CY TECH	ING2-GI : IA THEORIE ET ALGORITHMES NAÏVE BAYES
Rédigé par : Astrid Jourdan	A l'intention de : Elèves ING2-GSI
Durée : 3h	

Exercice 1

Considérons l'exemple très simple sur le football.

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

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 ?

1) Quelles sont les deux probabilités à calculer pour réponde à la question.
2) Transformer ces probabilités à l'aide de la formule de Bayes.
3) Appliquer l'hypothèse d'indépendance. Conclusion.

<u>Correction</u>

Notons

X_1 : Match à domicile / X_2 : Balance positive / X_3 : Mauvaises conditions climatiques / X_4 : Match précédent gagné

Y : Match gagné

On cherche

$$P_V=P(Y=V|X_1=V,X_2=V,X_3=F,X_4=V)$$
$$P_F=P(Y=F|X_1=V,X_2=V,X_3=F,X_4=V)$$

Avec la formule de Bayes, on a

$$P_V=P(X_1=V,X_2=V,X_3=F,X_4=V|Y=V)*P(Y=V)/\ P(X_1=V,X_2=V,X_3=F,X_4=V)$$
$$P_F=P(X_1=V,X_2=V,X_3=F,X_4=V|Y=F)*P(Y=F)/\ P(X_1=V,X_2=V,X_3=F,X_4=V)$$

On simplifie par le dénominateur qui est commun aux deux expressions (positif donc ne change rien au problème de maximisation), et on note

$$pp_V=P(X_1=V,X_2=V,X_3=F,X_4=V|Y=V)*P(Y=V)$$
$$pp_F=P(X_1=V,X_2=V,X_3=F,X_4=V|Y=F)*P(Y=F)$$

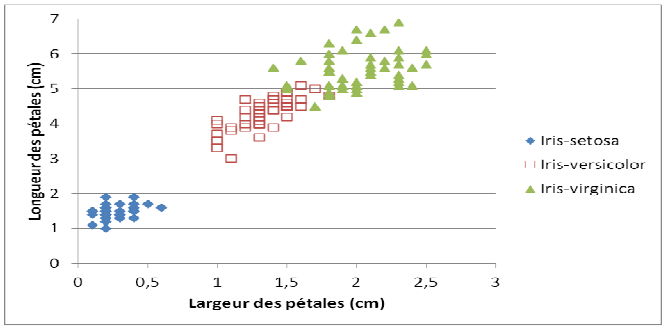
On suppose que les variables conditionnelles $X_i|Y=V$ et $X_i|Y=F$ sont indépendantes, d'où

$$pp_V=P(Y=V)\times P(X_1=V|Y=V)\times P(X_2=V|Y=V)\times P(X_3=F|Y=V)\times P(X_4=V|Y=V)$$
$$=(1/2)\times(3/4)\times(3/4)\times(1/2)\times(1/2)=9/128$$
$$pp_F=P(Y=F)\times P(X_1=V|Y=F)\times P(X_2=V|Y=F)\times P(X_3=F|Y=F)\times P(X_4=V|Y=F)$$
$$=(1/2)\times(1/2)\times(1/4)\times(1/4)\times(1/2)=1/128$$
$$\Rightarrow pp_V>pp_F$$

On peut donc pronostiquer un match gagnant !

Exercice 2

Considérons le fameux jeu de données « Iris » (http://archive.ics.uci.edu/ml/datasets/Iris) représenté ci-dessous. Il s'agit de trois types d'Iris caractérisés entre autre par la longueur et la largeur des pétales.



1) Calculez les moyennes et variances par classe et en déduire la forme du classifieur.
2) Déterminez la forme du classifieur et donnez la classe prédite pour les iris suivants

	longueur des pétales	largeur des pétales
Iris 1	2	0,2
Iris 2	5	1,65
Iris 3	7	3

3) A l'aide du logiciel R, faites un test pour savoir si les variables conditionnelles suivent une loi normale. Comparez la distribution des variables avec la distribution
4) L'hypothèse d'indépendance vous semble-t-elle vérifiée ?
5)

Correction

1) *Probabilités conditionnelles*

Notons Y le type d'iris, X1 la longueur et X2 la largeur. On cherche l'espèce la plus probable connaissant la longueur et la largeur des pétales :

$$P(Y=\text{setosa}|X_1,X_2) \overset{\text{Bayes}}{=} \frac{p(\text{setosa})\times P(X_1,X_2|\text{setosa})}{P(X_1,X_2)}$$

$$P(Y=\text{virginica}|X_1,X_2) \overset{\text{Bayes}}{=} \frac{p(\text{virginica})\times P(X_1,X_2|\text{virginica})}{P(X_1,X_2)}$$

$$P(Y=\text{versicolor}|X_1,X_2) \overset{\text{Bayes}}{=} \frac{p(\text{versicolor})\times P(X_1,X_2|\text{versicolor})}{P(X_1,X_2)}$$

