TP : Analyse Discriminante

1 Données artificielles

On considère le jeu de données artificielles (cf. Table 1) caractérisés par deux variables quantitatives x_1 et x_2 et une variable qualitative y à deux modalités, notées A et B. Le poids de chaque individu est choisi égal à 1/8.

	x_1	x_2	y
P_1	-2	0	A
P_2	2	-2	B
P_3	-1	1	A
P_4	-1	-1	A
P_5	-1	-2	B
P_6	0	0	A
P_7	-1	2	B
P_8	2	2	B

Table 1

Dans ce qui suit, on cherche à expliquer y en fonction de x_1 et x_2. Pour cela on réalise différentes classifications supervisées à l'aide du logiciel R.

1.1 Création des données dans le logiciel R

```
donnee1 <- matrix(c(-2,2,-1,-1,-1,0,-1,2,0,-2,1,-1,-2,0,2,2),ncol=2,byrow=F)
donnee2 <- matrix(c("A","B","A","A","B","A","B","B"),ncol=1,byrow=F)
donn<- cbind.data.frame(donnee1,donnee2)
nomligne<-c("P1","P2","P3","P4","P5","P6","P7","P8")
nomcol<-c("X1","X2","Y")
dimnames(donn)<-list(nomligne,nomcol)
donn
```

1.2 Analyse Discriminante Linéaire

La commande "lda" permet de réaliser une Analyse Discriminante Linéaire. Cette commande effectue une analyse discriminante bayésienne en supposant que chaque classe C_i est modélisée par une loi normale de même matrice variance, et possède une probabilité a priori dont la valeur peut être choisie de façon arbitraire. Il faut donc spécifier la probabilité a priori, notée p_i, de chaque classe ; en général, on choisit soit $p_i = cte = 1/q$ où q est le nombre de classes, soit la proportion d'observations dans C_i, notée p_i. L'argument "prior" est utilisé pour indiquer ce choix ; par défaut c'est le second choix qui est effectué.
La formule utilisée pour spécifier le modèle de discrimination est du type : $Y \sim X_1 + X_2$, si l'on veut expliquer la variable qualitative Y indiquant les classes par les variables quantitatives X_1 et X_2, et simplement "$Y \sim .$" si l'on veut expliquer Y par toutes les variables.

1. Charger d'abord la librairie MASS permettant d'effectuer les analyses discriminantes, puis effectuer une analyse discriminante linéaire (commande "lda") du jeu de données. On stockera les résultats dans "modele1" :

```
library(MASS)
modele1<- lda(Y ~ .,prior=c(0.5,0.5), donn)
predict(modele1, donn)
```

2. Parmi les résultats obtenus, identifier de façon précise :
   (a) les coordonnées du facteur discriminant,
   (b) le score de chaque individu ainsi que sa classe d'affectation,
   (c) les probabilités a posteriori pour chaque individu.
3. Par un calcul à la main, vérifier la valeur obtenue pour le score de P_1 :

1.3 Analyse Discriminante Quadratique

La commande "qda" réalise une analyse discriminante quadratique. Plus précisément, "qda" effectue une classification bayésienne en supposant que chaque classe C_i est modélisée par une loi normale de matrice variance estimée par les données, et permet de choisir arbitrairement la probabilité a priori de C_i. La commande "qda" s'utilise de la même façon que "lda", que ce soit pour sa syntaxe ou pour ses arguments.

1. Reprendre les questions 1. et 2. de la section précédente, mais cette fois en utilisant "qda" au lieu de "lda". On nommera "modele2" le modèle d'analyse discriminante quadratique ainsi obtenu, i.e. le résultat de "qda".

   ```
   modele2 <- qda(Y ~ .,prior=c(0.5,0.5), donn)
   predict(modele2, donn)
   ```

2. Les résultats sont-ils modifiés si l'on change les probabilités a priori ?

1.4 Evaluation de la qualité des modèles

1. Pour chacun des deux modèles, i.e. "modele1", "modele2" :
   (a) **Matrice de confusion :** Utiliser la commande "table" pour construire le tableau de classement (TC), c.-à-d. le tableau qui croise les variables classe d'appartenance (donn[,3]) et classe d'affectation (i.e. "predict(modele, donn)$class" pour les 2 analyses discriminantes, puis déterminer le pourcentage de bien classés lors de l'estimation du modèle. :

   ```
   (TC[1,1]+TC[2,2])/sum(TC)
   ```

   (b) **Leave one out CV :** Pour estimer (sans biais) le taux de mauvais classement par validation, on peut utiliser la validation croisée. Pour cela, il suffit d'ajouter l'argument "CV=TRUE" lors de l'appel de la commande "lda" ou "qda".
   (c) **CV :** Ecrire un scripte qui effectue une validation croisée avec un échantillon d'apprentissage de 80%.
   (d) **Courve ROC :** Installer le package ROCR, puis tracer la courbe ROC et évaluer l'AUC. Noter que la commande "prediction" permet de calculer les paramètres de base nécessaire à la définition de la courbe ROC.

   ```
   library(ROCR)
   modele.posterior<-predict(modele, donn)$posterior[,2]
   modele.pred<-prediction(modele.posterior, donn[,3])
   modele.roc<-performance(modele.pred, "tpr","fpr")
   plot(modele.roc,colorize=TRUE) # avec "add = TRUE" pour le 2 et 3¡ plot
   #   Evaluer l'AUC :
   modele.auc<-performance(modele.pred, "auc") ; modele.auc@y.values[[1]]
   ```

2. Quel modèle vous semble le plus performant ?

2 Données réelles

On considère le jeu de données contenu dans le fichier "don0.txt".
Les données portent sur un échantillon de 100 patients pour lesquels on a relevé les mesures suivantes :

- AGE (en années) ;
- POIDS (en kg) ;
- TAILLE (en cm) ;
- ALCOOL (en nombre de verres bus) ;
- SEXE (F ou H) ;
- RONFLE (ronflement : O = ronfle ; N = ne ronfle pas) ;
- TABA (tabac : O = fumeur ; N = non fumeur).

On cherche à expliquer/prédire la variable "RONFLE" à l'aide des autres variables.

2.1 Différents modèles possibles

1. Effectuer l'analyse discriminante linéaire avec toutes les variables et sans les variables "poids" et "taille".
2. Reprendre les question précédentes avec la fonction qda.
3. Comparer les différents modèles en rajoutant l'argument "CV=TRUE".

2.2 Prédiction

Soit quatre nouveaux patients pour lesquels les valeurs des variables explicatives sont les suivantes :

AGE	POIDS	TAILLE	ALCOOL	SEXE	TABA
42	55	169	0	F	N
58	94	185	4	H	O
35	70	180	6	H	O
67	63	166	3	F	N

Pour effectuer la prédiction, il faut d'abord créer un data-frame contenant les données sur les nouveaux patients : ce data-frame doit posséder la même structure que les données initiales.

1. Construire un data.frame appelé "n_donnes" contenant ces données.
2. En utilisant la commande "predict" sur n_donnees, donner les prédictions d'appartenance de ces quatre patients aux différentes classes, selon le meilleur modèle déterminé précédemment.