Université de Carthage
Ecole Supérieure de la Statistique et de l'Analyse de l'Information

Examen de Data Mining

3 ème année du cycle de formation d'ingénieurs

Durée de l'épreuve : 1 heure 30 - Documents non autorisés
Nombre de pages : 3 - Date de l'épreuve : 9 janvier 2020

Exercice 1 : On considère le tableau de données ci-dessous contenant les valeurs observées de deux variables quantitatives X^1 et X^2, et d'une variable qualitative Y possédant les deux modalités notées A et B, sur un échantillon I de huit individus notés $P_1, \ldots, P_8$. Chaque individu est muni du poids 1/8.

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
x^1	4	3	1	0	4	3	5	4
x^2	5	4	2	1	4	3	3	2
y	A	A	A	A	B	B	B	B

Par la suite, on applique différentes méthodes de classification supervisée à ces données afin d'expliquer Y en fonction de X^1 et X^2. Pour cela, on utilise les commandes du logiciel R. On note g, g_A, g_B les centres de gravité respectifs du nuage I et des classes $I_A = \{P_1, P_2, P_3, P_4\}$ et $I_B = \{P_5, P_6, P_7, P_8\}$.

A- Analyse factorielle discriminante

1- Calculer les centres de gravité g, g_A, g_B .

On effectue l'AFD linéaire du tableau de données.

2- Expliquer pourquoi il n'existe qu'un seul axe factoriel discriminant (non trivial).

3 - Quelle est la commande de R qui permet d'appliquer une AFD linéaire aux données. On notera "don" le data.frame dans lequel sont enregistrées les données. On précisera les arguments nécessaire pour cette fonction.
4- Sachant que le facteur discriminant a pour coordonnées :

```
X1 1.279204
X2 -1.066004
```

Indiquer les scores des 2 centres de gravité.

5- Les probabilités *a posteriori* et les scores des individus donnéees par l'AFD linéaire du tableau de données sont donnés ci-dessous :

```
$posterior
                A           B
P1 0.898604854 0.10139515
P2 0.938616893 0.06138311
P3 0.978504501 0.02149550
P4 0.987428061 0.01257194
P5 0.366919631 0.63308037
P6 0.500000000 0.50000000
P7 0.001434570 0.99856543
P8 0.002472623 0.99752738

$x
          LD1
P1 -0.8528029
P2 -1.0660036
P3 -1.4924050
P4 -1.7056057
P5  0.2132007
P6  0.0000000
P7  2.5584086
P8  2.3452079
```

Déterminer de deux manières différentes la classe d'affectation de chacun des individus.

B- Arbre de décision

On considère les résultats de la classification supervisée réalisée à l'aide de l'arbre de décision. L'arbre obtenu est présenté ci-dessous :

```
> arbre.full <- rpart(Y ~ ., data = donn, minsplit =3, method = "class")
> print(arbre.full)
n= 8

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 8 4 A (0.5000000 0.5000000)
  2) X1< 2 2 0 A (1.0000000 0.0000000) *
  3) X1>=2 6 2 B (0.3333333 0.6666667)
    6) X2>=3.5 3 1 A (0.6666667 0.3333333) *
    7) X2< 3.5 3 0 B (0.0000000 1.0000000) *
```

6- Commenter la ligne de commande qui a permis d'obtenir cet arbre.

7- Déterminer la classe prédite de chacun des huit individus et en déduire la matrice de confusion.

8- On considère la courbe ROC évaluant le modèle de l'arbre de décision obtenu pour prédire qu'un objet appartient à la classe B. Donner les coordonnées de trois points qui sont situés sur cette courbe.

Exercice 2 : Le traitement du cancer de la prostate change si le cancer a atteint ou non les noeuds lympatiques entourant la prostate. Pour éviter une investigation lourde un certain nombre de variables sont considérées comme explicatives de la variable Y : $Y = 0$ si le cancer n'a pas atteint le réseau lympatique et $Y = 1$ sinon. Le but de cette étude est donc d'expliquer et de prédire Y par les variables suivantes :

- `age` : âge du patient au moment du diagnostic ;
- `acide` : le niveau d'acide phosphate sérique ;
- `rayonx` : le résultat d'une analyse par rayon X, 0= négatif et 1= positif ;
- `taille` : la taille de la tumeur, 0= petite et 1= grande ;
- `grade` : l'état de la tumeur déterminé par biopsie, 0= moyen et 1= grave ;
- `log.acid` : le logarithme népérien du niveau d'acidité ;

On dispose d'une base de données constituée de 53 individus. Chacun des 53 individus est décrit par les 6 variables prédictives présentées ci-dessus ainsi que par sa valeur sur la variable Y. Par la suite, on applique la méthode SVM (Support Vector Machine) de classification supervisée à ces données afin d'expliquer Y par l'ensemble des variables explicatives que l'on note X. Pour cela, on utilise les commandes du logiciel *Python*.

On a appliqué le scripte suivant :

```
X_train,X_test,y_train,y_test=train_test_split(X,Y,test_size=0.25,random_state=0)

from sklearn import svm

svc = svm.SVC(C=1.0, kernel='rbf',
              degree= 3, gamma=1.0,
              coef0=0.0, shrinking=True,
              probability=True,tol=0.001,
              cache_size=200, class_weight=None,
              verbose=False,max_iter= -1,
              random_state=None)


from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
scaler.fit(X,Y)
x_train_scaled = scaler.transform(X_train)
svc.fit(x_train_scaled,Y_train)
```

1- Commenter les lignes de commandes précédentes.

2- Afin de déterminer le meilleur modèle, en terme d'erreur de prédiction, obtenu à partit de la méthode SVM, sur quels paramètres de la fonction `svm.SVC` devrions nous agir ? Quelle fonction *Python* permetterait d'obtenir les meilleures valeurs de ces paramètres. On expliquera le principe de cette fonction.