}{2}\Delta, \Delta)$ e se $\mathbf{X} \in G_2$ então $d(\mathbf{X}) \sim N(-\frac{1}{2}\Delta, \Delta)$ onde $\Delta = (\mu_1 - \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2)$.

• Discriminante Quadrática (QDA)

Na impossibilidade de se recorrer a qualquer resultado teórico, optar-se-á por:

- estimar E e R através da amostra teste, caso esta exista.
- estimar E e R através de uma amostra simulada a partir de uma distribuição normal multivariada, com parâmetros iguais aos valores amostrais das médias e matrizes de covariâncias associadas a cada classe.

• Discriminante Logística

Caso possamos admitir que $d(\mathbf{X})$ segue aproximadamente uma distribuição Normal então pode-se recorrer ao resultado que se segue para a obtenção de E e R. Se $\mathbf{X} \in G_i$, $d(\mathbf{X}) \sim N(\alpha_0 + \alpha^T \mu_i, \alpha^T \Sigma_i \alpha)$. Caso contrário tem de se proceder como na discriminante quadrática.

6 Estimação

Até agora admitiu-se que as probabilidades *a priori* e as densidades de probabilidade condicionais às classes eram conhecidas. No entanto, tanto nos exemplos apresentados como na maior parte das situações reais, este não é o caso.

Para estimação das probabilidades *a priori* optou-se pelo cálculo da frequência relativa de cada classe no seio da amostra de treino, i.e. $\hat{\pi}_i = \frac{n_i}{n} \quad i = 1, 2$, onde n_i representa o número de objectos na amostra de treino que pertencem à i-ésima classe e n a dimensão da amostra de treino.

Quanto à estimação dos parâmetros que caracterizam cada uma das discriminantes aqui abordadas, utilizou-se como estimativa da média da i-ésima classe, a média amostral. No que diz respeito à matriz de covariâncias, no caso da discriminante linear, utilizou-se a matriz de covariâncias combinada usual,

$$\hat{\Sigma} = \mathbf{S} = \frac{(n_1 - 1)\,\mathbf{S}_1 + (n_2 - 1)\,\mathbf{S}_2}{n - 2}$$

onde $\mathbf{S}_1$ e $\mathbf{S}_2$ representam as matrizes de covariâncias amostrais de cada classe.

No caso das discriminantes quadrática e logística as matrizes de covariâncias de cada classe, Σ_i, foram estimadas através das correspondentes matrizes amostrais.

7 Aplicações

7.1 Descrição dos conjuntos de dados

CHORON- Detecção de doença coronária (Macieira-Coelho, et al. ,1990)

. Amostra: 113 indivíduos
. Variáveis: 9 (4 variáveis clínicas e 5 relativas a provas de esforço)
. Classes:
  G_1- não sofre de doença coronária (30 indivíduos)
  G_2- sofre de doença coronária (83 indivíduos)
. Amostra teste: indisponível
. Amostra "simulada": 1000 observações para cada uma das classes

→ Posteriormente foram retiradas duas variáveis (ver Pires, et al., 1995) passando-se a trabalhar apenas com sete (do tipo binário, ordinal e contínuo).

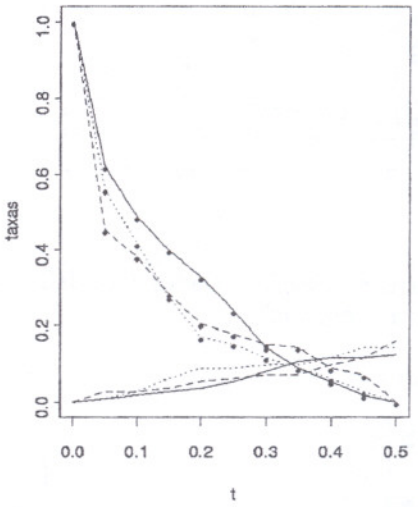
PIMA- Detecção de diabetes em mulheres Índias com mais de 21 anos da tribo "Pima", residentes no Arizona (Ripley,1996)

. Amostra: 200 mulheres
. Variáveis: 7
. Classes:
  G_1- tem diabetes (132 mulheres)
  G_2- não tem diabetes (68 mulheres)
. Amostra teste: 332 mulheres
. Amostra "simulada": 1000 observações para cada uma das classes

→ Uma selecção de variáveis pelo método *stepwise* (Ripley,1996) rejeitou duas passando-se a trabalhar apenas com 5 variáveis (número de vezes que esteve grávida / concentração de glucose no sangue / índice de massa muscular / concentração de insulina / idade)

7.2 Alguns resultados

Apresentam-se em seguida algumas figuras correspondentes às curvas de compromisso erro/rejeição em função do limiar t.



onde:

LDA erro — QDA erro — Logística erro
LDA rejeição — QDA rejeição — Logística rejeição

Figura 1: Taxas de erro e rejeição. Dados CHORON. Amostra de treino.

