大數據資料處理實務

李水彬

2023-09-01

Chapter 03 過濾、異常判定和處理

課程內容

- 重複的資料 duplicate data
- 遺漏值 missing value
- 異常判定和處理

重複的資料 duplicate data

- 重複資料是指多次紀錄同一事件，若是多次事件的結果相同，所記錄的資料就不能視為重複。

```
TaoW<-fread("桃園天氣2023.csv",header="auto")
```

重複的資料 duplicate data

- 2023/2/2 和2023/2/4 有資料重複了

date	week	temperature	humidity
2023/2/1	3	21.2	0.44
2023/2/2	4	15.6	0.90
2023/2/2	4	15.6	0.90
2023/2/3	5	18.4	0.74
2023/2/4	6	18.5	0.80
2023/2/4	6	18.5	0.80
2023/2/5	7	17.2	0.99
2023/2/6	1	14.4	0.99

資料重複的原因與排除

- 在紀錄事件時，為了避免遺漏常有重複性廣播機制，而這樣的機制往往是產生重複資料的原因。
- 紀錄資料的過程中，應該要有避免紀錄重複資料。例如，唯一性身分確認，排除已經存在的事件資料。
- 將暫存(或工作端)的資料匯入資料庫時，應該再次檢視匯入資料的是否有重複，或重複卻不明確的資料。

遺漏值(missing value)

遺漏值就其產生的原因，並無單一的定義。

- 有數值，但沒被記錄到：一種常發生在自動記錄機制的遺漏值，感應器偵測到環境訊息傳遞給終端設備時，因為感應器本身或傳輸網路問題造成資料無法傳達而遺漏。

遺漏値(missing value)

```
sushi<-fread("sushi.csv",header="auto")
```

```
head(sushi[29:35,])
```

date	week	dish	revenue	adj-revenue	weather	promotion	holiday	temperature	high-temperature	tempera
20140329	6	2111	63340	56440	rainy	1	1	24.6	27.7	
20140330	7	2650	79500	65000	cloudy	1	1	22.6	25.7	
20140331	1	1507	45200	42500	rainy	1	0	20.8	24.6	
20140401	2	861	25830	20530	cloudy	1	0	17.9	19.1	
20140402	3	1729	51865	47265	cloudy	1	0	18.7	21.0	
20140403	4	1511	45340	37840	rainy	1	0	18.8	20.9	

遺漏值(missing value)

- 沒有數值，故沒記錄：此常發生在正規型態的資料，每個被記錄的事件(顧客點餐資訊、每日氣候觀測、學生資料或問卷調查的受訪者)有設定應紀錄的變數，而因原有紀錄規範不宜造成無法得到觀測紀錄。

遺漏值(missing value)

```
d<-cbind(Customer = c("A001","A002","A003"),
         Gender=c("Female","", "Male"),
         Income=c(345,450,500))

head(d)
```

Customer	Gender	Income
A001	Female	345
A002		450
A003	Male	500

Customer A002 的 gender (性別) 是 空白。

異常判定和處理

資料異常分成兩種:

- 遺漏值：可以邏輯判斷找出來。

使用 == 邏輯運算。

```
d # 顧客資訊
```

```
##      Customer Gender   Income
## [1,] "A001"   "Female" "345"
## [2,] "A002"   ""       "450"
## [3,] "A003"   "Male"   "500"
```

```
d[,2]=="" #判定是否為空值
```

```
## [1] FALSE  TRUE FALSE
```

異常判定和處理

使用 is.na() 函數。

```
sushi[29:35,c(1,12)] # sushi濕度
```

```
##        date humidity
## 1: 20140329       86
## 2: 20140330       65
## 3: 20140331       NA
## 4: 20140401       83
## 5: 20140402       81
## 6: 20140403       86
## 7: 20140404       67
```

```
is.na(sushi[29:35,12]) #判定是否為空值
```

```
##      humidity
## [1,]    FALSE
## [2,]    FALSE
## [3,]     TRUE
## [4,]    FALSE
## [5,]    FALSE
## [6,]    FALSE
## [7,]    FALSE
```

異常判定和處理

資料異常分成兩種:

- 異常值：有紀錄(非遺漏)，但異於常值。異於常值並非錯誤，但也有可能是錯誤的。
- 資料無誤：因為氣候異常，造成雨量超過歷史紀錄。

  +一天狂降8個月雨量！ 利比亞古城水壩潰決兩千死 https://youtu.be/IBD-Ue4k4VM

  +海葵雨量破1000毫米超驚人 https://tw.nextapple.com/life/20230905/1C0B94AAE367FDDB07A90E0793BA34C2
- 資料有誤：

```
# 本系學生的身高
Height<-c(190, 160, 175, 180, 158, 172)
```

```
# 本系學生的身高
Height<-c(190, 160, 175, 80, 158, 172)
```


第四筆身高為 80 顯然不合理。

課堂練習5

找出 MealRecord2023(Assignment).csv 有哪幾天的遺漏氣溫資料?