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

Examen final - 8 novembre 2017
Durée 2h00 - Documents interdits

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

1) (2 pts) Décrire le fonctionnement de l'algorithme CART ?

2) (1 pt) Qu'est-ce que la technique du bagging ?

3) (1 pt) Expliquer comment l'on peut s'affranchir de l'utilisation de la méthode de la validation croisée pour estimer le taux d'erreur associé à un classifieur forêt aléatoire ?

4) (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 classification par arbre et des forêts aléatoires tout en comparant leurs performances.

Exercice 2 (8 pts) Soient Y un facteur binaire à expliquer et X une variable explicative qualitative à trois modalités $\{a, b, c\}$. Nous considérons la fonction de coût élémentaire suivante $h(y, d) = \mathbb{I}_{y \neq d}$. Nous notons $\pi_0 = \mathbb{P}(Y = 0)$, $\pi_1 = \mathbb{P}(Y = 1)$, $p_{0,x} = \mathbb{P}(X = x | Y = 0)$ et $p_{1,x} = \mathbb{P}(X = x | Y = 1)$ $(x \in \{a, b, c\})$.

1) (3 pts) Soit U une variable aléatoire uniforme sur $[0, 1]$ indépendante de X et Y. Nous considèrons les classifieurs aléatoires suivants $g_1(x) = \mathbb{I}_{U \leq 1/2}$ et $g_2(x) = \mathbb{I}_{U \leq \pi_1}$. Comparer les coûts de g_1 et g_2. Commenter.

2) (2 pts) Donner l'expression du classifieur optimal $g^{\#}$ en fonction de π_0, $p_{0,x}$ et $p_{1,x}$.

3) (1 pt) Nous supposons que $\pi_0 = 1/2$, $p_{0,a} = p_{0,c} = 1/2$ et $p_{1,a} = p_{1,b} = p_{1,c} = 1/3$. Comparer les coûts de $g^{\#}$, g_1 et g_2. Commenter.

4) (2 pts) Nous observons un échantillon de taille $n = 10$ avec $(x_1, y_1) = (a, 0)$, $(x_2, y_2) = (a, 0)$, $(x_3, y_3) = (c, 0)$, $(x_4, y_4) = (b, 1)$, $(x_5, y_5) = (c, 1)$, $(x_6, y_6) = (c, 0)$, $(x_7, y_7) = (c, 0)$, $(x_8, y_8) = (c, 1)$, $(x_1, y_1) = (a, 1)$ et $(x_{10}, y_{10}) = (b, 1)$.
Proposer un classifieur.

Exercice 3 (6 pts)

1) (4 pts) Expliquer en détails le résultat produit par le code donné ci-dessous.

```
>>> from sklearn import datasets
>>>
>>> digits = datasets.load_digits()
>>> X=digits.data
>>> y=digits.target
>>>
>>> from sklearn.model_selection import train_test_split
>>> X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.25,random_state=14)
>>>
>>> from sklearn.neural_network import MLPClassifier
>>> mlp = MLPClassifier(hidden_layer_sizes = (2000,2000,), activation = 'logistic' )
>>> digits_mlp = mlp.fit(X_train, y_train)
>>> print(1-digits_mlp.score(X_test,y_test))
0.0133333333333
>>>
>>> from sklearn.neighbors import KNeighborsClassifier
>>> knn = KNeighborsClassifier(n_neighbors=10)
>>> digit_knn=knn.fit(X_train, y_train)
>>> print(1-digit_knn.score(X_test,y_test))
0.0244444444444
>>>
>>> from sklearn.model_selection import GridSearchCV
>>> param=[{"n_neighbors":list(range(1,15))}]
>>> knn=GridSearchCV(KNeighborsClassifier(),param,cv=5,n_jobs=-1)
>>> digit_knn=knn.fit(X_train, y_train)
>>> digit_knn.best_params_["n_neighbors"]
1
>>>
>>> knn=KNeighborsClassifier(n_neighbors=digit_knn.best_params_["n_neighbors"])
>>> digit_knn=knn.fit(X_train, y_train)
>>> print(1-digit_knn.score(X_test,y_test))
0.0133333333333
```

