Génie Mathématique Année 2018-2019

Statistique GM5

Régression Non Linéaire avec Application sous R

TP8 – CART et Forêts Aléatoires

L'objet de ce TP est d'expérimenter, à l'aide des librairies `rpart` et `randomForest`, quelques techniques statistiques dans le cadre des modèles de régression par arbres CART et des Forêts Aléatoires. Le problème est issu d'un contrat de collaboration de recherche entre Air Normand, le laboratoire de mathématiques de l'INSA de Rouen et le laboratoire de mathématiques de l'université d'Orsay.

Il s'agit d'expliquer la pollution par les particules fines PM10 à l'aide de données météorologiques et de mesures de polluants primaires. On s'intéresse à une station de mesures d'Air Normand située dans l'agglomération du Havre sur la période 2004-2006. Il s'agit de la station HRI, station de fond urbain. Les données sont de deux types : des concentrations moyennes journalières de polluants (PM10, NO, NO2 et SO2) et des mesures de paramètres météorologiques (température, pression atmosphérique, humidité relative, vitesse et direction de vent, gradient température). Plus précisément, la variable à expliquer est la concentration moyenne journalière en PM10. Les variables explicatives sont :

- des polluants : NO, NO2, SO2
- des variables météorologiques : température (T.min, T.moy et T.max), vitesse de vent (VV.moy et VV.max), humidité relative (HR.min, HR.moy et HR.max), pression atmosphérique (PA.moy), direction de vent (DV.maxvv et DV.dom), gradient de température (GTrouen et GTlehavre), quantité de pluie (PL.som).

1. Récupérer les données à l'aide de la commande

```
library(rpart)
library(randomForest)
source("Perfopm10.R")
source("Tabdeppm10.R")
source("Fig_obspm10.R")
Data<-read.table("http://lmi2.insa-
rouen.fr/~bportier/Data/Data_HRI.txt",
header=TRUE,sep=";")
summary(Data)
```

et faire connaissance avec celles-ci.

2. Après avoir éliminé les données manquantes à l'aide de la commande

```
# Elimination des donnees
pm10data = na.omit(Data)
```

constituer, en procédant comme dans le TP6, un échantillon d'apprentissage et un échantillon test.

3. Construire l'arbre de régression CART maximal sur l'échantillon d'apprentissage. Commenter.

```
# Construction de l'arbre maximal
modcart <- rpart(formula(pm10data),data = pm10data[appri,])
summary(modcart)
print(modcart)
plot(modcart, branch = 0.3, uniform = T)
text(modcart, digit = 4,col=2)
title("Modélisation des PM10")
```

4. Évaluer les performances du modèle en estimation sur l'échantillon d'apprentissage et en prévision sur l'échantillon Test. On pourra utiliser pour cela les fonctions `Perfopm10`, `Fig_obspm10` et `Tabdeppm10` fournies dans le mail.

```
pm10est = predict(modcart)
Perfopm10(pm10data$PM10[appri],pm10est)
TabDeppm10(pm10data$PM10[appri],pm10est,30,50,30)
Titre = paste("Station HRI - Arbre maximal","Echantillon
d'apprentissage",
        sep="\n")
Fig_obspm10(pm10data$PM10[appri],pm10est,Titre,"Essai")
```

5. Construire l'arbre CART élagué et évaluer ses performances en estimation et en prévision sur l'échantillon Test. Comparer les résultats obtenus avec ceux obtenus pour l'arbre maximal.

6. Étudier à l'aide des forêts aléatoires l'importance des différentes variables explicatives et leur effet marginal sur les PM10. Commenter.

7. Étudier les performances en estimation sur l'échantillon d'apprentissage et en prévision sur l'échantillon test, du modèle issu de la forêt construite. Commenter les résultats obtenus. On pourra notamment comparer les résultats avec ceux que donnerait une modélisation linéaire des données avec un sous-ensemble de variables explicatives bien choisies.