TP1: Analyse statistique + Prétraitement

Dr TSOPZE

Données

- Nature: octroi de credit
- Nombre d'exemples : 6906.
- Nombre d'attributs : 15 + 1
- Présence des valeurs manquantes :
  - Notées: ?
  - A1, A2, A4, A5, A6, A7, A14
- Noms des attributs: A_i

Données

- **Var = read.table(chemin, header=F/T, dec=symbol, col.names=c(…), na.strings=c(‘cc'), sep=‘c’…)**

  cc → chaîne qui représente la valeur manquante

  Exemple:

  cr<- read.table(‘’crx.data.txt”,header=F, dec='.', na.strings=c(‘?'), sep=‘,’)

Données

- rownames(cr) = c(liste des noms de lignes)
- colnames(cr) =c(liste des noms de colonnes)
  - colnames(cr)=paste("A",1:16,sep='')
- Connaitre les dimensions: dim(cr)

Statistiques descriptive

- summary(var) → description de l'objet var.

Moyenne, médiane, mode,… de chaque attribut

Exemple: summary (cr)

- Accès à un attribut: var$attribut

Exemple: cr$A10

- Histogramme d'une variable

  hist(var,…)

  hist(cr$A11, prob = T)

  plot(cr$A13,col="blue")

  barplot(cr$A11)

- par() pour passer les paramètres à R

Exemple: par(mfrow=c(1,2)) pour mettre la fenêtre graphique sur une ligne et deux colonnes

qq.plot() trace le graphique de chaque variable et ses quartiles.

- boxplot(var) permet aussi de visualiser la distribution d'une variable.
- Brève description de quelques propriétés de l'attribut:
  - Les deux traits horizontaux de la boite sont le 1er et le 3^{e} quartiles
  - Le trait fort est la médiane
- abline() : tracer le trait horizontal représentant la moyenne

Outliers

- Comportement « étrange »
- Exemple: Analyse de l'attribut A3
  - plot(cr$A3, xlab = "")
  - abline(h = mean(cr$A3, na.rm = T), lty = 1)
  - abline(h = mean(cr$A3, na.rm = T) + sd(cr$A3, na.rm = T), lty = 2)
  - abline(h = median(cr$A3, na.rm = T), lty = 3)
  - identify(cr$A3)
- Identify () permet de sélectionner les points sur le graphique, on peut l'affecter à un objet puis l'utiliser dans la suite.
  - A=identify(cr$A3)
  - cr[A,]

Outliers

- objet [critère sur l'attribut]
- Pour éviter les valeurs manquantes:

!is.na (attribut)

Exemple: cr[!is.na(cr$A3) & cr$A3 >
…,]

boxplot(var) est limité à une valeur.

- valeurs nominales: pour chacune des valeurs de l'attribut nominal, afficher un box d'un attribut
- library(lattice)
- bwplot(attr_nom ~ autre_attr, ….)

Exemple:

- library(lattice)
- bwplot(A13 ~ A3, data=cr, ylab='un nominal A13',xlab='un numerique A3')

- Voir aussi library(Hmisc)

Valeurs manquantes - suppression

- Library(DMwR)
- Visualisation des lignes ayant des valeurs manquantes: var[!complete.cases(var),]

  Exemple:
- Compter le nombre de ligne: nrow()
  - Exemple: nrow(cr[!complete.cases(cr),])
- Suppression des lignes: na.omit()
  - Exemple: cr_clean=na.omit(cr)

NA – Remplacement

- Utiliser mean(), median(),…
  - Exemple:cr[is.na(cr$A14), ‘’A14"] <- median(cr$A14, na.rm = T)
  - vecteur[is.na(vecteur)]=mean(vecteur[!is.na(vecteur)]
- Utiliser les corrélations: cor()
- Exemple:
  - cor(cr[,14:15],use="complete.obs")
  - symnum(cor(cr[,14:15],use="complete.obs"))
- centralImputation(var): médiane (attribut numérique) ou la valeur le plus fréquente (attribut nominal)

NA - régression

- calculer les cœfficients de corrélation:

  $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n$

lm(attr ~ attr1+attr2+… data = var)

Exemple:lm(A14 ~ A15, data = cr)

- Utiliser ces coefficients dans la formule et faire une fonction pour remplir les attributs

  Exemple: cr[…] = …

- Utiliser les similarités:

  cr <- knnImputation(cr, k = 10)

Transformation

- Fonction **transform** (df, attr=expression): appliquer une fonction le calcul 'expression' à l'attribut 'attr' du data frame 'df' et ajouter cette nouvelle colonne à 'df'
- Fonction ***ifelse*** *(test= test_logique, yes=Valeur_vrai, no= valeur_Faux)* où 'test_logique' est un vecteur de booléen, retourne le composant i du vecteur 'Valeur_vrai' si le composant i dans 'test_logique' est TRUE et celui du vecteur 'Valeur_Faux' sinon;

  'Valeur_Vrai' et 'Valeur-Faux' peuvent être des expressions de calcul
- Fonction **cut**(vecteur_numérique, breaks=c(liste des points de coupe))
- Fonction **scale**(X): normalise et centre les éléments de X

Chaine de caractères

- Fonction **paste** : concaténer des éléments de vecteurs
- fonction **nchar** : compter le nombre de caractères dans des chaînes de caractères
- Fonctions **toupper** et **tolower** : transformer tous les caractères alphabétiques en majuscules et en minuscules respectivement
- fonction **strsplit (ch, split='sous_ch')** : séparer la chaîne de caractères 'ch' en sous-chaînes de caractères en coupant lors de la rencontre 'sous_ch'
- fonction **substr** : extraire une partie de chaînes de caractères en spécifiant les positions, dans les chaînes
- fonctions **sub** et **gsub** : chercher les occurrences d'un « motif » (argument pattern) dans des chaînes de caractères (argument x), et remplacer ces occurrences par un autre motif (argument replacement)

Chaine de caractères

- fonction **chartr**(old=x, new=y, x=ch) : remplacer dans la chaine ‘x’ les caractères de ‘old’ par ceux de ‘new’, en respectant l’ordre des caractères dans le premier vecteur
- fonction iconv (ch) : retirer des accents aux caractères de ch

Dates

- fonction **Sys.getlocale**(category = "LC_TIME") : géolocalise
- Fonction **Sys.setlocale**("LC_TIME", locale = « …") pour changer les paramètres locaux
- Fonction **as.Date**(x) : convertir la chaine x en une date
- fonction **Sys.Date** () : date courante
- fonction **difftime** (debut, fin) : calculer cette différence en jours, en semaines, en secondes, minutes ou heures si les dates comprennent aussi une heure
- Utiliser l'aide sur la fonction **strptime** pour connaitre les différents formats de date