Master MAS de Bordeaux
M2 : classification supervisée

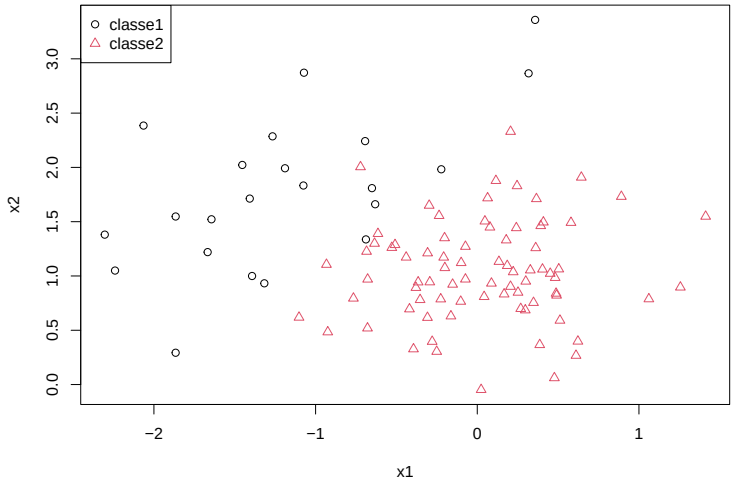
TP2 : analyse discriminante linéaire et quadratique

Exercice 1. Objectifs :

- découvrir l'analyse discriminante linéaire et quadratique.
- découvrir fonctions `lda` et `qda`.

On utilise dans cet exercice les données synth_train.txt et synth_test.txt.

```
train <- read.table(file="../data/synth_train.txt", header=TRUE)
Xtrain <- train[,-1]
Ytrain <- train$y
plot(Xtrain, pch=Ytrain, col=Ytrain)
legend("topleft", legend=c("classe1", "classe2"), pch=1:2, col=1:2)
```



1. En **analyse discriminante quadratique** on fait l'hypothèse paramétrique gaussienne que $X = (X^1, X^2) \sim \mathcal{N}(\mu_k, \Sigma_k)$:

$$f(x|Y=k) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} \exp(-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k))$$

Les paramètres inconnus μ_k, Σ_k et les probabilités à priori $\pi_k = \mathbb{P}(Y=k)$ pour $k = 1, 2$ sont estimés par :

$$\widehat{\pi}_k = \frac{n_k}{n}, \quad \widehat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

$$\widehat{\Sigma}_k = \frac{1}{n_k} \sum_{i:y_i=k} (x_i - \widehat{\mu}_k)(x_i - \widehat{\mu}_k)^T$$

Estimer ces paramètres sur les données d'apprentissage.

2. La règle de décision de Bayes (prédire la classe la plus probable à posteriori) s'écrit alors :

$$g(x) = \underset{\ell \in \{1,2\}}{\arg\max}\ Q_\ell(x) \tag{1}$$

avec

$$Q_\ell(x) = -\frac{1}{2}\log|\Sigma_\ell| - \frac{1}{2}(x-\mu_\ell)^T\Sigma_\ell^{-1}(x-\mu_\ell) + \log(\pi_\ell) \tag{2}$$

où Q_ℓ est appellée **fonction discriminante quadratique**.

Calculer $Q_1(x)$ et $Q_2(x)$ pour une nouvelle donnée $x = (-1, 1)$ et vérifier que x est bien affectée à la classe 2.

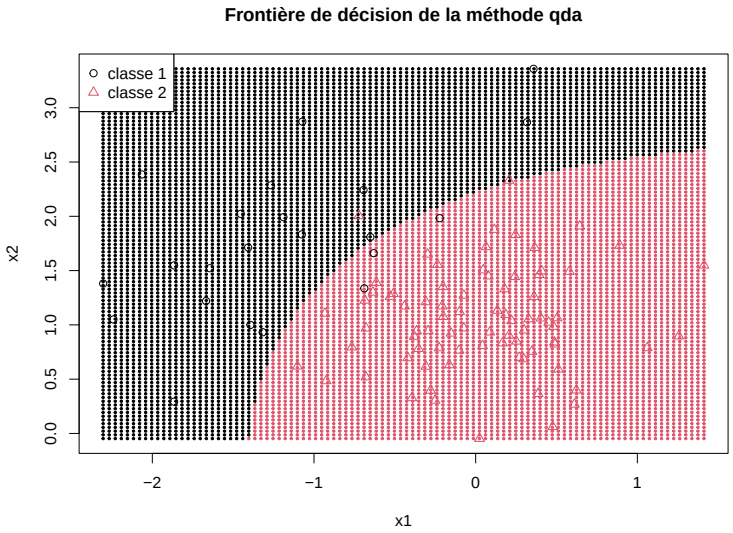
3. On sait de plus que les probabilités à posteriori des classes se calculent de la manière suivante :

$$\mathbb{P}(Y = k|X = x) = \frac{\exp\left(Q_k(x)\right)}{\sum_{\ell=1}^{K}\exp\left(Q_\ell(x)\right)}$$

Estimer les probabilités à posteriori pour $x = (-1, 1)$.

