105-1: EE4052

計算機程式設計

Computer Programming

Unit 12: 資料前處理

連 豐 力

臺 大 電 機 系

Sep 2016 - Jan 2017

資料庫

- 某個產品的銷售情況：
  - 臺北地區的銷售記錄，存在一個星期的空白
  - 臺中地區的銷售記錄，某天的數據是負的
  - 某個時段，高雄地區的銷售量遠小於屏東地區的銷售量
- 這些資訊，明顯與實際的情況不符，
- 因此，為了獲得準確的分析報告，
- 必須對這些不符合常理的情況進行處理。
- 一般的商務或日常實作之中，所汲取的資料通常是：
  - 不完整（缺少某部分）
  - 含有雜訊（錯誤或偏離期望）
  - 不一致（不同單位，不同編碼）
- 因此，通常需要進行所謂的前置處理，
  剔除雜訊，恢復資料完整性或一致性，
  才能進行下一步的分析

大綱

- 資料庫載入
- 資料遺漏值處理 - 刪除與插補
- 雜訊資料處理
- 資料轉換

計算機程式設計 – 2016F
Chap 12: 資料前處理
Feng-Li Lian @ NTU-EE

大綱

資料庫載入

軟體套件 與 資料庫

- install.packages("lattice") # 安裝 lattice 軟體套件
- install.packages("MASS") # 安裝 MASS 軟體套件
- install.packages("nnet") # 安裝 nnet 軟體套件
- library(lattice) # 載入 lattice 軟體套件
- library(MASS) # 載入 MASS 軟體套件
- library(nnet) # 載入 nnet 軟體套件

- install.packages("mice") # 安裝 mice 軟體套件
- library(mice) # 載入 mice 軟體套件
- data(nhanes2) # 取得 nhanes2 資料集

軟體套件 與 資料庫

- nrow(nhanes2) # nhanes2 資料集的橫列數
- ncol(nhanes2) # nhanes2 資料集的直行數
- summary(nhanes2) # nhanes2 資料集的概括資訊
- head(nhanes2)

```
> head( nhanes2 )
    age  bmi  hyp chl
1 20-39   NA <NA>  NA
2 40-59 22.7   no 187
3 20-39   NA   no 187
4 60-99   NA <NA>  NA
5 20-39 20.4   no 113
6 60-99   NA <NA> 184
```

```
> summary( nhanes2 )
    age          bmi          hyp          chl
 20-39:12   Min.   :20.40   no  :13   Min.   :113.0
 40-59: 7   1st Qu.:22.65   yes : 4   1st Qu.:185.0
 60-99: 6   Median :26.75   NA's: 8   Median :187.0
            Mean   :26.56             Mean   :191.4
            3rd Qu.:28.93             3rd Qu.:212.0
            Max.   :35.30             Max.   :284.0
            NA's   :9                 NA's   :10
```

- age: 年齡段，定性變數，3大類，沒有遺漏值
- hyp: 是否高血壓，定性變數，2大類，有 8 個遺漏值
- bmi: 身體品質指數 (kg/m^2)，定量變數，有 9 個遺漏值
- chl: 血清膽固醇總量 (mg/dL)，定量變數，有 10 個遺漏值

大綱

遺漏值處理 - 刪除 與 插補

遺漏值處理

- is.na(nhanes2) # 有遺漏值的數據列表
- sum(is.na(nhanes2)) # 有遺漏值的數據總數
- sum(complete.cases(nhanes2)) # 完整樣本的數量
- md.pattern(nhanes2) # 觀測遺漏值的情況

```
> md.pattern( nhanes2 )

   age hyp bmi chl
13   1   1   1   1  0
 1   1   1   0   1  1
 3   1   1   1   0  1
 1   1   0   0   1  2
 7   1   0   0   0  3
     0   8   9  10 27
```

遺漏值處理

- 處理缺失資料的方法：
  - 直接刪除
  - 用平均值或中位數取代
  - 多重補差法（利用變數間關係進行預測取代值）

```
> md.pattern( nhanes2 )
   age hyp bmi chl
13   1   1   1   1  0
 1   1   1   0   1  1
 3   1   1   1   0  1
 1   1   0   0   1  2
 7   1   0   0   0  3
     0   8   9  10 27
```

- imp <- mice(nhanes2, m = 4)　　# 產生四組完整的資料庫
- fit <- with (imp, lm(chl ~ age + hyp + bmi))　　# 產生回歸模型
- pooled <- pool(fit)　　# 對四組模型進行整理
- summary(pooled)　　# 展示內容

```
> summary( pooled )
                  est       se          t        df    Pr(>|t|)       lo 95     hi 95 nmis       fmi     lambda
(Intercept)  1.623221 61.55743  0.02636922 11.414325 0.979418234 -133.266142 136.51258   NA 0.3327949 0.22530537
age2        57.776554 22.03145  2.62245781  9.423247 0.026700135    8.277097 107.27601   NA 0.4069495 0.29315569
age3        71.438955 22.47045  3.17924050 16.059530 0.005804682   23.818082 119.05983   NA 0.1810950 0.08508939
hyp2       -13.012035 22.00083 -0.59143374  9.787785 0.567637105  -62.177396  36.15333   NA 0.3925400 0.27992001
bmi          5.817507  2.06504  2.81714081 12.230324 0.015295585    1.327552  10.30746    9 0.3049673 0.19990069
```


