EISTI : Cycle Ingénieurs : Deuxième Année - Parcours MI

Examen de fouille de données

29 janvier 2018

Durée 3H - Documents de cours et de TDs autorisés

Modalités : Vous devez rendre un document électronique contenant les réponses aux questions ainsi que le programme R associé à la dernière partie.

NOM Prénom :

Notations

Exercice 1.1 :
Exercice 1.2 :
Exercice 1.3 :
Exercice 1.4 :
Exercice 1.5 :
Exercice 1.6 :
Exercice 1.7 :
Exercice 1.8 :
Exercice 1.9 :

Exercice 2.1 :
Exercice 2.2 :
Exercice 2.3 :
Exercice 2.4 :
Exercice 2.5 :
Exercice 2.6 :

Question 3.1 :
Question 3.2 :

Question 3.3 :

Question 3.4 :

Question 3.5 :

Question 3.6 :

Question 3.7 :

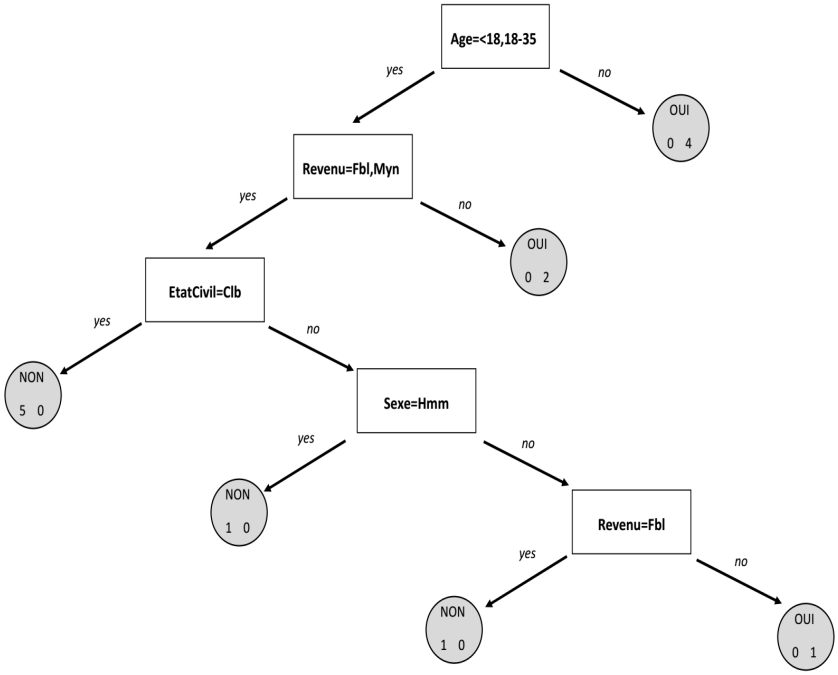
Notes globales :

Arbre de décision et Naive Bayes

On considère le jeu de données suivant sur le comportement d'achat de clients. On souhaite construire un arbre permettant de prévoir s'il y aura achat ou non en suivant de renseignement personnel sur le client.

	Sexe	Age	EtatCivil	Revenu	Achat
1	Homme	18-35	Marié	Moyen	Non
2	Homme	<18	Célibataire	Faible	Non
3	Femme	<18	Célibataire	Moyen	Non
4	Homme	18-35	Célibataire	Moyen	Non
5	Femme	18-35	Marié	Faible	Non
6	Femme	<18	Célibataire	Moyen	Non
7	Homme	18-35	Célibataire	Faible	Non
8	Homme >	35	Marié	Elevé	Oui
9	Femme	18-35	Célibataire	Elevé	Oui
10	Homme	18-35	Marié	Elevé	Oui
11	Homme	>35	Célibataire	Faible	Oui
12	Femme	>35	Célibataire	Moyen	Oui
13	Femme	>35	Marié	Elevé	Oui
14	Femme	18-35	Marié	Moyen	Oui

L'arbre complet est le suivant :



Exercice 1.1 Quel est le gain du 1er split si on considère l'indice de Gini ?

