Excercise_6_Data_Mining

Duncan Ferguson

10/20/2021

```
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse
1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------
tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Importing Table

```
df <- read.table("Table_8_1.csv", header=TRUE, sep=",")
df

##    RID         Age income student credit_rating Class..buys_computer
## 1    1       youth   high      no          fair                   no
## 2    2       youth   high      no     excellent                   no
## 3    3 middle_aged   high      no          fair                  yes
## 4    4      senior medium      no          fair                  yes
## 5    5      senior    low     yes          fair                  yes
## 6    6      senior    low     yes     excellent                   no
## 7    7 middle_aged    low     yes     excellent                  yes
## 8    8       youth medium      no          fair                   no
## 9    9       youth    low     yes          fair                  yes
## 10  10      senior medium     yes          fair                  yes
## 11  11       youth medium     yes     excellent                  yes
## 12  12 middle_aged medium      no     excellent                  yes
## 13  13 middle_aged   high     yes          fair                  yes
## 14  14      senior medium      no     excellent                   no

subset <- df[df$Age =="youth",]
subset

##    RID   Age income student credit_rating Class..buys_computer
## 1    1 youth   high      no          fair                   no
## 2    2 youth   high      no     excellent                   no
## 8    8 youth medium      no          fair                   no
```

```
## 9    9 youth    low     yes          fair                    yes
## 11  11 youth medium     yes     excellent                    yes
```

$$Gain(Income) = Info(D) - Info_{predictor}(D)$$

$$Info_{predictor}(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

$$Info(D) = -\sum_{i=1}^{m} p_i \, log_2 p_i$$

$$Info(D) = -[\frac{3}{5} log_2(\frac{3}{5})] - [\frac{2}{5} log_2(\frac{2}{5})] = 0.9709506$$

```
Info_i_D <- -((3/5)*log2(3/5))-((2/5)*log2(2/5))
Info_i_D

## [1] 0.9709506
```

Classifying Credit # Splitting on Credit Rating

```
subset_credit_e <- subset[subset$credit_rating =="excellent",]
subset_credit_f <- subset[subset$credit_rating =="fair",]
subset_credit_e

##    RID   Age income student credit_rating Class..buys_computer
## 2    2 youth   high      no     excellent                   no
## 11  11 youth medium     yes     excellent                  yes

subset_credit_f

##   RID   Age income student credit_rating Class..buys_computer
## 1   1 youth   high      no          fair                   no
## 8   8 youth medium      no          fair                   no
## 9   9 youth    low     yes          fair                  yes
```

$$Info_{Credit}(D) = \frac{2}{5} Info_{Credit=Excellent} + \frac{3}{5} Info_{Credit=Fair} = 0.9509775$$

$$Info_{Credit=Fair} = -\frac{1}{3} log_2(\frac{1}{3}) - \frac{2}{3} log_2(\frac{2}{3}) = 0.9182958$$

$$Info_{Credit=Excelent} = -\frac{1}{2} log_2(\frac{1}{2}) - \frac{1}{2} log_2(\frac{1}{2}) = 1$$

```
Info_Credit_Fair <- -((1/3)*log2(1/3))-((2/3)*log2(2/3))
Info_Credit_Fair

## [1] 0.9182958
```

```
Info_Credit_Excellent <- -((1/2)*log2(1/2))-((1/2)*log2(1/2))
Info_Credit_Excellent

## [1] 1

Info_Credit <- ((2/5)*Info_Credit_Excellent) + ((3/5)*Info_Credit_Fair)
Info_Credit

## [1] 0.9509775

Gain_Credit <- Info_i_D-Info_Credit
Gain_Credit

## [1] 0.01997309
```

$$Gain(Credit\ Rating) = Info(D) - Info_{income}(D) = 0.01997309$$

$$Gain(Credit\ Rating) = 0.9709506 - 0.9509775 = 0.01997309$$

Classifying Student # Splitting on Student

```
subset_student_y <- subset[subset$student == "yes",]
subset_student_n <- subset[subset$student == "no",]
subset_student_y