On a

$$p(\text{versicolor})=p(\text{virginica})=p(\text{setosa})=\frac{50}{150}=\frac{1}{3}$$

Donc on peut simplifier le calcul des probabilités conditionnelles par p(classe) et $P(X_1,X_2)$. Si on suppose que les variables conditionnelles $X_i|Y$=classe sont indépendantes, alors le problème consiste à calculer :

$$pp_{\text{setosa}} = P(X1=x_1|\text{setosa})\times P(X2=x_2|\text{setosa})$$

$$pp_{\text{virginica}} = P(X1=x_1|\text{virginica})\times P(X2=x_2|\text{virginica})$$

$$pp_{\text{versicolor}} = P(X1=x_1|\text{versicolor})\times P(X2=x_2|\text{versicolor})$$

2) *Moyennes et variances des variables conditionnelles*

```
data(iris) # les données iris sont déjà dans R
attributes(iris) # pour avoir des info sur le jeu de données (ex.noms des variables)
summary(iris) # résumé numérique de chaque variable

# pour la classe setosa
c.setosa=subset(iris, Species=="setosa")
```

```
mean(c.setosa$Petal.Length)
var(c.setosa$Petal.Length)
```

	Petal length		Petal width	
	mean	var	mean	var
Iris-setosa	1,46	0,03	0,24	0,01
Iris-versicolor	4,26	0,22	1,33	0,04
Iris-virginica	5,55	0,30	2,03	0,08

3&4) Construction du modèle

On suppose que $X_1|\text{setosa} \sim N(1{,}46 ; 0{,}03)$ et $X_2|\text{setosa} \sim N(0{,}24 ; 0{,}01)$. D'où

$$pp_{setosa} = \frac{1}{\sqrt{0{,}03 \times 2\pi}} e^{-\frac{1}{2}\frac{(x_1-1{,}46)^2}{0{,}03}} \times \frac{1}{\sqrt{0{,}01 \times 2\pi}} e^{-\frac{1}{2}\frac{(x_2-0{,}24)^2}{0{,}01}}$$

On suppose que $X_1|\text{versicolor} \sim N(4{,}26 ; 0{,}22)$ et $X_2|\text{versicolor} \sim N(1{,}33 ; 0{,}04)$. D'où

$$pp_{versicolor} = \frac{1}{\sqrt{0{,}22 \times 2\pi}} e^{-\frac{1}{2}\frac{(x_1-4{,}26)^2}{0{,}22}} \times \frac{1}{\sqrt{0{,}04 \times 2\pi}} e^{-\frac{1}{2}\frac{(x_2-1{,}33)^2}{0{,}04}}$$

On suppose que $X_1|\text{virginica} \sim N(5{,}55 ; 0{,}3)$ et $X_2|\text{virginica} \sim N(2{,}03 ; 0{,}08)$. D'où

$$pp_{virginica} = \frac{1}{\sqrt{0{,}03 \times 2\pi}} e^{-\frac{1}{2}\frac{(x_1-1{,}46)^2}{0{,}03}} \times \frac{1}{\sqrt{0{,}01 \times 2\pi}} e^{-\frac{1}{2}\frac{(x_2-0{,}24)^2}{0{,}01}}$$

3) Prévision

```
pp=function(x1,x2,espece)
{
  # Entrées :
  # x1 = longueur du pétal
  # x2 = largeur du pétal
  # espece = subset de iris en fonction de setosa, versicolor, virginica
  # sortie : la "probabilité" (simplifiée) d'appartenir à la classe de espece pour le
nouvel iris (x1,x2)

  mu1=mean(espece$Petal.Length)
  mu2=mean(espece$Petal.Width)
  sigma1=var(espece$Petal.Length)
  sigma2=var(espece$Petal.Width)
  exp(-((x1-mu1)^2)/(2*sigma1)-((x2-mu2)^2)/(2*sigma2))/(sqrt(sigma1*sigma2))
}
```