Fixando t pode-se fazer uma leitura dos valores de E e R. Por exemplo nos dados PIMA, amostra teste (Figura 3) e na discriminante logística, para t=0.3 reduz-se a taxa de erro para cerca de metade: passa de 19,6% para 9,6% , mas temos uma taxa de rejeição de 25% . Se permitirmos apenas que cerca de 15% das observações sejam rejeitadas, já baixamos a taxa de erro de 19,6% para 13,5% .

Um outro aspecto que pode ser tomado em consideração e que ajudará a encontrar o melhor valor de t é o conhecimento dos custos relativos do erro e

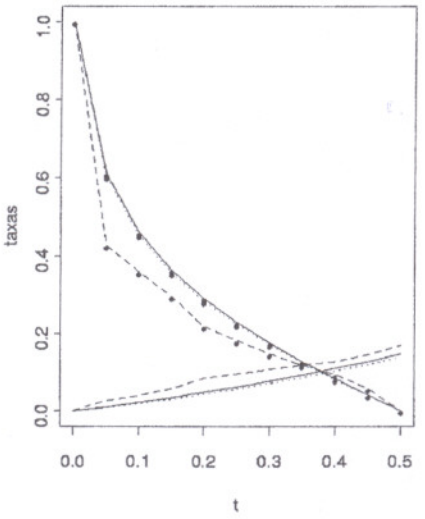


Figura 2: Taxas de erro e rejeição. Dados CHORON. LDA e Logística - "teóricas". QDA - amostra "simulada".

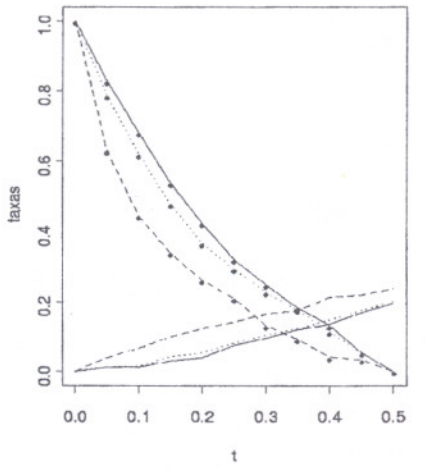


Figura 3: Taxas de erro e rejeição. Dados PIMA. Amostra de teste.

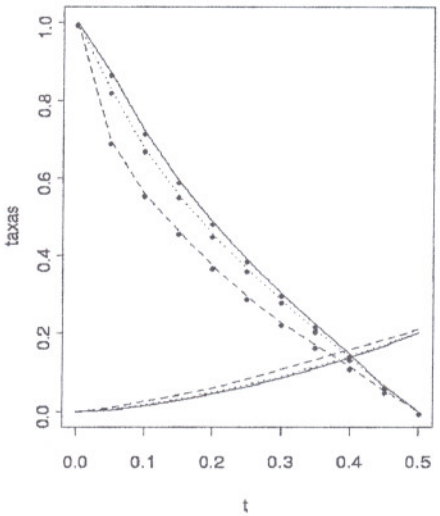


Figura 4: Taxas de erro e rejeição. Dados PIMA. LDA e Logística - "teóricas". QDA - amostra "simulada".

da rejeição. Este assunto irá ser alvo de trabalho futuro.

Bibliografia

[1] Chow, C.K. (1957). An optimum character recognition system using decision functions. *IRE Trans. Electron. Computers* 6:247-254.

[2] Chow, C.K. (1970). On optimum recognition error and reject tradeoff. *IEEE Trans. Inform. Theory*16:41-46.

[3] Devijver, P.A. e Kittler J. (1982). *Pattern Recognition: a Statistical Approach.* Prentice-Hall International, London.

[4] Leondes, C. T. (1998). *Image Processing and Pattern Recognition*. Academic Press, U.S.A

[5] Macieira-Coelho, et al. (1990). Diagnóstico de cardiopatia isquémica no doente ambulatório: análise multivariada de dados clínicos e electrocardiográficos. *Acta Médica Portuguesa*3, 277-282.

[6] Pires, A.M., Branco, J.A. e Amaral-Turkman, M.A. (1995). Comparação de métodos de análise discriminante no diagnóstico da doença coronária. Em Mendes-Lopes et al. (editores) *Actas do II Congresso Anual da Sociedade Portuguesa de Estatística. Coimbra*.pp. 319-343.

[7] Ripley, B.D. (1996). *Pattern Recognition and Neural Networks*. Cambrige University Press, U.K.