Exercice 1.2 Quelle est la matrice de confusion pour la base test suivante ?

	Sexe	Age	EtatCivil	Revenu	Achat
15	Homme	18-35	Marié	Moyen	Non
16	Homme	<18	Célibataire	Faible	Non
17	Femme	>35	Marié	Elevé	Oui
18	Femme	18-35	Marié	Moyen	Oui

Exercice 1.3 Élaguez l'arbre après le 2ème split.

Exercice 1.4 Énoncez les résultats de l'arbre élagué sous la forme de règles *SI ALORS* .

Exercice 1.5 Combien a-t-on d'individus mal classés dans la base d'apprentissage avec ce nouvel arbre ?

On utilise maintenant le regroupement de modalités (classes) effectué dans l'arbre de décision. On a donc les variables (attributs) suivantes :

```
X1 : Sexe : H=1 et F=0
X2 : Age :  "  35 "=1 et " >35 "=0
X3 : EtatCivil :  Marié=1 et Célibataire=0
X4 : Revenu : " Faible-Moyen "=1 et Elevé=0
Y : Achat : Oui=1 et Non=0
```

Nous allons appliquer la méthode Naive Bayes.

Exercice 1.6 Quelles sont les deux probabilités que l'on doit calculer pour prédire la classe d'un nouvel individu dont on connait les valeurs des Xi ?

Exercice 1.7 Appliquez la formule de Bayes pour transformer ces probabilités.

Exercice 1.8 Quelle hypothèse doit-on supposer pour utiliser la méthode naive Bayes ? Que deviennent alors les deux probabilités ?

Exercice 1.9 Quelle est la classe prédite pour l'individu 15 de la base test (justifiez vos calculs) ?

Evaluation des résultats de classifieur

Nous considérons une méthode d'apprentissage supervisé appliquée à un ensemble d'apprentissage $D = \{(xi, ci)\}$. La classe ci prend deux valeurs : $Vrai$ et $Faux$.
Pour étudier l'efficacité de notre méthode, nous utilisons la matrice de confusion qui se présente comme suit :

	V	F
V	TP	FP
F	FN	TN

où les colonnes représentent les valeurs réelles de la classe et les lignes ses valeurs prédites par la méthode. Par exemple, TP représente le nombre d'exemples positifs classés $Vrai$ par la méthode.

Plusieurs critères de mesure de la performance de notre méthode peuvent être utilisées. Parmi lesquelles la **sensibilité** (Se) et la **spécificité** (Sp), respectivement définies comme suit :

— $Se = \frac{TP}{TP+FN}$
— $Sp = \frac{TN}{FP+TN}$

Exercice 2.1 À quel intervalle appartient chacune de ces grandeurs ? et quels sens leur donnez-vous ?

Exercice 2.2 En vous appuyant sur un exemple simple, expliquez pourquoi il ne suffit pas d'avoir une "bonne" valeur de la sensibilité (ou de la spécificité) pour conclure à l'efficacité de la méthode ?

Exercice 2.3 Expliquez pourquoi pour une bonne méthode, la sensibilité et la spécificité sont forcément strictement supérieures à 0.5.

On définit la **précision** d'un algorithme comme étant la probabilité que la valeur réelle soit égale à $Vrai$ lorsque la valeur prédite est égale à $Vrai$.

Exercice 2.4 Exprimez la précision en fonction de p (la probabilité que la valeur réelle soit égale à $Vrai$), Se et Sp puis en fonction des valeurs TP, TN, FP et FN.

Exercice 2.5 Donnez la sensibilité, la spécificité et la précision dans le cas où la matrice de confusion est la suivante :

	V	F
V	60	40
F	10	470

Exercice 2.6 Donnez la précision lorsque la sensibilité est égale à 0.99, la spécificité égale à 0.9 et la probabilité d'avoir la valeur vraie est égale à 0.01. Refaites la calcul en donnant à la spécificité les valeurs suivantes : 0.99 puis 0.999. Conclusion ?

