MÉTODOS ESTATíSTICOS EM DATA MINING
- Folha 4: Análise Discriminante Linear e Quadrática -

33. Considere o conjunto de dados salmon (ver página web da cadeira em http://www.fc.up.pt/pessoas/jpcosta/MEDM.html). Suponha que estamos interessados em distinguir os peixes do Alasca dos peixes Canadianos. Escolha aleatóriamente 20% dos peixes e deixe-os de lado. Considere o modelo normal e determine a regra linear e a regra quadrática com os dados restantes. Determine uma estimativa da taxa de erro de ambas as regras, usando os dados deixados de lado (ignore a variável género). Repita a análise anterior desta vez usando a variável género codificada com 0 para um dos sexos e 1 para o outro.

34. Considere o conjunto de dados "wine" (http://archive.ics.uci.edu/ml/ ou página da disciplina) contendo análises químicas de vinhos produzidos na mesma região de Itália, mas de três tipos diferentes. Este conjunto de dados têm 14 atributos, sendo o primeiro a classe e os restantes atributos predictivos(Alcohol, Malic acid, Ash, Alcalinity of ash, Magnesium, Total phenols, Flavanoids, Nonflavanoid phenols, Proanthocyanins, Color intensity, Hue, OD280/OD315 of diluted wines, Proline).

    (a) Começe por fazer um diagrama de dispersão das variáveis predictivas e, para ilustração gráfica, escolha as duas que lhe parecem mais predictivas de acordo com os gráficos.

    (b) Corra a Análise Discriminante Linear com essas duas variáveis.

    (c) Faça um diagrama de dispersão dessas duas variáveis com cores diferentes para as classes e desenha as fronteiras da LDA.

    (d) Repita a alínea anterior para QDA.

35. Considere o conjunto de dados spam (library(kernlab) ou página da disciplina) contendo informação sobre 4601 e-mails (1813 spam e 2788 nonspam). Corra a regressão logística nestes dados e verifique que variáveis são importantes. Analise os resultados previstos pelo modelo considerando que sempre que a probabilidade é superior a 0.5, temos spam; faça o mesmo para probabilidade=0.95. Dado que os resultados anteriores foram obtidos no mesmo conjunto de dados, não são fiáveis. Divida aleatoriamente os dados em treino (70%) e teste (30%)e corra a RL no treino. Use o modelo obtido para prever os valores do teste (para assim termos mais confiança nos resultados) e também no treino, para comparar os erros (deve-se esperar que o erro no teste seja maior). Compare com o valor obtido por LDA. Para uma melhor comparação, use as curvas ROC, isto é o gráfico da probabilidade de se detectar os verdadeiros positivos (sensibilidade) versus os falsos positivos (1 - especificidade) ou versus a especificidade (verdadeiros negativos). A área sob a curva ROC, que varia entre 0 e 1, dá uma medida da capacidade do modelo em discriminar entre os dois valores.

36. Duas distribuições normais são especificadas pelos seguintes parâmetros:
$\mu_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$; $\mu_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$; $\Sigma_1 = \Sigma_2 = \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$; $\pi_1 = \pi_2 = 0.5$

    (a) Calcule o erro do classificador de Bayes e desenhe esse classificador.

    (b) Gere 2 amostras de tamanho 100 para cada classe e suponha que desconhecia μ_1, μ_2 e Σ.

        i. Utilize os EMV na regra de Bayes e calcule o erro aparente.

        ii. Calcule uma estimativa do erro usando uma parte para treino e outra para teste.

37. Para uma amostra aleatória de tamanho n, $X_1, X_2, \ldots, X_n$, considere a média e matriz de variância amostrais:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_k; \qquad \hat{\Sigma}_n = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X}_n)(X_k - \bar{X}_n)^T$$

Mostre que o efeito de adicionar uma nova amostra X_{n+1} pode ser calculado pelas relações recursivas

$$\bar{X}_{n+1} = \bar{X}_n + \frac{1}{n+1}(X_{n+1} - \bar{X}_n) \text{ e}$$

$$\hat{\Sigma}_{n+1} = \frac{n-1}{n}\hat{\Sigma}_n + \frac{1}{n+1}(X_{n+1} - \bar{X}_n)(X_{n+1} - \bar{X}_n)^T$$

38. As relações dadas no exercicio anterior são uteis para actualizar estimativas da matriz de covariâncias. No entanto, frequentemente, estamos também interessados na inversa dessa matriz, o que é mais demorado. Começando por demonstrar a identidade matricial

$$(A + XX^T)^{-1} = A^{-1} - \frac{A^{-1}XX^TA^{-1}}{1 + X^TA^{-1}X}$$

e usando os resultados do problema anterior, mostre que

$$\hat{\Sigma}_{n+1}^{-1} = \frac{n}{n-1}\left[\hat{\Sigma}_n^{-1} - \frac{\hat{\Sigma}_n^{-1}(X_{n+1} - \bar{X}_n)(X_{n+1} - \bar{X}_n)^T\hat{\Sigma}_n^{-1}}{\frac{n^2-1}{n} + (X_{n+1} - \bar{X}_n)^T\hat{\Sigma}_n^{-1}(X_{n+1} - \bar{X}_n)}\right]$$