計算機程式設計 – 2016F
Chap 12: 資料前處理
Feng-Li Lian @ NTU-EE

遺漏值處理

- 刪除法：
  - 刪除觀測樣本
  - 刪除整個變數
  - 使用不同權數進行加權
- 補差法：
  - 平均值補差
  - 回歸補差
  - 二階補差
  - 熱平台補差
  - 冷平台補差
  - 抽樣填補

遺漏值處理

- 隨機抽樣補差法：
- nhanes2[, 4] # 針對第4組數據
- sub <- which(is.na(nhanes2[, 4]) == TRUE)
- dataTR <- nhanes2[-sub,]
- dataTE <- nhanes2[sub,]
- dataTE
- dataTE[, 4] <- sample(dataTR[, 4], length(dataTE[, 4]), replace = T)
- # 在非遺漏值之中，簡單抽樣之後的值，取代之
- dataTE

遺漏值處理

- 平均值補差法：
- nhanes2[, 4] # 針對第4組數據
- sub <- which(is.na(nhanes2[, 4]) == TRUE)
- dataTR <- nhanes2[-sub,]
- dataTE <- nhanes2[sub,]
- dataTE
- dataTE[, 4] <- mean (dataTR[, 4])
- # 用非遺漏值之平均值取代之
- dataTE

遺漏值處理

- 回歸模型預測值補差法：
- nhanes2[, 4]　　# 針對第4組數據
- sub <- which(is.na(nhanes2[, 4]) == TRUE)
- dataTR <- nhanes2[-sub,]
- dataTE <- nhanes2[sub,]
- dataTE
- lmout <- lm(chl ~ age, data = dataTR)
  - # 利用 dataTR 中 age 為引數，chl 為因變數，建構線性回歸模型
- dataTE[, 4] <- round(predict(lmout, dataTE))
  - # 用回歸模型預測值取代之
- dataTE

遺漏值處理

- 熱平台補差法：在非遺漏值資料中，找到一個與遺漏值所在樣本相似的樣本
- accept <- nhanes2[which(apply(is.na(nhanes2), 1, sum) != 0),]
  - # 存在遺漏值的樣本
- donate <- nhanes2[which(apply(is.na(nhanes2), 1, sum) == 0),]
  - # 無遺漏值的樣本
- accept[1,]
- donate[1,]
- sa <- donate[which(donate[, 1] == accept[2, 1] & donate[, 3] == accept[2, 3] & accept[2, 4]),]
  - # 找尋與 accept 中第2個樣本相符的樣本

遺漏值處理

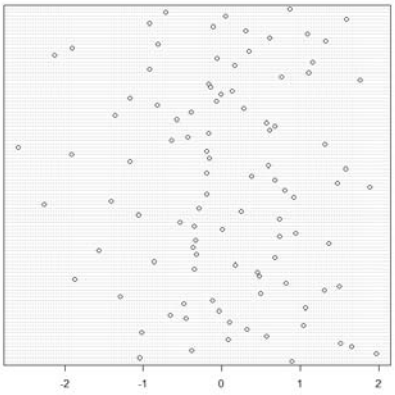
- 冷平台補差法：將資料分層，在層中對遺漏值使用平均值取代
- levelhyp <- nhanes2[which(nhanes2[, 3] == "yes"),]
  - # 按照 hyp 分層
- levelhyp
- levelhyp[4, 4] <- mean(levelhyp[1:3, 4])
  - # 用層內平均值代替第4個樣本的遺漏值

大綱

雜訊資料處理

雜訊資料處理

- 雜訊：是量測過程中，隨機錯誤或偏差所獲得的數據。
- 使用 outliers 軟體套套件中的 outlier 函數尋找雜訊資料
- 主要是利用：尋找資料中與其他觀測值及平均值差距最大的點，當作異常值



- install.packages("outliers")
- library(outliers)
- set.seed(1)
- s1 <- .Random.seed
- y <- rnorm(100)
- outlier(y) # 找出最遠離群值
- outlier(y, opposite = T) # 找出最遠離群值相反的值
- dotchart(y)

雜訊資料處理

- 去除雜訊：是採用分群（分箱），回歸，檢查等方法，去平滑化一小群的數據，以去除掉雜訊。
- 分群（分箱）法：
- set.seed(1); s1 <- .Random.seed
- x <- rnorm(12)
- x <- sort(x)
- dim(x) <- c(3, 4)
- x[1,] <- apply(x, 1, mean)[1] # 用第1橫列的平均值代替第1橫列中的資料
- x[2,] <- apply(x, 1, mean)[2] # 用第2橫列的平均值代替第2橫列中的資料
- x[3,] <- apply(x, 1, mean)[3] # 用第3橫列的平均值代替第3橫列中的資料
- x