4. Utiliser maintenant les fonctions `qda` et `predict.qda` du package `MASS` pour prédire la classe du point $x = (-1, 1)$. Vérifier que vous retrouvez les résultats obtenus précédemment.

5. La méthode `qda` construit une frontière de décision quadratique que l'on peut représenter graphiquement.



6. En **analyse discriminante linéaire**, on ajoute l'hypothèse que **les matrices de covariance sont égales**. L'estimateur de la matrice de covariance $\Sigma = \Sigma_1 = \Sigma_2$ est :

$$\widehat{\Sigma} = \frac{1}{n}\sum_{k=1}^{2} n_k\widehat{\Sigma}_k$$

Estimer cette matrice sur les données d'apprentissage.

7. La règle de décision de Bayes (prédire la classe la plus probable à posteriori) s'écrit alors :

$$g(x) = \underset{\ell \in \{1,2\}}{\arg\max}\ L_\ell(x)$$

avec

$$L_\ell(x) = x^T \Sigma^{-1} \mu_\ell - \frac{1}{2} \mu_\ell^T \Sigma^{-1} \mu_\ell + \log(\pi_\ell)$$

où L_ℓ est appellée **fonction discriminante linéaire**.

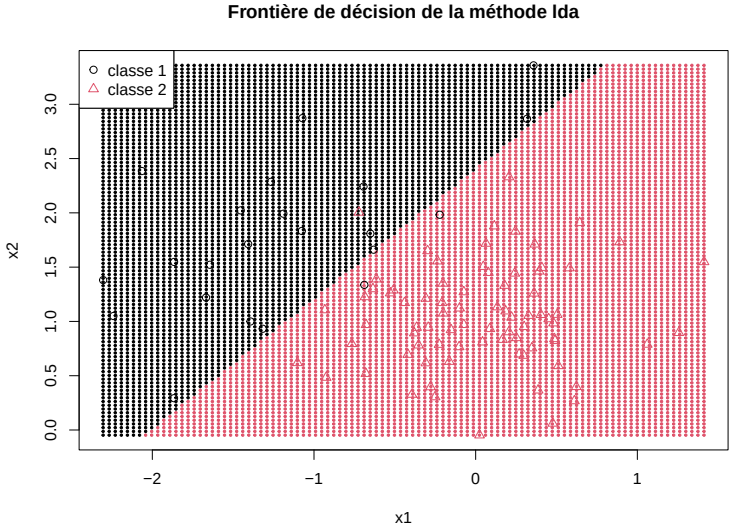
Calculer $L_1(x)$ et $L_2(x)$ pour la nouvelle donnée $x = (-1, 1)$ et vérifier que x est bien affecté à la classe 2.

8. On sait de plus que les probabilités à posteriori des classes se calculent de la manière suivante :

$$\mathbb{P}(Y = k | X = x) = \frac{\exp\left(L_k(x)\right)}{\sum_{\ell=1}^{K} \exp\left(L_\ell(x)\right)}$$

Estimer les probabilités à posteriori pour $x = (-1, 1)$.

9. Utiliser maintenant les fonctions `lda` et `predict.lda` pour prédire la classe du point $x = (-1, 1)$. Vérifier que vous retrouvez les résultats obtenus précédemment.
10. La méthode `lda` construit une frontière de décision linéaire que l'on peut représenter graphiquement. Retrouver cette représentation.



Exercice 2. Faire de la reconnaissance automatique de caractères manuscrits.

