TP 3: Arbre de classification CART

Exercice 1: Données synthétiques

Récupérer le jeu de données d'apprentissage synth_train.txt . On a Y *dans* $\{1, 2\}$et X *dans* R^2. On dispose de 100 données d'apprentissage.

1. Charger le jeu de données dans R. Transformer la variable de sortie y en facteur.

```
rm(list=ls())
data<-    read.table(file="../../data/synth_train.txt",header=TRUE)
dim(data)

## [1] 100      3

data$y<-    as.factor(data$y)
```

2. Charger le package rpart (**rpart** = **r**ecursive **part**itioning) et consulter l'aide de la fonction rpart.

```
library(rpart)
help(rpart)
```

3. Construire un arbre de classification `t` à l'aide de la fonction `rpart` (attention la fonction demande une formule du style y ~ et des données data=data_synth). Faire afficher l'arbre en tapant `t`, tracer l'arbre à l'aide des fonctions `plot(t)` puis `text(t)`.

4. Calculer l'erreur d'apprentissage du prédicteur obtenu.

5. Charger le jeu de données test puis calculer le taux d'erreur test.

6. Faire de même avec l'arbre maximal (on pourra regarder la fonction `rpart.control` qui permet de régler les règles de construction d'un arbre).

Exercice 2 : Reconnaissance automatique de caractères manuscrits

Récupérer le jeux de données d'apprentissage zip_train.txt. L'objectif est d'identifier dans une image noir et blanc de 16 × 16 pixels un chiffre de 0 à 9. A chaque pixel d'une image est associé un nombre réel entre −1 (noir) et 1 (blanc). On a $Y \in \{0, \ldots, 9\}$ et $X \in [-1, 1]^{256}$.

7. Charger les jeu de données zip_train.txt et zip_test.txt.
8. Calculer les taux d'erreur empirique et taux d'erreur test d'un arbre CART construit sur ce jeu de données.