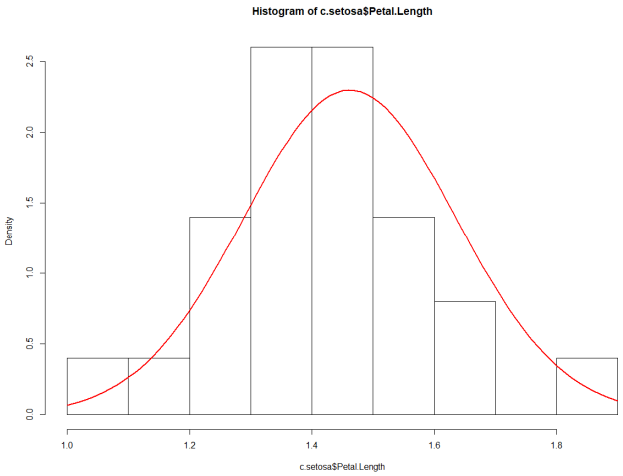
	longueur des pétales	largeur des pétales	pp_{setosa}	$pp_{versicolor}$	$pp_{virginica}$
Iris 1	2	0,2	0.41	10^{-12}	10^{-18}
Iris 2	5	1,65	10^{-127}	0.81	1,57
Iris 3	7	3	0	10^{-22}	$4x10^{-4}$

Dans le cas de l'iris 2, le type prédit est versicolor mais il n'y a pas beaucoup de différence entre le $pp_{versicolor}$ et $pp_{virginica}$. Ce qui est normal étant donné que l'iris 2 est à la frontière entre ces deux classes.

4) Vérification des hypothèses

```
#pour la classe setosa
hist(c.setosa$Petal.Length,proba=T) # histogramme
x=seq(min(c.setosa$Petal.Length),max(c.setosa$Petal.Length),by=0.01) # vecteur de valeurs
pour la longueur des petals avec un pas de 0.01
```

```
lines(x,dnorm(x,mu,sigma),col="red",lwd=2) # dnorm(x) donne la densité d'une loi normale au point x / lines ajoute une courbe à un graphique
```



La distribution de la longueur des pétales au sein de la classe Setosa (histogramme) est proche de la distribution gaussienne estimée (courbe rouge).

5) Matrice de corrélation par classe

```
cor(c.setosa[,3:4]) # matrice de corrélation entre les colonnes 3 et 4
             Petal.Length Petal.Width
Petal.Length      1.00000     0.33163
Petal.Width       0.33163     1.00000
```

Pour la classe Setosa, on note une faible corrélation entre la longueur et la largeur. L'hypothèse d'indépendance des variables n'est donc pas contredite (ce qui ne veut pas dire qu'elle est confirmée).
Faire la même analyse avec les autres classes.

Exercice 3

Il s'agit maintenant d'utiliser le logiciel R pour construire le classifieur bayésien. Pour cela, nous allons utiliser les fonctions `naiveBayes` et `predict` du package `e1071.`

Installez et chargez le package `e1071`

1) Exemple guidé avec le jeu de données Iris.

L'instruction

```
naiveBayes(Y ~ X1+X2+…,data=Mydata)
```

permet de construire le classifieur bayésien où Y est la classe, X1,X2,… les variables explicatives du jeu de données Mydata.

```
model=naiveBayes(Species ~ Petal.Length+Petal.Width,data=iris))
```

Le modèle est constitué des probabilités de chaque classe (A-priori probabilities) et des moyennes et écart-types des variables dans chaque classe (Conditional probabilities).

L'instruction

```
predict(model,newdata=Mynewdata)
```

permet de prédire la classe des nouveaux points de Mynewdata avec le modèle obtenu à l'aide de `naiveBayes.`
Si on ajoute l'argument `type="raw",` on obtient la probabilité de chacune des classes.

```
# Nouveaux iris (cf. exo 2)
newiris=matrix(0,3,2)
```

```
newiris[1,]=c(2,0.2)
newiris[2,]=c(5,1.65)
newiris[3,]=c(7,3)
newiris=as.data.frame(newiris)
names(newiris)=names(iris)[3:4] # "Petal.Length" et "Petal.Width"

predict(model,newdata=newiris,type="raw")
```

On retrouve bien les conclusions de l'exercice 2, à savoir une quasi-équiprobabilité pour l'iris 2 d'être dans deux classes.

