Apprentissage supervisé

TP3 : Bayésien naif

Eercice 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.

```r
library(e1071)
## Données d'entrée binaires
data(HouseVotes84, package = "mlbench")
help(HouseVotes84,package = "mlbench")
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)
names(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 TP2 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 le jeu de données d'apprentissage habituel dispose de 100 données d'apprentissage. On note X^1 et X^2 les deux coordonnées de X. On rappelle que le bayésien naïf est une approche générative où 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. Donner les estimateurs du maximum de vraisemblance de tous les paramètres du bayésien naïf considéré.
2. Ecrire la règle de décision associée.
3. Charger le jeu de données dans R. Transformer la variable de sortie y en facteur.
4. Implémenter la méthode du bayésien naïf.
   (a) On pourra commencer par créer une fonction bn_estim qui prend en argument les données et calcul les estimateurs des différents paramètres associés à la modélisation du bayésien naïf.
   (b) Puis, on pourra écrire une fonction bn_predict permettant de prédire la classe associée à une observation x (cette fonction utilisera les paramètres estimés par la fonction précédente).
5. Tester les fonctions : appliquer la fonction bn_estim avec l'échantillon d'apprentissage, puis utiliser la fonction bn_predict pour prédire les points de coordonnées $(0,1)$ et $(-2,2)$.
6. Calculer le taux d'erreur d'apprentissage du bayésien naïf.
7. Charger le jeu de données test synth_test.txt puis calculer le taux d'erreur test du bayésien naïf.