Université Paris Nanterre
Master2 ISEFAR Data Mining
Classification (Apprentissage Supervisée)
CM et TP : Ana Karina FERMIN

Introduction

En TP l'étudiant doit être capable de savoir implémenter et utiliser les méthodes de classification sous le logiciel R, sur des exemples synthétiques et sur des données réelles. Il faut comprendre les commandes de chaque TP et interpréter les résultats et graphiques.

Le logiciel R et Rstudio :

- **Logiciel R** :

  Avant de commencer ce TP, il est nécessaire de s'assurer que le logiciel R est installé sur votre machine. Le logiciel R est gratuit et peut être téléchargé sur le site du CRAN (Comprehensive R Archive Network) à l'adresse suivante https://cran.r-project.org/. Il existe de nombreuses sources d'informations disponibles en ligne.

- À propos de **RStudio**:

  RStudio est une sur-couche de R (également libre et gratuit) rendant son utilisation plus conviviale. Il peut-être téléchargée sur https://rstudio.com (après avoir téléchargé le logiciel R).

  Lorsque vous ouvrez 'RStudio' pour la première fois, l'interface est divisée par défaut en 3 fenêtres:

  - la console interactive 'R' qui sert à exécuter le code (à gauche) - l'environnement / histoire qui contient l'espace de travail et l'historique des commandes (en haut à droite) - Les Files / Plots / Packages / Help / Viewer (en bas à droite).

- **Script R et Rmarkdown** : Le logiciel R utilise des lignes de commandes. Il faut donc taper les commandes dans la console pour les exécuter. Plutôt que de saisir les commandes directement dans la console, on vous conseille de les enregistrer directement dans un script de commandes R, ce qui permet de reproduire les analyses ultérieurement.

  Rmarkdown est un langage à balise qui permet des générer des rapports en mélangeant du texte, du code R et les résultats produits par ce code. En plus de produire des rapports agréables à lire (le code et ses résultats ne sont pas séparés des commentaires qui leur sont associés), son principal avantage est qu'il produit des rapports qui sont dynamiques et qui rendent les travaux reproductibles. C'est un simple fichier texte dont l'extension est '.Rmd'.

  A chaque séance de TP créer un fichier scrip R ou un Rmarkdown

- **Aide en ligne** : R possède un système d'aide en ligne, incluant un moteur de recherche et un index des commandes associées. Pour accéder à l'aide pour n'importe

quelle fonction, on utilise la commande `help()`. Le logiciel R ouvre une fenêtre avec les informations relatives à la fonction demandée.

- **Packages (commande library)** : R dispose d'un certain nombre de packages de base qui sont installés lors de l'installation de R. Il est possible d'installer des packages additionnels. Le chargement d'un package s'effectue à l'aide de la commande `library()`, en indiquant le nom du package.

- **Doc sur initiation à R** : A vous de regarder la doc !

  Par exemple vous pouvez aller sur la site

  http://www.math.univ-toulouse.fr/ besse/Wikistat/pdf/st-tutor2-R-init.pdf

- Quelques packages importants à utiliser dans ce cours

  - Le package tidyverse :

    Le terme 'tidyverse' est une contraction de tidy (traduit comme "bien rangé") et de universe. Le 'tidyverse' est un ensemble d'extensions pour R conçues pour fonctionner ensemble. Elle facilitent un très grand nombre d'opérations courantes dans R. Plusieurs packages constituent le coeur de 'tidyverse':

    * ggplot2 (visualisation)
    * dplyr (manipulation des données)
    * tidyr (remise en forme des données)
    * purrr (programmation)
    * readr (importation de données)
    * tibble (tableaux de données)

  - le package caret : classification and regression trainning

TP1 : Introduction à la classification supervisée

Exercice 1

1. **Simulation d'un mélange Gaussian à 2 classes dans $\mathbb{R}^2$**

   Soit $\mathbb{X} \subseteq \mathbb{R}^2$ l'espace d'entré et $\mathbb{Y} = \{C_1, C_2\}$ l'ensemble de classes. Simuler des données dans $\mathbb{R}^2$ issues d'un mélange de 2 vecteurs gaussien. Pour générer ces données, il suffit de simuler une variable Y qui prend par exemple les valeurs 1 ou 2 avec probabilité *a priori* $p_1 = \mathbb{P}(C_1)$ et $p_2 = \mathbb{P}(C_2)$ respectivement. Puis, conditionnellement a la valeur de Y simuler un vecteur X qui suit une loi normale en dimension 2 (i.e. si $Y = j$, $X \sim N(\mu_j, \Sigma_j)$ pour $j = 1, 2$).

   (a) Générer des données en 2D à partir d'un mélange Gaussien à deux classes et créer une liste de vecteurs ou *data frame*. Commencer pour écrire dans un script ou autre le code suivante :

   ```
   n <- 550; p1 <- 0.4; p2 <- 0.6; mu1 <- c(0.1,2); mu2 <- c(1.5,0.5)
   Sigma1 <- diag(c(0.4,0.3)); Sigma2 <- diag(c(0.1,0.5))
   y <- sample(c(1,2),size=n,prob=c(p1,p2),replace=TRUE)
   y <- sort(y)
   n1 <- sum(y==1);n2 <- sum(y==2)

   library(mvtnorm)
   x <- rbind(rmvnorm(n1,mu1,Sigma1),rmvnorm(n2,mu2,Sigma2))
   D <- cbind(x,y)
   colnames(D)=c("x1","x2","y")
   D <- data.frame(D)
   ```

   Que font les fonctions `sample` et `rmvnorm` ?

   (b) Tracer les deux ensembles de points sur un même graphique.

2. **Données réelles, Iris [Edgar Anderson]**

   Les données Iris, stockées sous `R`, sont très célèbres et souvent utilisées dans la littérature de Data-Mining. Les variables considérérées sont mesurées sur 150 fleurs (trois espèces d'iris : iris setosa, iris versicolor et iris virginica) :

   - largeur du sépale (Sepal.Width)
   - longueur du pétale (Petal.Length)
   - largeur du pétale (Petal.Width)
   - longueur du sépale (Sepal.Length)

   ```
   data(iris); attach(iris)
   names(iris)
   summary(iris)
   ```

(a) Pour afficher les données, taper sur la console de `R` la commande `iris`.

(b) Considérer les longueurs et la largeurs des pétales. Tracer les trois ensembles de points sur un même graphique.

(c) Considérer les longueurs et la largeurs des pétales et sépales pour deux types d'iris : versicolor et iris virginica. Tracer les données.