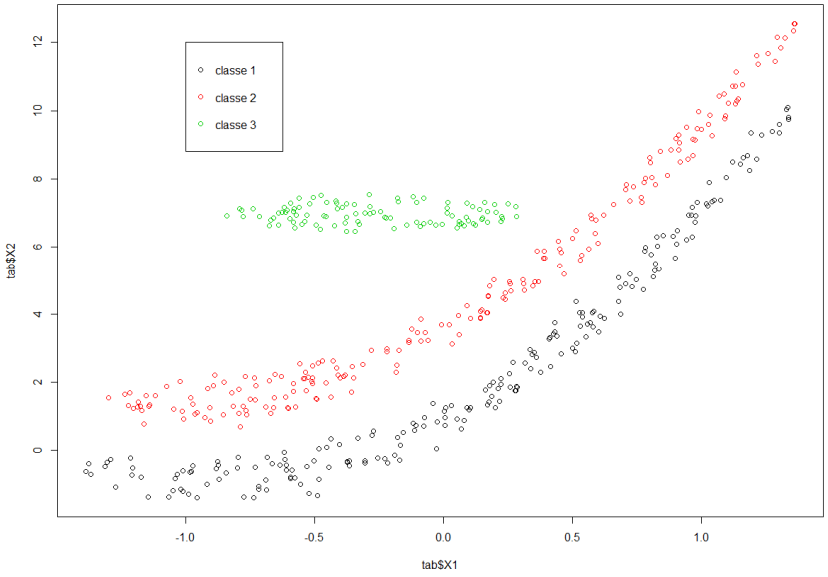
2) Application sur un jeu de données simulé en 2D : Test_Classif_Correl.txt

```
tab=read.table("Test_Classif_Correl.txt",header=T)
summary(tab)
```

a) Représentez le nuage de points. Pensez-vous que la méthode naïve Bayes est appropriée à ce jeu de données ?
b) Construisez le modèle
c) Calculez la matrice de confusion (sur la base d'apprentissage). Est-ce que les résultats étaient prévisibles ?

Correction

```
# Graphique des données
plot(tab$X1,tab$X2,col=tab$Y)
legend(min(tab$X1),max(tab$X2),c("classe 1","classe 2","classe 3"),col=c(1,2,3),pch=1)

# Verification des distributions
classe1=subset(tab,tab$Y==1)
classe2=subset(tab,tab$Y==2)
classe3=subset(tab,tab$Y==3)
shapiro.test(classe1$X1)

#       Shapiro-Wilk normality test

#data:  classe1$X1
#W = 0.96436, p-value = 6.038e-05

boxplot(classe1$X1,classe2$X1,classe3$X1,main="X1")
boxplot(classe1$X2,classe2$X2,classe3$X2,main="X2")
# les distributions des variables dans chaque classe ne sont pas gaussiennes d'après le
test de Shapiro. Cependant les boxplot montrent qu'elles sont suffisamment sympathiques
pour être supportées par la méthode.

# Vérification de l'indépendance
cor(classe1$X1,classe1$X2)
cor(classe2$X1,classe2$X2)
cor(classe3$X1,classe3$X2)
# Cela confirme les conclusions sur le nuage de points. Les variables sont très corrélées
dans les classes 1 et 2. Elles ne sont pas corrélées dans la classe 3, ce qui ne veut pas
dire qu'elles sont indépendantes mais faute de preuve, on peut le supposer. On note aussi
que si on peut établir que la distribution est gaussienne dans la classe 3 alors non
corrélation entraine indépendance (mais uniquement dans le cas gaussien !!!)

# transformation des variables
tab[,1:2]=scale(tab[,1:2]) # Centre et réduit les variables quantitatives
tab$Y=as.factor(tab$Y) # Transformation de la classe en variable qualitative

# Naive Bayes
model=naiveBayes(Y ~. ,data=tab) # construction du modèle
prev= predict(model,tab[,-3]) # prédiction sur la base d'apprentissage
table(tab$Y,prev) # matrice de confusion
predict(model,tab[,-3],type="raw") # probabilité de chaque classe
```



Matrice de confusion

```
prev
      1   2   3
1   147  53   0
2   105  95   0
3     0   0 100
```

Aucune surprise, il y a une très grande corrélation entre X1 et X2 dans les classes 1 et 2 et il a beaucoup de mal à les séparer.

Proba pour chaque classe

```
                1           2             3
[1,] 0.55882199 0.44117801  1.275083e-46
[2,] 0.66956539 0.33043461 1.672407e-143
[3,] 0.53977553 0.46022447  1.190663e-38
[4,] 0.24944958 0.75055042  1.547304e-11
[5,] 0.64554382 0.35445618 2.656447e-105
[6,] 0.66009505 0.33990495 9.826320e-119
```

…

Exercice 4

Mettre en place la méthode Naïve Bayes sur le jeu de donnée Landsat ou Frogs.