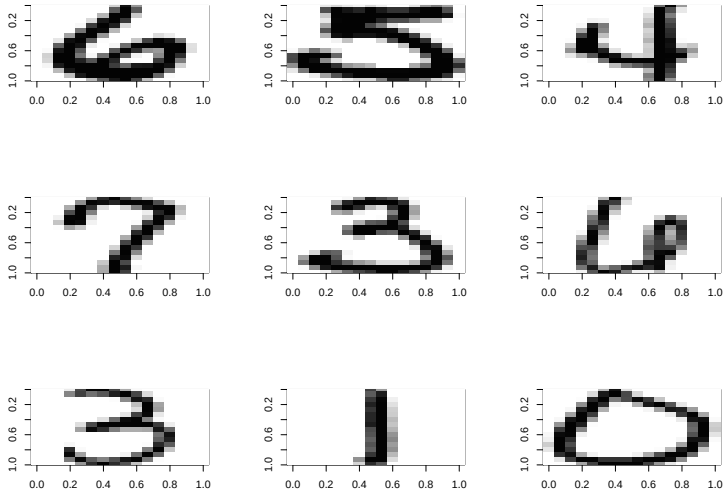
Récupérer les jeux de données numbers_train.txt et numbers_test.txt. Chaque fichier contient 500 images de dimension 16×16 et chaque image représente un caractère manuscript (un chiffre entre 0 et 9). On a donc $Y \in \{0, \ldots, 9\}$ et $X = (X^1, \ldots, X^{256}) \in \mathbb{R}^{256}$. Il s'agit d'image en niveaux de gris où chaque pixel prend une valeur entre 0 (noir) et 1 (blanc).

1. Importer les 500 images du fichier numbers_train.txt.

```
data <- read.table("../data/numbers_train.txt", header=TRUE)
Xtrain <- as.matrix(data[,-1])
Ytrain <- as.factor(data[,1])
```

2. Visualiser les neuf premières images.

```
par(mfrow=c(3,3))
for (i in 1:9){
  image(matrix(Xtrain[i,],16,16), col=gray(1:100/100), ylim=c(1,0))
}
```



3. Prédire avec la méthode `lda` les classes des 500 images de l'ensemble d'apprentissage et calculer le taux d'erreur d'apprentissage.
4. Importer les 500 images du fichier numbers_train.txt.
5. Prédire maintenant les classes des 500 images de l'ensemble test et calculer le taux d'erreur test.
6. Essayer matinenant d'utiliser la méthode `qda`. Que constatez-vous ? Expliquer pourquoi ?
7. Enfin, estimer l'erreur de la méthode `lda` par validation croisée LOO (Leave One Out) puis par validation croisée 5-folds. Commentez vos résultats.

Exercice 3. L'objectif est d'implémenter l'analyse discriminante quadratique.

— Implémenter une fonction `adq_estim` qui prend en entrée une matrice de données X et un vecteur de classes Y et estime le vecteur de paramètres inconnus $\theta = (\pi_1, \ldots, \pi_K, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K)$. La sortie de cette fonction doit être une liste de taille K et chaque élément de cette liste doit être la liste des estimations de π_k, μ_k et Σ_k. N'oubliez pas de nommer les éléments des listes afin de

rendre les résultats plus lisibles. Vérifiez enfin que vous retrouvez les résultats de la question 1) de l'exercice 1 avec cette fonction.

— Implémenter une fonction `adq_pred` qui prend en entrée les paramètres estimés avec la fonction `adq_estim` et une matrice de nouvelles données à prédire. Cette fonction doit fournir en sortie une liste avec le vecteur des prédictions et la matrice des probabilités à posteriori de ces nouvelles observations. Vérifiez enfin que vous retrouvez les résultats des questions 2) et 3) de l'exercice 1 avec cette fonction.

Exercice 4. Sélectionner des variables.

On utilise dans cet exercice des données où $n = 1260$ exploitations agricoles sont décrites par :

— 22 variables quantitatives correspondant à 22 critères économiques et financiers,

— une variable qualitative à deux modalités qui indique si l'exploitation agricole est en difficulté de paiement (0=sain, 1=défaillant).

Les données sont présentées ici : `http://www.modulad.fr/archives/numero-31/desbois-31/desbois-31.pdf`

```
load("../data/Desbois_complet.rda")
colnames(data)

##  [1] "DIFF" "R1"   "R2"   "R3"   "R4"   "R5"   "R6"   "R7"   "R8"   "R11"
## [11] "R12"  "R14"  "R17"  "R18"  "R19"  "R21"  "R22"  "R24"  "R28"  "R30"
## [21] "R32"  "R36"  "R37"
```

