EISTI

TP ACP

Machine Learning

BOGALHO Jessy
24/01/2017

Initialisation/recuperation des donnees
Dataset=read.table("FrenchCities.csv", header=T,sep=';',row.names=1)

Liste des attributs
attributes(Dataset)

Affiche une colonne (ici la 1ere) à partir de son attribut
Dataset$NO2
Affiche une colonne (ici la 1ere) à partir de son numero
Dataset[,1]
Affiche une colonne (ici la 1ere) à partir de son nom
Dataset[,"NO2"]

Selection de plusieurs lignes (10, 11 et 12)
Dataset[10:12,]

Affiche les lignes repondant à une expression logique (ici N02>50)
subset(Dataset,NO2>50)

Données sans les 5 premieres lignes
Dataset[-(1:5),]

Enregistre toutes les données de Dataset dans un fichier (ici "name.file")
write.table(Dataset,file='name.file',sep=';',row.names=T,col.names=T)

Type de chaque variable (int, num, ...)
str(Dataset)

Transforme la variable GEO en facteur (au lieu de integer)
Dataset$GEO=as.factor(Dataset$GEO)

Renvoi vrai si la variable (GEO) correspond au type (factor), faux sinon
is.factor(Dataset$GEO)

Pour les facteurs, renvoi un vector de chaine de caracteres avec les differentes valeurs prises par le facteur
levels(Dataset$GEO)

Modifie les valeurs prises par le facteurs
levels(Dataset$GEO)=c('South','East','West','North')

Donnes des indicateurs numériques (médiane, moyenne, quartiles,...) pour les variables quantitatives; donne la frequence des variables qualitatives
summary(Dataset)

Valeur moyenne de la variable NO2
mean(Dataset$NO2)

Variance de la variable NO2
sd(Dataset$NO2)

Mediane de NO2
median(Dataset$NO2)

Quartiles de NO2
quantile(Dataset$NO2)

Matrice des corrélations des 5 premieres variables
cor(Dataset[,1:5])

Boite de Tuckey avec variable TEMP
boxplot(Dataset$TEMP,main='Boxplot of TEMP',ylab='TEMP',col='grey')

Histogramme avec TEMP
hist(Dataset$TEMP,main='Histogram of TEMP',xlab='TEMP',col='blue')

Graphique 2D montrant la latitude en fonction de la temperature
plot(Dataset$LATITUDE,Dataset$TEMP,main='TEMP vs
LATITUDE',xlab='LATITUDE',ylab='TEMP',col="red",lwd=2)

Affiche le nom de chaque point
text(Dataset$LATITUDE,Dataset$TEMP,row.names(Dataset))

Nuages de dispersion pour les 5 premieres variables
pairs(Dataset[,1:5])

Boite de Tuckey de la variable NO2
boxplot(Dataset$NO2)

Supprime la/les lignes ayant des valeurs de TEMP <= 0
Dataset=subset(Dataset,TEMP>0)

Tableau avec le nombre d elements concernés par chaque valeur du facteur GEO (tableau de contingence)
tab=table(Dataset$GEO)

Idem mais sous forme de camembert
pie(tab)

Creation de 2 categories de variables, ici suivant 2 conditions : une sur la colonne 31 et l autre sur la colonne 1
shiny=Dataset[,31]>2000&Dataset[,1]>50

Changement du type de shiny : facteur
shiny=as.factor(shiny)

Tableau de contingence avec colonne GEO, avec 2 categories de variables définies par shiny
tab=table(shiny,Dataset$GEO)

Compilation de la bibliothèque FactoMineR
library(FactoMineR)

PCA : analyse en composantes principales

Renvoi les "eigenvalues" de la matrice de covariance : % de variance et de variance cumulée (% d inertie)
res$eig

% de variance expliquée par les deux axes principaux
barplot(res$eig[,2])

Liste de matrices contenant tous les resultants des variables (coordonnés, corrélations entre variables et axes, contributions...)
res$var

Affiche un cercle de correlation entre les axes 1 et 2
plot.PCA(res,choix="var",axes=c(1,2))

Suppression de MONTHr et MONTHdr des estimations, on les ajoute en variables supplementaires, ce qui change les axes principaux
res1=PCA(Dataset.PCA , scale.unit=TRUE, ncp=5, quanti.sup=c(3:14,16:27), graph = FALSE)

Affichage des composantes principales sur les axes 1 et 2
plot.PCA(res,choix="ind",axes=c(1,2))
plot.PCA(res,choix="var",axes=c(1,2))

Suppression de Biarritz (outlier potentiel), ajout comme point supplémentaire
res2<-PCA(Dataset.PCA, scale.unit=TRUE, ncp=5, ind.sup=5,graph = FALSE)

scale : centre et reduit les variables
dist : matrice des distances
hclut(dist) : (clustering) hierarchique (CAH - classification hiérarchique ascendante par exemple)
kmeans(dataset) : clustering k-means
cutree : creation d'un vecteur avec les clusters