39. A expressão

$$J_1 = \frac{1}{n_1 n_2}\sum_{i\in\omega_1}\sum_{j\in\omega_2}(y_i - y_j)^2$$

mede claramente a dispersão entre-grupos de duas amostras, uma contendo n_1 observações da classe w_1 e a outra contendo n_2 observações da classe w_2. De forma análoga,

$$J_2 = \frac{1}{n_1^2}\sum_{i\in\omega_1}\sum_{j\in\omega_1}(y_i - y_j)^2 + \frac{1}{n_2^2}\sum_{i\in\omega_2}\sum_{j\in\omega_2}(y_i - y_j)^2$$

mede claramente a dispersão total dentro-dos-grupos. Mostre que $J_1 = (\mu_1 - \mu_2)^2 + \frac{1}{n_1}s_1^2 + \frac{1}{n_2}s_2^2$ e $J_2 = \frac{2}{n_1}s_1^2 + \frac{2}{n_2}s_2^2$, onde s_i^2 é a dispersão da classe ω_i.

40. Descreva o método de análise discriminante linear de Fisher.

41. Seja

$$S_B = \sum_{i=1}^{K} n_i(\bar{X}_i - \bar{X})(\bar{X}_i - \bar{X})^t$$

a matriz de dispersão entre-grupos para o caso de c classes. Mostre que $S_B = [(n_1n_2)/n](\bar{X}_1 - \bar{X}_2)(\bar{X}_1 - \bar{X}_2)^t$ para o caso de $K = 2$.

42. * Considere um problema com duas classes em que projecta os dados num espaço de dimensão 1, $Y = V^TX + v_0$, de forma a optimizar um critério de separação entre as classes projectadas.

    (a) Mostre que os valores de V e v_0 que optimizam o critério $f(\hat{m}_1, \hat{m}_2, s_1^2, s_2^2) = \frac{\pi_1\hat{m}_1^2 + \pi_2\hat{m}_2^2}{\pi_1 s_1^2 + \pi_2 s_2^2}$ são $V = (\pi_1\hat{\Sigma}_1 + \pi_2\hat{\Sigma}_2)^{-1}(\hat{\mu}_2 - \hat{\mu}_1)$ e $v_0 = -V^T(\pi_1\hat{\mu}_1 + \pi_2\hat{\mu}_2)$

(b) Faça o mesmo para o critério de Fisher (1936), $f(\hat{m}_1, \hat{m}_2, s_1^2, s_2^2) = \frac{(\hat{m}_1 - \hat{m}_2)^2}{s_1^2 + s_2^2}$. Este critério não depende de v_0 e por isso na prática é comum tomar para v_0 o ponto médio: $v_0 = (\hat{m}_1 + \hat{m}_2)/2$

43. Considere uma máquina linear com funções discriminantes $g_i(x) = w_i^t x + w_{io}$, $i = 1, \ldots, c$. Mostre que as regiões de decisão são convexas mostrando que se $x_1 \in \mathcal{R}_i$ e $x_2 \in \mathcal{R}_i$, então $\lambda x_1 + (1 - \lambda)x_2 \in \mathcal{R}_i$, com $0 \leq \lambda \leq 1$.

44. Regressão Linear como classificador. Seja $Y = \begin{cases} n_2/n & \text{para o grupo 1} \\ -n_1/n & \text{para o grupo 2} \end{cases}$

Considere uma regressão linear da variável binária Y em $(X - \overline{X})$ e que passe pela origem (é mais fácil começar por centrar os dados e portanto sssumir que $\overline{X} = 0$ e fazer a regressão em X em vez de $(X - \overline{X})$).

(a) Mostre que as equações normais para estimar o vector das constantes da regressão linear, $\hat{\beta} = (X^T X)^{-1} X^T Y$, neste caso dão $\hat{\beta} = c\hat{\Sigma}^{-1}(\overline{X}_1 - \overline{X}_2)$ e encontre c.

(b) Se $n_1 = n_2$ podemos classificar um novo x como sendo da classe 1 se $\hat{\beta}(x - \overline{X}) > 0$ e como sendo da classe 2 no caso contrário. Mostre que isto é equivalente à LDA dada na aula que assume normalidade e com $\pi_1 = \pi_2$.

45. Suponha que, num problema de Análise Discriminante Paramétrica, as funções densidade de probabilidade de cada classe são normalmente distribuidas, $p_i(x) = N_p(x; \mu_i, \Sigma_i)$, com matrizes de covariância diagonais. Encontre os EMV de μ_i e Σ_i, $i = 1, \ldots, K$, e a regra de classificação correspondente. Descreva um algoritmo para implementar essa regra.

46. * Suponha que, num problema de Análise Discriminante Paramétrica, as funções densidade de probabilidade de cada classe são normalmente distribuidas, $p_i(x) = N_p(x; \mu_i, \Sigma)$ $i = 1, \ldots, K$ (matrizes de covariância iguais).

