Université de Montpellier Année 2015-2016
M2 MIND - Discrimination et scoring

Examen partiel - 4 novembre 2015
Durée 1h30 - Documents interdits

Exercice 1 (7 pts) Soient X le vecteur aléatoire de dimension d contenant des variables explicatives quantitatives et Y le facteur binaire à valeurs dans $\{0,1\}$ à expliquer.

1) (1 pt) Donner la probabilité que $Y=1$ sachant $X=x$ en fonction de la loi marginale de Y et de la loi conditionnelle de X.

2) (3 pts) Pour la fonction de coût élémentaire

$$h(y,d)=\begin{cases} 0 & \text{si} \quad y=d \\ a & \text{si} \quad y=1 \quad \text{et} \quad d=0 \\ 10a & \text{si} \quad y=0 \quad \text{et} \quad d=1 \end{cases}$$

donner le prédicteur $g(X)$ minimisant le coût moyen.

3) (3 pts) Expliquer en détails le résultat produit par le code `R` donné ci-dessous.

```
pi0=0.3;pi1=0.7
mu0=c(3,9);mu1=c(5,5)
Sigma0=matrix(c(3,3*0.5,3*0.5,3),2,2)
Sigma1=matrix(c(5,-5*0.9,-5*0.9,5),2,2)
library(mvtnorm)
print(c(pi0*dmvnorm(c(4,4),mu0,Sigma0),pi1*dmvnorm(c(4,4),mu1,Sigma1)))
```

Exercice 2 (6 pts) Expliquer en détails le résultat produit par le code `R` donné ci-dessous.

```
n=200;pi0=0.3;pi1=0.7;mu0=c(3,9);mu1=c(5,5)
Sigma0=matrix(c(3,3*0.5,3*0.5,3),2,2)
Sigma1=matrix(c(5,-5*0.9,-5*0.9,5),2,2)
y=sample(c(0,1),size=n,prob=c(pi0,pi1),replace=TRUE);y=sort(y)

n0=sum(y==0);n1=sum(y==1);library(mvtnorm)
x=rbind(rmvnorm(n0,mu0,Sigma0),rmvnorm(n1,mu1,Sigma1))
plot(x,pch="*",cex=4,col=y+1)

donnees=data.frame(y=y,x1=x[,1],x2=x[,2])

nrep=100;npart=n/5
errorreg=rep(0,nrep)
for (l in 1:nrep)
{
indi=sample(1:n,n)
for (i in 1:5)
{
out=indi[(npart*(i-1)+1):(i*npart)]
model.lm=lm(y~x1+x2,data=donnees[-out,])
classif[out]=as.numeric(predict(model.lm,donnees[out,])>1/2)
}
errorreg[l]=mean(classif!=y)
}
```

Exercice 3 (7 pts) On se place dans le contexte de la classification supervisée binaire.

1) (2 pts) Expliquer en quoi consiste la méthode des k-plus-proches-voisins ?

2) (2 pts) Quel est le modèle de l'analyse discriminante linéaire ?

3) (4 pts) Nous supposons que nous observons 100 réalisations indépendantes $(x_1, y_1), \ldots, (x_n, y_n)$ du couple (X, Y) stockées dans les objets `R` : `x` et `y`, `y` est le vecteur binaire à expliquer.
Donner le code R mettant en oeuvre la technique de la validation croisée à 2 ensembles pour calculer l'erreur associée à l'analyse discriminante quadratique.

Correction examen partiel 04/11/2015 HMMA303

Exercice 1

1| $\mathbb{P}(Y=1|X=x) = \dfrac{f_1(x)\pi_1}{f_0(x)\pi_0 + f_1(x)\pi_1}$

avec $\mathbb{P}(Y=i)=\pi_i$, $\forall i=0,1$
et $f_i(x)$ la densité de la loi conditionnelle de X sachant $Y=i$, $\forall i=0,1$.

2| $C(g) = \mathbb{E}\left[h(Y,g(X))\right]$ (*)

$= \mathbb{E}\left[\mathbb{E}\left[h(Y,g(X))|X=x\right]\right]$

$$\mathbb{E}\left[h(Y, g(X)) \mid X = x\right]$$

$$= \alpha\, \mathbb{P}(Y=1 \mid X=x)\, \mathbb{1}\{g(x)=0\}$$

$$+ 1 0 \alpha\, \mathbb{P}(Y=0 \mid X=x)\, \mathbb{1}\{g(x)=1\} \quad (**)$$

Minimiser $(*)$ en $g(\cdot)$ est équivalent à minimiser $(**)$ en $g(\cdot)$ uniformément en x.

$$g^*(x) \in \arg\min_{g(\cdot)} \mathbb{E}\left[h(Y, g(X)) \mid X = x\right]$$

$$g^*(x) = \begin{cases} 0 & \text{si } \alpha\, \mathbb{P}(Y=1 \mid X=x) \leq 10\alpha\, \mathbb{P}(Y=0 \mid X=x) \\ 1 & \text{sinon} \end{cases}$$

(3)

$$g^*(x) = \begin{cases} 0 \text{ si } \mathbb{P}(Y=0|X=x) > \frac{1}{11} \\ 1 \text{ si } \mathbb{P}(Y=0|X=x) \leq \frac{1}{11} \end{cases}$$

On veut surtout pas se tromper lorsque $Y=0$.

3] Nous avons

$\pi_0 = \mathbb{P}(Y=0) = 0.3$

$\pi_1 = \mathbb{P}(Y=1) = 0.7$

$$X|Y=0 \sim N_2\left(\begin{pmatrix} 3 \\ 9 \end{pmatrix}, \begin{bmatrix} 3 & 3/2 \\ 3/2 & 3 \end{bmatrix}\right)$$

$$X|Y=1 \sim N_2\left(\begin{pmatrix} 5 \\ 5 \end{pmatrix}, \begin{bmatrix} 5 & -\frac{45}{10} \\ -\frac{45}{10} & 5 \end{bmatrix}\right)$$

Le code calcule le vecteur

$$(\pi_0 f_0((4,4)), \pi_1 f_1((4,4)))$$

Exercice 2

Simulation de $n = 200$ observées suivant le modèle présenté exercice 1 question 3.

Calcul de l'erreur associé au classifieur par régression linéaire par une méthode de validation croisée à 5 ensembles répété 100 fois.

Exercice 3

1] Voir cours

2] Voir cours

3] model.lda ← train(y~.,
data = data.frame(y = y, x = x),
method = "lda", metric = "Accuracy",
trControl = trainControl(method = "repeatedcv", number = 5,
repeats = 100))