Universal Bank

Abhinav Reddy

2024-03-08

This below R code is Loading the required libraries for data manipulation, tidying data and for the Naive Bayes classifier.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(e1071)
```

This below R is reading a CSV file named UniversalBank.csv using read_csv() function and assigns it to the variable **data**.

```
library(readr)
data<- read_csv("C:/Users/Abhinav Reddy/Desktop/FML/Assignmnet 3/UniversalBank.csv")
```

```
## Rows: 5000 Columns: 14
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (14): ID, Age, Experience, Income, ZIP Code, Family, CCAvg, Education, M...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

This below R code is partitioning the data into training (60%) and validation (40%) sets.

```
set.seed(123)
tr_ind <- sample(1:nrow(data), 0.6*nrow(data))
training_set <- data[tr_ind, ]
val_set <- data[-tr_ind, ]
```

Question A : Create a pivot table for the training data with Online as a column variable, CC as a row variable, and Loan as a secondary row variable. The values inside the table should convey the count. In R use functions melt() and cast(), or function table()?

Answer A:

This below R code is creating a pivot table, the training data with Online as a column variable, CC as a row variable, and loan as a secondary row variable and displaying the counts for each combination of these variables.

```
pivotTable_<-pivotTable <- table(training_set$CreditCard, training_set$Online, training_set$"Personal Lo

print(pivotTable_)
```

```
## , ,  = 0
##
##
##        0    1
##   0  785 1145
##   1  317  475
##
## , ,  = 1
##
##
##        0    1
##   0   65  122
##   1   34   57
```

Qustion B: Consider the task of classifying a customer who owns a bank credit card and is actively using online banking services. Looking at the pivot table, what is the probability that this customer will accept the loan offer? [This is the probability of loan acceptance (loan = 1) conditional on having a bank credit card (cc = 1) and being an active user of online banking services (online = 1)]?

Answer B:

This below R code is calculating the probability of a customer accepting the loan offer given that they own a bank credit card (cc = 1) and are actively using online banking services (online = 1), using the counts obtained from the pivot table.

```
count_lco_ <- pivotTable[1, 1, 1]

tcount_co_ <- sum(pivotTable[1, 1, ])

prob_loan_acpt_co <- count_lco_ / tcount_co_

print(prob_loan_acpt_co)
```

```
## [1] 0.9235294
```

Question C: Create two separate pivot tables for the training data?

Answer C:

This below R code is creating 2 pivot tables named as pivotTable_1_ and pivotTable_2_, pivotTable_1_ with Loan as a function of Online and pivotTable_2_ with Loan as a function of cc.

```
pivotTable_1_ <- table(training_set$"Personal Loan", training_set$Online)
print(pivotTable_1_)
```

```
##
##        0    1
##   0 1102 1620
##   1   99  179
```

```
pivotTable_2_<- table(training_set$"Personal Loan", training_set$CreditCard)
print(pivotTable_2_)
```

```
##
##        0    1
##   0 1930  792
##   1  187   91
```

Question D: Compute the following quantities: i. P(cc = 1 | loan = 1) ii. P(Online = 1 | loan = 1) iii. P(loan = 1) iv. P(cc = 1 | loan = 0) v. P(online = 1 | loan = 0) vi. P(loan = 0) ?

Answer D:

```
# (i)P(cc = 1 | loan = 1)
p_C1L1_ <- pivotTable[2, , 2] / sum(pivotTable[, , 2])
p_C1L1_
```

```
##         0         1
## 0.1223022 0.2050360
```

```
# (ii) P(Online = 1 | loan = 1)
p_O1L1_ <- pivotTable[, 2, 2] / sum(pivotTable[, , 2])
p_O1L1_
```

```
##         0         1
## 0.4388489 0.2050360
```

```
# (iii) P(loan = 1)
p_L1_ <- sum(pivotTable[, , 2]) / sum(pivotTable)
p_L1_
```

```
## [1] 0.09266667
```

```r
# (iv) P(cc= 1 | loan = 0)
p_C1L0_ <- pivotTable[2, , 1] / sum(pivotTable[, , 1])
p_C1L0_
```

```
##         0         1
## 0.1164585 0.1745040
```

```r
# (v) P(online = 1 | loan = 0)
p_O1L0_<- pivotTable[, 2, 1] / sum(pivotTable[, , 1])
p_O1L0_
```

```
##         0         1
## 0.4206466 0.1745040
```

```r
# (vi) P(loan = 0)
p_L0_ <- sum(pivotTable[, , 1]) / sum(pivotTable)
p_L0_
```

```
## [1] 0.9073333
```

Question E: Use the quantities computed above to compute the naive Bayes probability P(loan = 1 | cc= 1, online = 1)?