(a) Mostre que a melhor regra linear de classificação (regra de Bayes) consiste em primeiro fazer uma certa transformação linear do espaço (branqueamento); de seguida, atribuir uma nova observação X à classe de cujo centro está mais próxima (a menos de uma constante).

(b) Mostre então que neste caso apenas precisamos de considerar um espaço de dimensão $K - 1$.

(c) Se $K > 2$, sugira outros subespaços de dimensão $L < K - 1$, que sejam óptimos para esta Análise Discriminante, no sentido em que os centros estejam o mais dispersos possivel (veremos mais tarde que isto é equivalente à regra linear de Fisher).

(d) Sugira um algoritmo para encontrar esses subespaços e variáveis discriminantes correspondentes (também conhecidas por variáveis canónicas).

34) EM R:

```
wine=read.table("wine.data",sep=",")
dim(wine)
pairs(wine[,2:14],col=wine[,1])

#Vamos escolher a 2 e a 13 e a 1 que contém a classe
wine1=wine[,c(2,13,1)]
plot(wine1[,1:2],col=wine1[,3],pch=20,cex=1.5,cex.lab=1.4)

#correr o LDA
wine.lda = lda(x=wine1[,1:2],grouping=wine1[,3])

#crear uma grelha para fazer as fronteiras da LDA x=seq(10,15,0.01)
y=seq(0.5,4.5,0.01)
z = as.matrix(expand.grid(x,y),0)
m =length(x)
n =length(y)

#como as classes são 1,2 e 3 vamos crias os contornos em 1.5 e 2.5
wine.ldp = predict(wine.lda,z)$class
contour(x,y,matrix(wine.ldp,m,n),levels=c(1.5,2.5), add=TRUE, d=FALSE, lty=2)

# Vamos fazer agora as fronteiras com QDA.
wine.qda = qda(x=wine1[,1:2],grouping=wine1[,3])
wine.qdp = predict(wine.qda,z)$class
plot(wine1[,1:2],col=wine1[,3],pch=20,cex=1.5,cex.lab=1.4)
contour(x,y,matrix(wine.qdp,m,n),levels=c(1.5,2.5), add=TRUE, d=FALSE, lty=2)

# CLARO QUE ISTO FOI APENAS USANDO 2 VAR. PREDICTIVAS.
```

35) EM R:

```
library(kernlab)
data(spam)

# Vamos escolher 0 para nonspam e 1 para spam
Y = as.numeric(spam[, ncol(spam)])-1 # (O R atribui 2 a spam e 1 a nonspam)
X = spam[ ,-ncol(spam)]
gl = glm(Y   ., data=X,family=binomial)
summary(gl)

proba = predict(gl,type="response")
predicted.spam = as.numeric( proba > 0.5)
table(predicted.spam,Y)
#
```

```
# Ele classificou 194 spams como non spam e, mais importante, 122 nonspam como
spam!!!
# agora para 0.99
predicted.spam = as.numeric( proba>0.99)
table(predicted.spam,Y)
# Agora claro erra muito mais e ainda assim atribui como spam 12 que eram non-
spam!


    n = length(Y)
s=sample(1:n)
q=round(0.70*n)
train=s[1:q]
test=s[(q+1):n]
gl = glm(Y[train]    ., data=X[train,],family=binomial)
proba.train = predict(gl,newdata=X[train,],type="response")
proba.test = predict(gl,newdata=X[test,],type="response")
predicted.spam.train = as.numeric(proba.train > 0.99)
predicted.spam.test = as.numeric(proba.test > 0.99)
table(predicted.spam.train, Y[train])
table(predicted.spam.test, Y[test])


    # LDA
library(MASS)
lda.res = lda(x=X[train,],grouping=Y[train])
proba.lda = predict(lda.res,newdata=X[test,])$posterior[,2]
predicted.spam.lda = as.numeric(proba.lda > 0.99)


    # Curvas ROC; em vez de 0.5 e 0.99 apenas, vamos usar muitos cut-off entre 0
e 1
cvec = seq(0.001,0.999,length=1000)
cvec = seq(0.001,0.999,length=1000)
specif= numeric(length(cvec))
sensit= numeric(length(cvec))
for (cc in 1:length(cvec))
+ sensit[cc]= sum( proba.lda > cvec[cc] & Y[test]==1)/sum(Y[test]==1)
+ specif[cc]= sum( proba.lda<=cvec[cc] & Y[test]==0)/sum(Y[test]==0)
+ }
plot(specif,sensit,main="LDA=linha preta e RL= linha vermelha",
+ xlab="SPECIFICITY",ylab="SENSITIVITY",type="l",lwd=2)
for (cc in 1:length(cvec))
+ sensit[cc]= sum( proba.test> cvec[cc] & Y[test]==1)/sum(Y[test]==1)
+ specif[cc]= sum( proba.test<=cvec[cc] & Y[test]==0)/sum(Y[test]==0)
+ }
lines(specif,sensit,col="red",type="l",lwd=2)


    # Agora a RL parece ser melhor para estes dados.
```