TP de classification supervisée : Fiche 2 (TP 2-3)

1 Evaluation de la règle de classement (iris de Fisher)

1. Diagnostic apparent :
   (a) A partir des résultats du TP précédent, calculer la matrice de confusion à l'aide de la fonction `table` sur la classe réelle et la classe prédite.
   (b) Calculer le taux de bon classement (TBC) et le taux de mauvais classement (TMC).
   (c) Expliquer pourquoi la méthode utilisée peut souffrir d'un biais d'optimisme.
2. Diagnostic sur un échantillon test :
   (a) Découper les données iris en un échantillon d'apprentissage (70% des données) et un échantillon test (30% des données) pour construire le tableau de score.
   (b) Faire la prédiction sur l'échantillon test (30%). Ici on écrit une fonction nommée `calcalpha` qui apprend le tableau des coefficients α.

```
calcalpha <-function(X, Y){
d=ncol(X)
k=nlevels(Y)
W=matrix(0,d,d)
ni=table(Y)
for (i in levels(Y)){
W=W+cov.wt(X[Y==i,],method="ML")$cov*ni[i]
}
W=W/sum(ni)
moyennes=by(X,Y,colMeans)
G=matrix(unlist(moyennes),k,d,byrow=T)
B=cov.wt(G,wt = as.vector(table(Y)),method="ML")$cov
alpha=matrix(0,(d+1),k)
rownames(alpha) = c("intercept",colnames(X))
colnames(alpha) = levels(Y)
for (i in 1:k) {
barXi=matrix(G[i,],d,1)
alpha[1,i]=-t(barXi)%*%solve(W)%*%barXi
alpha[2:(d+1),i]=2*solve(W)%*%barXi
}
return (alpha)
}
```

   On peut aussi écrire une fonction `predictY` qui à partir de la matrice des α, prédit la classe pour un tableau X.

```
predictY<- function(X,alpha){
s=as.matrix(cbind(1,X))%*%alpha
Ypredit=colnames(alpha)[apply(s,1,which.max)]
}
```

En déduire les classes prédites sur l'échantillon test.

(c) Calculer le taux de bon classement (TBC) et le taux de mauvais classement (TMC).

3. Par validation croisée leave-one-out, procéder de manière similaire à ci-dessus ; à chaque étape, toutes les données sauf une serviront d'échantillon d'apprentissage, la donnée mise à l'écart servant d'échantillon test.

2 Analyse discriminante linéaire (iris de Fisher)

1. Charger le package `MASS`, puis utiliser la fonction `lda` qui permet d'ajuster le modèle d'analyse discriminante linéaire. En utilisant les fonctions `predict` et `table`, réaliser la matrice de confusion.

2. Evaluer le taux de mauvais classement par validation croisée leave-one-out, en utilisant l'option `CV = TRUE` dans la fonction `lda`. Réaliser aussi la matrice de confusion. On remarque que les classes d'affectation et les probabilités a posteriori sont directement renvoyées dans l'objet de sortie, sans faire appel à `predict`.

3. Utiliser à nouveau la fonction lda sans préciser l'option CV=TRUE. Dans les sorties on remarque que la fonction retourne les coefficients linéaires discriminants LD1 et LD2. Ceux-ci peuvent être récupérés par le champ scaling de l'objet retourné par la fonction lda (cette sortie n'est pas disponible dans le cas où on a CV=TRUE). En multipliant la matrice de données par la matrice des coefficients linéaires discriminants, obtenir une projection des individus sur ces axes discriminants. Faire le graphique permettant de visualiser ces données.