Université de Montpellier Année 2017-2018
M2 MIND - Discrimination et scoring

Examen de contrôle continu - 4 octobre 2017
Durée 1h30 - Documents interdits

Exercice 1 (6 pts) On se place dans le contexte de la classification supervisée binaire.

1) (2 pts) Décrire le fonctionnement de la méthode de l'estimation de densité par la méthode du noyau uniforme.

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) (2 pts) Nous supposons que nous observons n réalisations indépendantes $(x_1, y_1), \ldots, (x_n, y_n)$ du couple (X, Y) stockées dans le `data.frame` nommé `cc` contenant `y` le vecteur binaire à expliquer, et `x` la matrice des variables explicatives quantitatives. Donner le code `R` mettant en oeuvre les techniques de la méthode des 3-plus proches voisins et de l'analyse discriminante linéaire tout en comparant leurs performances.

Exercice 2 (8 pts) Soient Y le facteur binaire à expliquer et X une variable explicative réelle. Nous supposons que

$$\mathbb{P}(Y = 0) = \mathbb{P}(Y = 1) = 1/2\,,$$

$$X|Y = 0 \sim \mathcal{N}(-2, 1)\,,$$
$$X|Y = 1 \sim \mathcal{N}(1, 1)\,.$$

1) (2 pts) Calculer $\mathbb{P}(Y = 0|X = x)$ et $\mathbb{P}(Y = 1|X = x)$.

2) (2 pts) Donner le classifieur optimal $g^{\#}(\cdot)$ pour la fonction de coût élémentaire $h(y, d) = \mathbb{I}_{y \neq d}$.

3) (2 pts) Soit l'échantillon d'apprentissage $(x_1, y_1) = (-1, 0)$, $(x_2, y_2) = (1, 1)$, $(x_3, y_3) = (2, 1)$. Donner le classifieur des 1-plus proches voisins noté $g^1(\cdot)$.

4) (2 pts) Calculer les coûts moyen de $g^{\#}$ et g^1 et les comparer.

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

```
load(file="MNIST.train.Rdata")
dim(mnist.train)
# [1] 60000   785
subsamp.train <- sample(1:60000,5000)
mnist.train <- mnist.train[subsamp.train,]
# [1] 5000   785

load(file="MNIST.test.Rdata")
dim(mnist.test)
# [1] 10000   785
subsamp.test <- sample(1:10000,1000)
mnist.test <- mnist.train[subsamp.test,]
dim(mnist.test)
# [1] 1000   785

library(caret)
model1 <- knn3(y~.,data=mnist.train,k=20)
res.model1 <- predict(model1,mnist.test)

class.model1 <- rep(0,1000)
for (k in 1:1000) class.model1[k] <- order(res.model1[k,])[10]
mean(as.factor(class.model1-1)==mnist.test[,785])
# [1] 0.963
```

Pour vous aider à comprendre :

```
# sample(1:5,3)
# [1] 4 5 2

# head(res.model1)
#      0   1 2 3 4 5 6 7    8    9
# [1,] 0 0.8 0 0 0 0 0 0 0.15 0.05
# [2,] 0 0.0 0 0 0 0 1 0 0.00 0.00
# [3,] 0 0.0 0 0 0 0 0 1 0.00 0.00
# [4,] 0 1.0 0 0 0 0 0 0 0.00 0.00
# [5,] 0 1.0 0 0 0 0 0 0 0.00 0.00
# [6,] 1 0.0 0 0 0 0 0 0 0.00 0.00

# order(c(5,7,8,2))
# [1] 4 1 2 3
#  order(c(5,7,8,2))[4]
# [1] 3
```

Correction examen de contrôle continu
4/10/2017
HMMA303

Exercice 1

1] Voir cours

2] Voir cours

3] library (caret)

model.knn ← train(y ~ .,
data = cc, method = "knn", k=3,
metric = "Accuracy", trControl =
trainControl(method = "repeatedcv",
number = 10, repeats = 100))

Cette commande met en œuvre le modèle knn avec $k = 3$ et évalue son erreur.

model.lda <- train(y ~ .,
data = cc, method = "lda", metric = "Accuracy",
trControl = trainControl(method = "repeatedcv",
number = 10, repeats = 100))

Cette commande met en œuvre le modèle lda et évalue son erreur par validation croisée à 10 ensembles répétée 100 fois.

Exercice 2

1] $\mathbb{P}(Y=0|X=x) = \frac{f_0(x)\pi_0}{f_0(x)\pi_0 + f_1(x)\pi_1}$

$$\mathbb{P}(Y=0|X=x) = \frac{e^{-\frac{1}{2}(x+2)^2}}{e^{-\frac{1}{2}(x+2)^2} + e^{-\frac{1}{2}(x-1)^2}}$$

$$\mathbb{P}(Y=1|X=x) = \frac{e^{-\frac{1}{2}(x-1)^2}}{e^{-\frac{1}{2}(x+2)^2} + e^{-\frac{1}{2}(x-1)^2}}$$

2] La fonction de coût est la fonction de coût symétrique. Dans ce cas le classifieur optimal est le classifieur de Bayes, il consiste à effectuer la prédiction à la classe de probabilité la plus forte.