Etude de cas avec R

Cette dernière partie nécessite la rédaction d'un programme R ainsi que la rédaction des réponses aux questions associées.

Soit l'ensemble de données joint nommé "wineData.csv" décrivant des données correspondant au résultat d'une analyse chimique de vins cultivés dans la même région en Italie mais issus de trois cultivars différents. L'analyse a déterminé les quantités de 13 constituants trouvés dans chacun des trois types de vins :

1. Alcohol
2. Malic acid
3. Ash
4. Alcalinity of ash
5. Magnesium
6. Total phenols
7. Flavanoids
8. Nonflavanoid phenols
9. Proanthocyanins
10. Color intensity
11. Hue
12. OD280/OD315 of diluted wines
13. Proline

Question 3.1 Lire ce fichier avec la commande

```
wine <-read.csv(file="wineData.csv",sep=";")
```

Pour que cette commande marche, il faut que le fichier contenant les données soit dans le même répertoire de travail de R. Pour connaître le répertoire de travail de R, exécuter la commande

```
getwd()
```

Vous pouvez ensuite copier le fichier des données dans le répertoire de travail. une deuxième solution consiste à spécifier le répertoire de travail PATH [1] dans la commande ainsi :

```
flags <-read.csv(file="PATH/wineData.csv.csv",sep=";")
```

1. Ici, PATH est une variable contenant le répertoire de travail trouvé par la commande getwd()

Effectuer ensuite les transformations de types suivantes :

```
wine$Class<-as.factor(wine$Class)
```

Quels sont les types des attributs présents dans ce jeu de données ? Choisir une variable de chaque type et représenter graphiquement la distribution de ses valeurs.

Question 3.2 Notre but est d'apprendre les valeurs de la variable **Class** en fonction des autres attributs (ou d'un sous ensemble). Diviser l'ensemble de données en un jeu d'apprentissage et un jeu de test en prenant 70% pour l'apprentissage et 30% pour le test. Vérifier que toutes les valeurs de la classe sont représentées dans les deux jeux.

Question 3.3 A l'aide du package rpart, trouver l'arbre de décision appris à partir de l'ensemble d'apprentissage avec le paramètre de contrôle $minsplit = 10$. Représenter graphiquement cet arbre.

quels sont les attributs les plus discriminants pour la détermination de la classe ? Écrire l'ensemble des règles qui prédisent la sortie.

Question 3.4 Trouver la matrice de confusion sur les données de l'ensemble de test. Quelle classe a la meilleure précision ? Quelle est celle qui a le meilleur rappel ?

Question 3.5 Appliquer l'algorithme de Random Forest sur le jeux d'apprentissage, donner la précision du classifieur pour chaque valeur de classe en se basant sur l'ensemble d'apprentissage. A L'aide des deux fonctions

```
importance(rf)  varImpPlot(rf)
```

que vous appliquez au modèle résultant, trouver les sept meilleurs attributs selon le gain de Gini.

Question 3.6 Nous souhaitons appliquer l'algorithme k-means sur la totalité des données ; choisir parmi les sept attributs trouvés dans la question précédente, les attributs opportuns. Appliquer l'algorithme k-means sans prendre en compte la variable classe, avec K=3, K=4,k=5. Relevez l'inertie totale de l'ensemble de données. Relevez l'inertie inter et l'inertie intra pour chacune des trois valeurs du paramètre k. Pour construire un dataFrame en prenant en compte certains attributs d'un autre data frame, vous pouvez utiliser :

```
newDF<-data.frame(oldDF$att1, .., oldDF$attk, ..)
```

Question 3.7 Trouver la matrice de correspondance classes-clusters pour k=4. Est-ce qu'il y a des clusters purs ? Est-ce qu'il a des vins de classes différentes qui se ressemblent ? Commentez cette matrice.