大綱

資料轉換

資料轉換

- 光滑：去掉資料中的雜訊，可以透過分箱、回歸、或分群等技術實現
- 屬性建構：由指定的屬性建構出新屬性，並增加到資料集中。
  例如：透過『銷售額』和『成本』建構出『利潤 』，
  只需要對對應屬性資料進行簡單轉換。
- 聚集：對資料進行整理。
  例如：可以透過『日銷售額』資料，計算『月』和『年』的銷售資料。
- 規範化：把資料按照某種比例縮放，實質落入一個特定的小區間。
  例如：-1.0 ~ 1.0 或 0.0 ~ 1.0。
  常態分布之標準化是常見的方法。

資料轉換

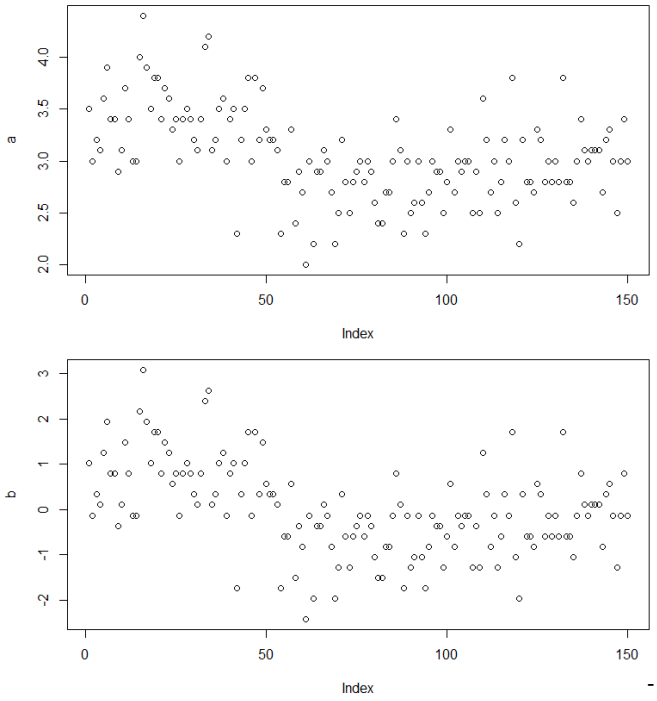
- 規範化：把資料按照某種比例縮放，實質落入一個特定的小區間。

  例如：-1.0 ~ 1.0 或 0.0 ~ 1.0。

  常態分布之標準化是常見的方法。

- a <- iris[,2]
- plot(a)
- b <- scale(a)
- plot(b)
- # 對該數據標準化

```
attr(,"scaled:center")
[1] 3.057333
attr(,"scaled:scale")
[1] 0.4358663
```



資料轉換

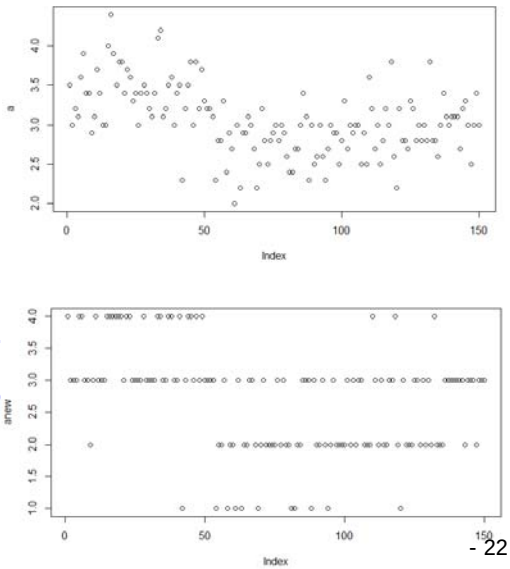
- 離散化：把數值屬性（例如：年齡）的原始值，

  用區間標籤（例如：0~10，11~20 等），

  或者是概念標籤（例如：youth, adult, senior）取代。

  可以實現將定量資料向定性資料轉化，將連續類型資料離散化。

- a <- iris[,2]; plot(a)
- n <- length(a)
- anew <- rep(0, n)
- which(a < 2.5)
- anew[which(a < 2.5)] <- 1
- anew[which(a >= 2.5 & a < 3.0)] <- 2
- anew[which(a >= 3.0 & a < 3.5)] <- 3
- anew[which(a >= 3.5)] <- 4
- plot(anew)



資料轉換

- 由額定資料產生概念分層：

屬性（例如：street）可以泛化到較高的概念層（例如：city, country 等）。

或者是概念標籤（例如：youth, adult, senior）取代。

可以實現將定量資料向定性資料轉化，將連續類型資料離散化。

- 資料泛化可以視為資料合併，
- 以城市為例，1 = 臺北，2 = 臺中，3 = 高雄，等等，
- 可以透過資料合併，
- 將 1, 2, 3 等合併為大城市，6, 7, 8 等等合併為中城市。

```
city <- c( 6, 7, 2, 3, 1, 5, 4, 2, 8, 9, 2, 3, 8, 1, 2, 8, 8, 6 )
citytype <- rep( 0, 18 )
citytype[ which( city <= 5 ) ] <- 1
citytype[ which( city >= 6 ) ] <- 2
```