Answer E:

This below R code will print the value of the predicted probability for P(loan = 1 | cc= 1, online = 1).

```r
naivebayes_Model_ <- naiveBayes(training_set$"Personal Loan" ~ Online + CreditCard,
                                data = training_set)

pred_nb_probabilitie_ <- predict(naivebayes_Model_,
                                 newdata = data.frame(Online = 1,CreditCard = 1),type = "raw")

print("Naive Bayes probability P(loan = 1 | cc = 1, online = 1):")
```

```
## [1] "Naive Bayes probability P(loan = 1 | cc = 1, online = 1):"
```

```r
print(pred_nb_probabilitie_[1])
```

```
## [1] 0.8843065
```

Question F: Compare this value with the one obtained from the pivot table in (B). Which is a more accurate estimate?

Answer F:

```r
print("Probability from pivot table (Question B):")
```

```
## [1] "Probability from pivot table (Question B):"
```

```r
print(prob_loan_acpt_co)
```

```
## [1] 0.9235294
```

```r
print("Naive Bayes probability (Question E):")
```

```
## [1] "Naive Bayes probability (Question E):"
```

```r
print(pred_nb_probabilitie_[1])
```

```
## [1] 0.8843065
```

- The probability obtained from the pivot table in Question B is **0.9235294**.
- The probability obtained from the Naive Bayes model in Question E is approximately **0.8843065**.

While the pivot table value seems higher in this specific case, Naive Bayes could be a better general approach for making predictions due to its ability to capture relationships between features.

Question G: Which of the entries in this table are needed for computing P(loan = 1 | cc = 1, online = 1)? Run naive_bayes on data. Examine model output on training data, and find the entry that corresponds to p(loan = 1 | cc = 1, online = 1). Compare this to the number you obtained in (E)?

Answer G:

For computing P(loan = 1 | cc = 1, online = 1), we need following entries from the Naive Bayes model:

- P(loan = 1): It is the overall probability of a person getting a loan, regardless of their credit card or online application status.
- P(cc = 1 | loan = 1): It is the probability of a person having a credit card (cc = 1) given that they were approved for a loan (loan = 1).
- P(online = 1 | loan = 1): It is the probability of a person applying online (online = 1) given that they were approved for a loan (loan = 1).

```r
nb_Model <- naiveBayes(training_set$"Personal Loan" ~ Online + CreditCard,
                        data = training_set)

pred_probabilities_tdata<- predict(nb_Model,
                                   newdata = training_set, type = "raw")

p_loan_given_L1C1O1 <- pred_probabilities_tdata[2, 2]


print("Naive Bayes probability P(Loan = 1 | CC = 1, Online = 1):")
```

```
## [1] "Naive Bayes probability P(Loan = 1 | CC = 1, Online = 1):"
```

```r
cat(p_loan_given_L1C1O1)
```

```
## 0.1156935
```

Comparing this to the number obtained in (E):

- Question E: In Question E, we are directly predicting the probability of loan = 1 given cc = 1 and online = 1 using the Naive Bayes model.
- Question G: In Question G, we are predicting the probabilities for the training data and extracting the probability of loan = 1 given cc = 1 and online = 1 from the model's output.

```r
print("Naive Bayes probability P(loan = 1 | cc = 1, online = 1) from Question E:")
```

```
## [1] "Naive Bayes probability P(loan = 1 | cc = 1, online = 1) from Question E:"
```

```r
print(pred_nb_probabilitie_[1])
```

```
## [1] 0.8843065
```

```r
print("Naive Bayes probability P(loan = 1 | cc = 1, online = 1) from Question G:")
```

```
## [1] "Naive Bayes probability P(loan = 1 | cc = 1, online = 1) from Question G:"
```

```r
cat(p_loan_given_L1C1O1)
```

```
## 0.1156935
```

Thank You!!!