Classification supervisé

TP6 : Bayésien naif

Exercice 1. Objectifs :

- Découvrir les fonctions R qui font du Bayésien naif.
- Comparer les performances du Bayésien naif au LDA et au QDA sur les données "Desbois".

Dans R, la méthode du Bayésien naif est implementée dans le package `e1071` avec la fonction `naiveBayes` et dans le package `klaR` avec la fonction `NaiveBayes`. La fonction du package `klaR` implémente en plus l'estimation non paramètrique de densité (à noyau) pour les variables d'entrée quantitatives.

1. Exectuez pas à pas le code ci-dessous et commentez le.

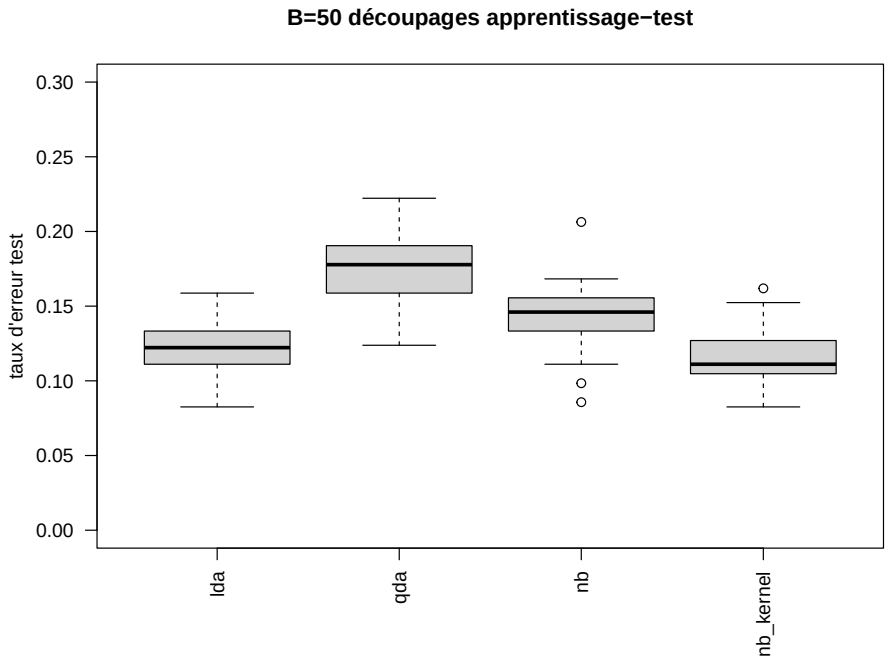
```
library(e1071)
## Données d'entrée binaires
data(HouseVotes84, package = "mlbench")
?mlbench::HouseVotes84
g <- naiveBayes(Class ~ ., data = HouseVotes84)
g$apriori
g$tables
predict(g, HouseVotes84[1,])
predict(g, HouseVotes84[1,], type = "raw")
pred <- predict(g, HouseVotes84)
table(pred, HouseVotes84$Class)

# Données d'entrée quantitatives
data(iris)
g <- naiveBayes(Species ~ ., data = iris)
## ou encore:
#m <- naiveBayes(iris[,-5], iris[,5])
g$apriori
g$tables
table(predict(g, iris), iris[,5])

library(klaR)
?NaiveBayes
m <- NaiveBayes(Species ~ ., data = iris)
mnames(predict(m))
table(predict(m)$class, iris[,5])

m2 <- NaiveBayes(Species ~ ., data = iris, usekernel=TRUE)
names(predict(m2))
table(predict(m2)$class, iris[,5])
```

2. Reprendre les données "Desbois" de l'exercice 3 de la feuille de TP5 et ajouter la méthode Bayésien naif aux boxplots de la question 5.



Exercice 2. Objectif : Implémenter le Baysésien naif pour des données d'entrées quantitatives.

On considère de nouveau le jeu de données "synth_train.txt" et "synth_test.txt". On note X^1 et X^2 les deux coordonnées de X. On rappelle que le bayésien naïf est une approche indirecte où l'on fait l'hypothèse d'indépendance des variables d'entrée conditionnellement à Y :

$$\forall x \in \mathbb{R}^2,\ \forall k \in \{1,2\} \quad f_k(x) = f_{k,1}(x_1) f_{k,2}(x_2)$$

où f_k est la densité conditionnelle de X sachant $\{Y = k\}$, et $f_{k,1}$ et $f_{k,2}$ sont les densités conditionnelles respectivement de X^1 et X^2 sachant $\{Y = k\}$. De plus on suppose que pour tout $k \in \{1,2\}$ et tout $j \in \{1,2\}$ la loi de X^j sachant $\{Y = k\}$ est $\mathcal{N}(\mu_{kj}, \sigma^2_{kj})$.

1. Quels sont les paramètres de ce modèle bayésien naïf et leurs estimateurs ?
2. Quelle est alors la règle de classification de la méthode Bayésien naif.
3. Implémenter cette règle de décision.
   (a) On pourra commencer par créer une fonction bn_estim qui estime les paramètres de ce modèle bayésien naïf sur des données d'apprentissage.
   (b) Puis, on pourra écrire une fonction bn_predict qui prédit la classe de données test (cette fonction utilisera les paramètres estimés par la fonction précédente).
4. Tester vos fonctions : utiliser la fonction bn_estim pour estimer les paramètre avec les données d'apprentissage "synth_train.txt", puis utiliser la fonction bn_predict pour prédire la classe des points $(0,1)$ et $(-2,2)$.

```
train <- read.table(file="../data/synth_train.txt", header=TRUE)
train$y <- as.factor(train$y)
chap <- bn_estim(train[,-1], train$y)
pred <- bn_predict(matrix(c(0,1,-2,2), nrow=2, byrow=TRUE), chap)
pred

## [1] 2 1
```

5. Calculer le taux d'erreur d'apprentissage.

```
## [1] 0.04
```

6. Charger le jeu de données test synth_test.txt puis calculer le taux d'erreur test.

```
test <- read.table(file="../data/synth_test.txt", header=TRUE)
pred <- bn_predict(xtest=test[,-1], chap)
sum(pred!=test$y)/length(test$y)

## [1] 0.05
```