

Fouille de Données T.P. N° 2

Etude du classifieur bayésien

20 octobre 2021

1 Application du théorème de Bayes

Exercice 1 Soit P la population des champignons. Il y a deux classes de champignons : vénéneux et comestible. on s'intéresse à l'attribut binaire volve. Soit le tableau suivant :

classe k	1 : vénéneux	2 : comestible
$P(k)$	0.05	0.95
$P(\text{volve} k)$	0.9	0.2

On ramasse les champignons si la règle de Bayes détermine leur classe comme comestible. Est ce qu'on ramasse les champignons ayant une volve.

Exercice 2 Les anglais et les américains orthographient le mot rigueur respectivement *rigour* et *rigor*.

Un homme ayant pris une chambre dans un hotel a écrit ce mot sur un bout de papier. Une lettre est prise au hasard dans ce mot : c'est un voyelle. Quelle la probabilité que cet homme soit anglais ?
Répéter la même question sachant que l'hotel héberge 60% d'américains, les autres étant des anglais.

2 Classifieur bayésien naif

Soit l'ensemble d'apprentissage suivant

Match à domicile ?	Balance positive ?	Mauvaises conditions climatiques ?	Match précédent gagné ?	Match gagné
V	V	F	F	V
F	F	V	V	V
V	V	V	F	V
V	V	F	V	V
F	V	V	V	F
F	F	V	F	F
V	F	F	V	F
V	F	V	F	F

Exercice 3 Quel est votre pronostic pour le match de samedi sachant qu'il est à domicile, que la balance est positive, les conditions climatiques seront bonnes et le match précédent a été gagné ?

3 Etude pratique en utilisant Weka

Exercice 4 En utilisant le classeur Bayésien , ainsi que la base d'exemples Weather. Tester le classeur sur la base de test contenant les trois exemples positifs suivants :

- la journée est ensoleillée et le vent est faible.
- la journée est ensoleillée, le vent est faible, la température est de 23 degrés et l'humidité est de 70 %.
- la température est de 23 degrés.

Exercice 5 Effectuer une étude comparative, en mode validation croisée, des algorithmes k-plus proche voisin, arbres de décision(ID3 ou J48) et le classeur bayésien sur la base de données IRIS. Commenter les résultats.

4 Etude pratique avec R

Le but de cette section est d'appliquer le classifieur bayésien sur le dataset IRIS :

1. **Exploration de l'ensemble des données IRIS :**

Caractéristiques :

```
dim(iris)
names(iris)
str(iris)
attributes(iris)
```

Contenu par ligne ou par colonne :

```
iris[1:5,]
head(iris)
tail(iris)
iris$Sepal.Length[1:10]
```

Statistiques :

```
summary(iris)
var(iris$Sepal.Length)
hist(iris$Sepal.Length)
quantile(iris$Sepal.Length)
quantile(iris$Sepal.Length, c(.1, .3, .65))
table(iris$Species)
pie(table(iris$Species))
```

Plot :

```
plot(density(iris$Sepal.Length))
barplot(table(iris$Species))
boxplot(Sepal.Length~Species, data=iris)
with(iris, plot(Sepal.Length, Sepal.Width, col=Species, pch=as.numeric(Species)))

library(scatterplot3d)
scatterplot3d(iris$Petal.Width, iris$Sepal.Length, iris$Sepal.Width)

library(MASS)
parcoord(iris[1:4], col=iris$Species)
```

2. Devision en train et test

```
set.seed(1234)
ind <- sample(2, nrow(iris), replace=TRUE, prob=c(0.7, 0.3))
trainData <- iris[ind==1,]
testData <- iris[ind==2,]
trainData
testData
```

3. Classifieur bayésien

```
library(e1071)
model=naiveBayes(Species ~ .,data=trainData)
attributes(model)
model$apriori
```

4. Prédiction et matrices de confusion

```
trainPredBayes <- predict(model, newdata = trainData)
trainPredBayes
table(trainPredBayes, trainData$Species)
testPredBayes <-predict(model,newdata=testData)
table(testPredBayes, testData$Species)
```

Exercice 6 Parmi les attributs du modèle bayésien : nous retrouvons les attributs *a priori* et *tables*. Que représentent ces attributs et à quoi ils servent ?

Exercice 7 En examinant la matrice de confusion sur l'ensemble de test, donner la précision et le rappel de chaque classe ainsi que l'erreur sur l'ensemble de test.

Exercice 8 Ecrire une fonction Evaluator qui prend en paramètre trois argument : un modèle, un ensemble de test ainsi que le nombre de colonnes et qui affiche la matrice de confusion, la précision et le rappel de chaque classe ainsi que l'erreur.

Exercice 9 Refaire le même travail sur le dataset *PimaIndiansDiabetes* de la librairie *mlbench*