Université de Montpellier Année 2016-2017
M2 MIND - Discrimination et scoring

Examen de contrôle continu - 17 octobre 2016
Durée 1h30 - Documents interdits

Exercice 1 (8 pts) On se place dans le contexte de la classification supervisée binaire.

1) (2 pts) Nous supposons que l'ensemble des prédicteurs contient une variable qualitative à trois modalités et une variable quantitative. Décrire comment fonctionne la méthode de l'analyse discriminante non paramétrique.

2) (2 pts) Nous supposons que l'ensemble des prédicteurs contient deux variables quantitatives. Quel modèle probabiliste est associé à la technique de l'analyse discriminante quadratique ?

3) (4 pts) Nous supposons que nous observons n 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 et `x` une variable quantitative. Donner le code `R` mettant en oeuvre les techniques d'analyses discriminantes linéaire, quadratique et non paramétrique et comparant leurs performances.

Exercice 2 (6 pts) Soient X une variable aléatoire de loi uniforme sur $[0, 1]$ et Y le facteur binaire à valeurs dans $\{0, 1\}$ à expliquer. Nous supposons que

$$\mathbb{P}(Y = 1 | X = x) = 1/3, \quad \forall x \in [0, 1]\,.$$

1) (1 pt) Calculer

$$\mathbb{P}[X \in [a, b] \cap \{Y = 1\}]$$

pour tout $0 < a < b < 1$.

2) (2 pts) Soit l'échantillon d'apprentissage $(x_1, y_1) = (0.4, 1)$, $(x_2, y_2) = (0.2, 0)$, $(x_3, y_3) = (0.7, 0)$. Donner les classifieurs des k-plus proches voisins pour $k = 1$ et $k = 3$.

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

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

calculer le coût moyen des deux classifieurs mis en évidence à la question précédente.

Exercice 3 (6 pts) Expliquer en détails le résultat produit par le code `R` donné ci-dessous.

```
data(iris)
n <- 100
x <- scale(iris[1:n,1:2])
y <- as.numeric(iris[1:n,5])-1
fisher <- data.frame(y=as.factor(y),x1=x[,1],x2=x[,2])
```

```
# y            x1                x2
# 0:50   Min.   :-1.8248   Min.   :-2.2956
# 1:50   1st Qu.:-0.7340   1st Qu.:-0.6246
#        Median :-0.1106   Median :-0.1024
#        Mean   : 0.0000   Mean   : 0.0000
#        3rd Qu.: 0.6685   3rd Qu.: 0.6287
#        Max.   : 2.3827   Max.   : 2.7176

len <- 500
x1 <- seq(min(x[,1]),max(x[,1]),length=len)
x2 <- seq(min(x[,2]),max(x[,2]),length=len)
grille <- expand.grid(x1=x1,x2=x2)

library(caret)

model.knn <- knn3(y~x1+x2,data=fisher,k=1)
mean(predict(model.knn,fisher,type="class")!=y)

# 0

classif.knn <- predict(model.knn,grille,type="class")

model.reg <- lm(as.numeric(fisher$y)-1 ~ .,data=fisher)
res.reg <- predict(model.reg,grille)
classif.reg <- rep(0,len*len)
classif.reg[res.reg > 0.5] <- 1
table(classif.reg,classif.knn)

#            classif.knn
# classif.reg      0      1
#           0 112599   7854
#           1   9015 120532

tune.knn <- train(y~x1+x2,data=fisher,method="knn",metric="Accuracy",
trControl=trainControl(method="repeatedcv",number=2,repeats=100),
tuneGrid=data.frame(k=seq(1,21,by=2)))
print(tune.knn)

k-Nearest Neighbors
100 samples
  2 predictor
  2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (2 fold, repeated 100 times)
Summary of sample sizes: 50, 50, 50, 50, 50, 50, ...
Resampling results across tuning parameters:

  k   Accuracy  Kappa
   1  0.9868    0.9736
   3  0.9904    0.9808
   5  0.9912    0.9824
   7  0.9917    0.9834
   9  0.9921    0.9842
  11  0.9926    0.9852
  13  0.9936    0.9872
  15  0.9932    0.9864
  17  0.9910    0.9820
  19  0.9893    0.9786
  21  0.9894    0.9788

Accuracy was used to select the optimal model using  the largest value.
The final value used for the model was k = 13.
```

Correction examen contrôle continu

17 octobre 2016

HPPA303

Exercice 1 Voir cours

Exercice 2

1) $\forall 0 < a < b < 1$

$$\mathbb{P}\left[X \in [a,b] \mid \{Y=1\}\right] = \frac{\mathbb{P}(Y=1 \mid X \in [a,b])}{\mathbb{P}(X \in [a,b])}$$

$$= \mathbb{P}(X \in [a,b]) \left(\frac{1}{3}\right)$$

$$= \left(\frac{b-a}{3}\right)$$

2] Soit $g_1(x)$ le classifieur
des 1-plus-proches-voisins

$$g_1(x) = \begin{cases} 0 & \text{si } x \in [0, 0.3] \\ 1 & \text{si } x \in]0.3, 0.55] \\ 0 & \text{si } x \in]0.55, 1] \end{cases}$$

Soit $g_3(x)$ le classifieur
des 3-plus-proches-voisins

$$g_3(x) = 0 \quad \forall x \in [0,1]$$

3] $\mathbb{E}\left[h(Y, g_1(X)) \right] = C_2$ est le
-cût moyen du -classifieur $g_2(\cdot)$

Nous remarquons que
$\mathbb{E}\left[h(Y, g_1(X)) \right] = \Pi\left[Y \neq g_1(X) \right]$
sont l'erreur de -classification.

Nous avons

$$C_1 = \mathbb{P}\left[\{g_2(X)=1\} \cap \{Y=0\}\right] + \mathbb{P}\left[\{g_1(X)=0\} \cap \{Y=1\}\right]$$

Comme $\mathbb{P}(Y=1|X=x) = \frac{1}{3}$, $\forall x \in [0,1]$ les variables aléatoires X et Y sont $\perp\!\!\!\perp$.

$$C_1 = \frac{2(0.55-0.3)}{3} + \frac{(0.3-0)+(1-0.55)}{3}$$

$$C_1 = \frac{0.5+0.75}{3} = \frac{1.25}{3}$$

Par ailleurs,

$$C = \mathbb{P}\left[\{g(X)=1\} \cap \{Y=0\}\right] + \mathbb{P}\left[\{g_3(X)=0\} \cap \{Y=1\}\right]$$

$$\Leftrightarrow C_3 = \mathbb{P}(Y=1) = \frac{1}{3}$$

Exercice 3

model.knn = _classifieur des 1-plus-proches voisins sur les données iris de Fisher

Si l'on évalue ses performances sur le même échantillon que celui ayant permis de le déterminer, on trouve une erreur de classification nulle. C'est le phénomène de sur-apprentissage.

model.reg = _classifieur par regression linéaire

On compare par matrice de confusion les deux classifications basées sur les modèles précédents des points d'une grille de 25 000 points.

time.knn = par validation croisée à 2 ensembles répétée 100 fois, on trouve que la meilleure méthode est celle du 13 plus-proches-voisins