$$g^{\#}(x) = \begin{cases} 1 \text{ si } \mathbb{P}(Y=1|X=x) \geq \mathbb{P}(Y=0|X=x) \\ 0 \text{ sinon} \end{cases}$$

$$\Leftrightarrow g^{\#}(x) = 1 \text{ si } e^{-\frac{1}{2}(x-1)^2} \geq e^{-\frac{1}{2}(x+2)^2}$$

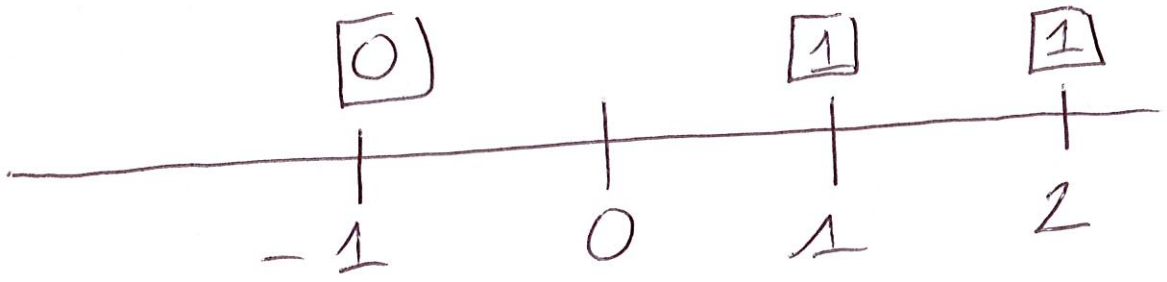
$$\Rightarrow g^{\#}(x) = 1 \text{ si } (x-1)^2 \leq (x+2)^2$$

$$\Rightarrow g^{\#}(x) = 1 \text{ si } -2x+1 \leq 4x+4$$

$$\Rightarrow g^{\#}(x) = 1 \text{ si } x \geq -\frac{1}{2}$$

$$g^{\#}(x) = \mathbb{1}\{x \geq -\tfrac{1}{2}\}$$

3]



on en déduit que

$$g^{\wedge}(x) = \mathbb{1}\{x \geq 0\}$$

4] $\mathbb{E}[\ell(Y, g^{\#}(X))]$

$= \mathbb{P}(Y \neq g^{\#}(X))$

$= \mathbb{P}(\{Y=0\} \cap \{g^{\#}(X)=1\})$

$+ \mathbb{P}(\{Y=1\} \cap \{g^{\#}(X)=0\})$

$= \mathbb{P}(\{Y=0\} \cap \{X \geq -\tfrac{1}{2}\})$

$+ \mathbb{P}(\{Y=1\} \cap \{X < -\tfrac{1}{2}\})$

$$\Leftrightarrow \mathbb{E}\left[h(Y, g^{\#}(X))\right] = C(g^{\#})$$

$$= \mathbb{P}(X \geq -\tfrac{1}{2} \mid Y=0)\,\mathbb{P}(Y=0)$$
$$+ \mathbb{P}(X < -\tfrac{1}{2} \mid Y=1)\,\mathbb{P}(Y=1)$$

$$= \frac{1}{2} - \frac{1}{2}\,\mathbb{P}(\mathcal{N}(0,1) < \tfrac{3}{2})$$
$$+ \frac{1}{2}\,\mathbb{P}(\mathcal{N}(0,1) < -\tfrac{3}{2})$$

$$= \frac{1}{2} - \frac{1}{2} F_{\mathcal{N}(0,1)}\left(\tfrac{3}{2}\right) + \frac{1}{2} F_{\mathcal{N}(0,1)}\left(-\tfrac{3}{2}\right)$$

$$= 1 - F_{\mathcal{N}(0,1)}(1{,}5)$$

$$C(g^{1}) = \mathbb{E}\left[h(Y, g^{1}(X))\right]$$

$$= \mathbb{P}(\{Y=0\} \cap \{X \geq 0\}) + \mathbb{P}(\{Y=1\} \cap \{X < 0\})$$
$$= \mathbb{P}(X \geq 0 \mid Y=0)\,\mathbb{P}(Y=0) + \mathbb{P}(X < 0 \mid Y=1)\,\mathbb{P}(Y=1)$$
$$= \frac{1}{2}\,\mathbb{P}(\mathcal{N}(0,1) > 2) + \frac{1}{2}\,\mathbb{P}(\mathcal{N}(0,1) < -1)$$
$$= \frac{1}{2} - \frac{1}{2} F_{\mathcal{N}(0,1)}(2) + \frac{1}{2} F_{\mathcal{N}(0,1)}(-1)$$

$$C(g^1) = 1 - \frac{1}{2} F_{N(0,1)}(2) - \frac{1}{2} F_{N(0,1)}(1)$$

Sur $[0, \infty[$, la fonction de répartition de loi normale centrée réduite est strictement concave donc

$$\frac{1}{2} F_{N(0,1)}(2) + \frac{1}{2} F_{N(0,1)}(1) < F_{N(0,1)}\left(\frac{3}{2}\right)$$

On retrouve bien que

$$C(g^\#) < C(g^1)$$

Exercice 3

Sur la base de données MNIST, dont les échantillons d'apprentissage et de test sont ramenés à des tailles de 5000 et 1000 individus, mise en œuvre de la méthode des 20-plus-proches-voisins et calcul d'erreur sur le test.