##    RID   Age income student credit_rating Class..buys_computer
## 9    9 youth    low     yes          fair                  yes
## 11  11 youth medium     yes     excellent                  yes

subset_student_n

##   RID   Age income student credit_rating Class..buys_computer
## 1   1 youth   high      no          fair                   no
## 2   2 youth   high      no     excellent                   no
## 8   8 youth medium      no          fair                   no
```

$$Info_{Student}(D) = \frac{2}{5} Info_{Student=Yes} + \frac{3}{5} Info_{Studen=No} = 0$$

$$Info_{Studen=Yes} = -\frac{2}{2} log_2(\frac{2}{2}) - \frac{0}{2} log_2(\frac{0}{2}) = 0$$

$$Info_{Studen=No} = -\frac{0}{3} log_2(\frac{0}{3}) - \frac{3}{3} log_2(\frac{3}{3}) = 0$$

```
Info_Student_Yes <- -((2/2)*log(2/2))-((0/2)*log2(0/2))
Info_Student_Yes

## [1] NaN

Info_Student_No <- -((0/3)*log(0/3))-((3/3)*log2(3/3))
Info_Student_No

## [1] NaN
```

```
Info_Student <- ((2/5)*0)+((3/5)*0)
Info_Student
```

```
## [1] 0
```

```
Gain_Student <- Info_i_D - Info_Student
Gain_Student
```

```
## [1] 0.9709506
```

$$Gain(Student) = Info(D) - Info_{Student}(D) = 0.9709506$$

$$Gain(Student) = 0.9709506 - 0 = 0.9709506$$

Classifying Income

Splitting on Income

```
subset_income_high <- subset[subset$income == "high",]
subset_income_medium <- subset[subset$income == "medium",]
subset_income_low <- subset[subset$income == "low",]
subset_income_low
```

```
##   RID   Age income student credit_rating Class..buys_computer
## 9   9 youth    low     yes          fair                  yes
```

```
subset_income_medium
```

```
##    RID   Age income student credit_rating Class..buys_computer
## 8    8 youth medium      no          fair                   no
## 11  11 youth medium     yes     excellent                  yes
```

```
subset_income_high
```

```
##   RID   Age income student credit_rating Class..buys_computer
## 1   1 youth   high      no          fair                   no
## 2   2 youth   high      no     excellent                   no
```

$$Info_{Income}(D) = \frac{1}{5}Info_{Income=Low} + \frac{2}{5}Info_{Income=Medium} + \frac{2}{5}Info_{Income=High} = .4$$

$$Info_{Income=Low} = -\frac{1}{1}log_2(\frac{1}{1}) - \frac{0}{1}log_2(\frac{0}{1}) = 0$$

$$Info_{Income=Medium} = -\frac{1}{2}log_2(\frac{1}{2}) - \frac{1}{2}log_2(\frac{1}{2}) = 1$$

$$Info_{Income=High} = -\frac{0}{2}log_2(\frac{0}{2}) - \frac{2}{2}log_2(\frac{2}{2}) = 0$$

```
Info_Income_Low <- -(1/1)*log2(1/1) - 0
Info_Income_Low
```

```
## [1] 0
```

```
Info_Income_Medium <- -((1/2)*log2(1/2))-((1/2)*log2(1/2))
Info_Income_Medium
```

```
## [1] 1
```

```
Info_Income_High <- 0-((2/2)*log2(2/2))
Info_Income_High
```

```
## [1] 0
```

```
Info_Income <- ((1/5)*Info_Income_Low) + ((2/5)*Info_Income_Medium) +
((2/5)*Info_Income_High)
Info_Income
```

```
## [1] 0.4
```

```
Gain_Income <- Info_i_D - Info_Income
Gain_Income
```

```
## [1] 0.5709506
```

$$Gain(Income) = Info(D) - Info_{Income}(D) = 0.5709506$$

$$Gain(Student) = 0.9709506 - .4 = 0.5709506$$