1. Utiliser la code ci-dessous pour estimer le taux d'erreur test de la méthode `lda`.

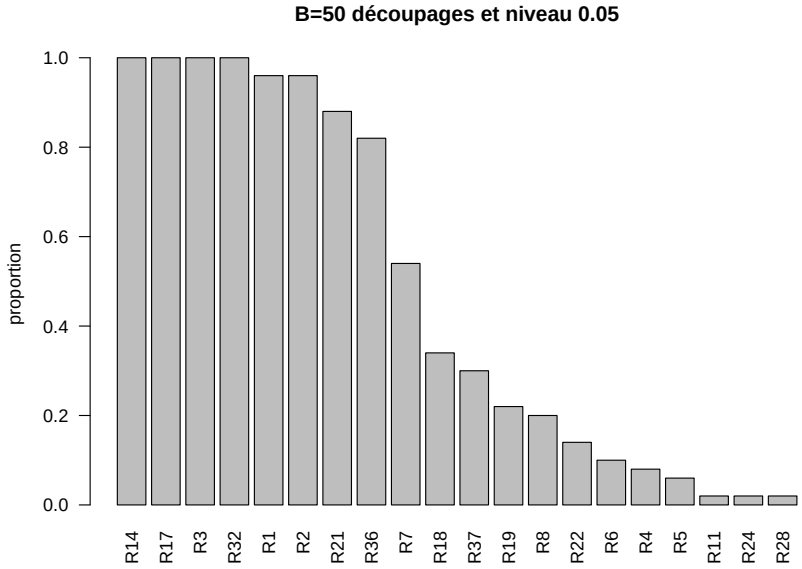
```
library(MASS)
# Indices des 75% de données d'apprentissage
set.seed(10)
tr <- sample(1:nrow(data),945)
# Métode lda avec la syntaxe utilisant les formules
g <- lda(DIFF~., data=data[tr,])
pred <- predict(g, data[-tr,-1])$class
sum(pred != data[-tr,1])/length(pred)
# Taux d'erreur test (pour ce découpage) de 15.2%
```

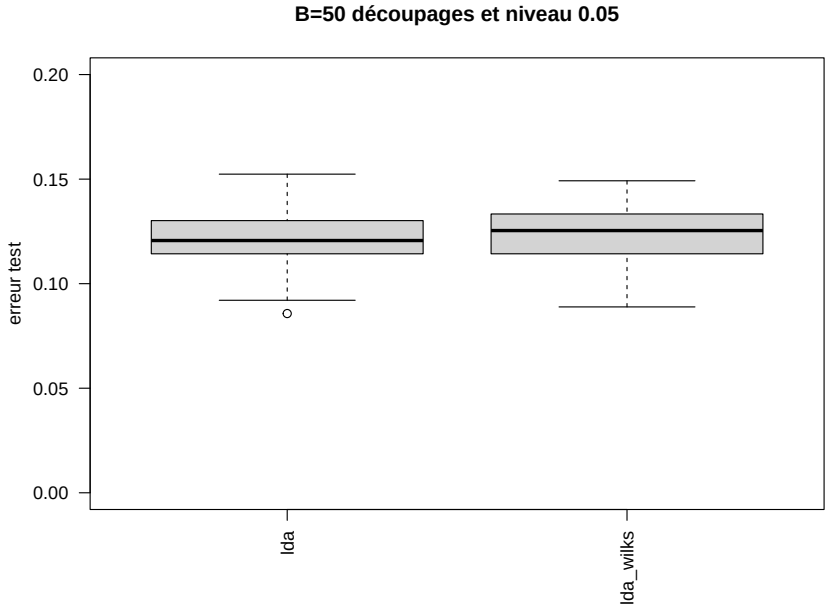
2. Recommencer avec la méthode `greedy.wilks` du package `klar`. Que fait cette fonction ?

```
library(klaR)
?greedy.wilks
formula <- greedy.wilks(DIFF~., data=data[tr,])$formula
formula
g <- lda(formula, data=data[tr,])
pred <- predict(g, data[-tr,-1])$class
sum(pred != data[-tr,1])/length(pred)
# Taux d'erreur test (avec ces variables et pour ce découpage) de 14.6%
```

3. A quoi correspond l'argument `niveau` ? Que se passe-t-il lorsqu'on le modifie ?
4. Refaire les trois questions précédentes avec un autre découpage apprentissage/test. Les variables sélectionnées sont-elles différentes ? Les taux d'erreur test sont-ils différents ?

5. Ecrire le code permettant de retrouver les graphiques ci-dessous et commenter ces résutats.





6. On décide de choisir la méthode `lda_wilks` ? Quelles sont les variables finalement sélectionnées ? Prédire avec cette méthode si la 3ème exploitation agricole est saine ou défaillante. Quelle confiance a-t-on dans cette prédiction ?