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

```
> library(MASS)
> data(Pima.tr)
> data(Pima.te)
> library(e1071)
> calibration <- tune(svm,type~.,data=Pima.tr,
+ ranges=list(gamma=seq(0.001,0.2,length=5),cost=seq(0.1,2,length=5)))
> model.svm <- svm(type~.,data=Pima.tr,gamma=calibration$best.parameters[1],
+ cost=calibration$best.parameters[2])
> mean(predict(model.svm,Pima.te)==Pima.te$type)
[1] 0.7951807
```

Correction examen
HPPA 303
08/11/2017

Exercice 1

1) Arbres binaires de régression ou de classification

2) Bagging : Bootstrap Aggregating

3) Out-of-bag

4) library(caret) ; library(rpart)
library(randomForest)

```
model.cart <- train(y ~ ., data=cc,
                    method="rpart", ----)

model.rf <- train(y ~ ., data=cc,
                  method="rf", ----)
```

Exercice 2

1] $Y \in \{0,1\}$ et $X \in \{a,b,c\}$

$h(y,a) = \mathbb{1}\{y \neq a\}$

$C(g_1) = \mathbb{E}\left[h(Y, g_1(X)\right] = \mathbb{P}\left(Y \neq g_1(X)\right)$

$= \mathbb{P}(Y=0 \cap U \leq 1/2) + \mathbb{P}(Y=1 \cap U \geq 1/2)$

$= \frac{\pi_0}{2} + \frac{\pi_1}{2} = \frac{1}{2}$

$C(g_2) = \mathbb{P}(Y \neq g_2(X)) = \pi_0 \pi_1 + \pi_1 \pi_0$

$= 2\pi_0\pi_1 = 2(\pi_0 - \pi_0^2) \leq \frac{2}{4}$

$\Rightarrow C(g_1) \geq C(g_2)$

Le classifieur aléatoire basé sur les proportions théoriques est meilleur.

2] Fonction de coût élémentaire symétrique $\Rightarrow$ le classifieur optimal est le classifieur de Bayes.

$$\mathbb{P}(Y=1|X=x) \propto \pi_1 p_{1,x}$$

$$\mathbb{P}(Y=0|X=x) \propto \pi_0 p_{0,x}$$

$$\Rightarrow \mathbb{P}(Y=0|X=x) = \left[\frac{\pi_0 p_{0,x}}{\pi_0 p_{0,x} + \pi_1 p_{1,x}}\right]$$

$$g^{\#}(x) = \mathbb{1}\{p_{1,x} \geq \pi_0 (p_{1,x} + p_{0,x})\}$$

3] $\pi_0 = \pi_1 = \frac{1}{2}$, $g_1 = g_2$, $C(g_1) = C(g_2)$

$= \frac{1}{2}$

Sinon, ailleurs,

$$g^{\#}(x) = \begin{cases} 0 & \text{si } x = a \text{ ou } x = c \\ 1 & \text{si } x = b \end{cases}$$

$$C(g^{\#}) = \mathbb{P}(Y=1 \cap X=a) + \mathbb{P}(Y=0|X=b) + \mathbb{P}(Y=1 \cap X=c)$$

$$= \frac{1}{2} \times \frac{1}{3} + \frac{1}{2} \times \frac{1}{3} = \frac{1}{3} < \frac{1}{2}$$

4] $\widehat{\pi_0} = \widehat{\pi_1} = \frac{1}{2}$

$\widehat{p_{0,a}} = \frac{2}{5}$, $\widehat{p_{0,c}} = \frac{3}{5}$

$\widehat{p_{1,a}} = \frac{1}{5}$, $\widehat{p_{1,b}} = \frac{2}{5}$, $\widehat{p_{1,c}} = \frac{2}{5}$

$$\hat{g}_{10}(x) = \begin{cases} 0 & \text{si } x = a \text{ ou } x = c \\ 1 & \text{si } x = b \end{cases}$$

Exercice 3

1] Comparaison des techniques :

- perceptron multi-couches : 2 couches cachées avec chacune 2000 neurones ;
- 10 plus-proches-voisins ;
- 1 plus-proche-voisin ;

par utilisation d'un ensemble de test.

2] Mise en oeuvre du classifieur SVM sur les données pima.