Assignment 3 Naive Bayes

wliu16

2022-10-19

```
#select, rename and clean the data
bank_data <- bank.df[, c(10, 13, 14)]
bank_data$Online <- as.factor(bank_data$Online)
bank_data$CC <- as.factor(bank_data$CreditCard)
bank_data$Loan <- as.factor(bank_data$Personal.Loan)
bank_data <- bank_data[, c(-1, -3)]

#Data partition with training: validation = 6:4
set.seed(123)
index_train <- createDataPartition(bank_data$Loan,
                                   p = 0.6,
                                   list = FALSE)
train_data <- bank_data[index_train, ]
validate_data <- bank_data[-index_train, ]
```

A. Pivot table of Online, CC and Loan for training data

```
#molted <- melt(train_data, id = c("Online"))
#tibble(molted)
#x.cast <- dcast(molted, Online~variable, fun.aggregate = length)
#x.cast
```

```
xtab <- ftable(xtabs(~CC+Loan+Online, data = train_data))
xtab
```

```
##          Online    0    1
## CC Loan
## 0  0             791 1144
##    1              79  125
## 1  0             310  467
##    1              33   51
```

B. P(Loan=1 | CC=1, Online=1) = 51/(467+51) = 9.85%

C. Create two separate pivot tables

```
melt.x1 <- melt(train_data, id = "Loan", variable = "Online")
cast.x1 <- dcast(melt.x1, Loan~Online, fun.aggregate = length)
cast.x1
```

```
##   Loan Online   CC
## 1    0   2712 2712
## 2    1    288  288
```

```
ftable(train_data[, c(3,1)]) #Loan as row vs Online as column
```

```
##      Online    0    1
```

```
## Loan
## 0            1101 1611
## 1             112  176
```

```
ftable(train_data[, c(3,2)]) #Loan as row vs CC as column
```

```
##      CC    0    1
## Loan
## 0        1935  777
## 1         204   84
```

D. Compute P(A | B)

```
ftable(train_data[, c(2,3)]) #CC | Loan
```

```
##    Loan    0    1
## CC
## 0        1935  204
## 1         777   84
```

```
ftable(train_data[, c(1,3)]) #Online | Loan
```

```
##        Loan    0    1
## Online
## 0            1101  112
## 1            1611  176
```

```
ftable(train_data[, 3]) #Loan distribution
```

```
##     0    1
##
##  2712  288
```

i. P(CC = 1 | Loan =1) = 84/(204+84) = 0.2917
ii. P(Online =1 | Loan =1) = 176/(176+112) = 0.6111
iii. P(Loan =1) = 288/(288+2712) = 0.096
iv. P(CC=1| Loan =0) = 777/(1935+777) = 0.2865
v. P(Online =1 | Loan =0) = 1611/(1611+1101) = 0.5940
vi. P(Loan = 0) = 2712/(2712+288) = 0.904

E. Naive Bayes probability P(Loan =1 | CC=1, Online =1) = P(CC = 1 | Loan =1)*P(Online =1|Loan =1)*P(Loan =1)/ (P(CC = 1 | Loan =1)*P(Online =1|Loan =1)*P(Loan =1)+ P(CC=1| Loan =0)*P(Online =1 | Loan =0)*P(Loan = 0))= 0.2971*0.6111*0.096/(0.2971*0.6111*0.096+0.2865*0.5940*0.904)= 0.1018 =10.18%

F. Naive Bayes result of 10.18% is higher than pivot table calculation of 9.84%

G. Run NB

```
result <- naiveBayes(Loan~Online+CC, data = train_data)
result
```

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
##
## A-priori probabilities:
## Y
##     0     1
```

```
## 0.904 0.096
##
## Conditional probabilities:
##    Online
## Y           0         1
##   0 0.4059735 0.5940265
##   1 0.3888889 0.6111111
##
##    CC
## Y           0         1
##   0 0.7134956 0.2865044
##   1 0.7083333 0.2916667
```

The entries needed for NB calculation: P(loan) 0.904 and 0.096 P(Loan =0 | Online =1) = 0.5940 P(Loan =1 | Online =1) = 0.6111 P(Loan =0 | CC =1) = 0.2865 P(Loan =1 | CC =1) = 0.2917