Naive Bayes Classification

Akhil Kornala

2023-11-06

```
knitr::opts_chunk$set(echo = TRUE, comment = NA)
```

```
library(caret)
```

```
Loading required package: ggplot2

Loading required package: lattice
```

```
library(e1071)
library(ISLR)
library(reshape2)
library(readr)
#loading data set required
Universal_Bank <- read_csv("UniversalBank-1.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.
```

Conversion of all the variables into factor

```
Universal_Bank$Personal.Loan<-factor(Universal_Bank$Personal.Loan)
Universal_Bank$Online<-factor(Universal_Bank$Online)
Universal_Bank$CreditCard<-factor(Universal_Bank$CreditCard)
```

Splitting data into two sets for training and validation.

```
set.seed(1237)
training<-createDataPartition(Universal_Bank$Personal.Loan,p=0.6,list = FALSE)
training_setPart<-Universal_Bank[training,]
validation_setPart<-Universal_Bank[-training,]
nrow(training_setPart)
```

```
[1] 3000
```

```r
nrow(validation_setPart)
```

```
[1] 2000
```

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(). In Python, use panda dataframe methods melt() and pivot().

```r
table1<-xtabs(~CreditCard+Personal.Loan+Online,data=training_setPart)
ftable(table1)
```

```
                         Online    0    1
CreditCard Personal.Loan                 
0          0                     783 1133
           1                      82  115
1          0                     319  477
           1                      37   54
```

Question-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)]

```r
46/(46+460)
```

```
[1] 0.09090909
```

Question-C Create two separate pivot tables for the training data. One will have Loan (rows) as a function of Online (columns) and the other will have Loan (rows) as a function of CC.

```r
table(Personal.Loan=training_setPart$Personal.Loan,
      Online=training_setPart$Online)
```

```
             Online
Personal.Loan    0    1
            0 1102 1610
            1  119  169
```

```r
table(Personal.Loan=training_setPart$Personal.Loan,
      CreditCard=training_setPart$CreditCard)
```

```
             CreditCard
Personal.Loan    0    1
            0 1916  796
            1  197   91
```

```r
table(Personal.Loan=training_setPart$Personal.Loan)
```

```
Personal.Loan
   0    1
2712  288
```

Question-D Compute the following quantities [P(A | B) means "the probability ofA given B"]: i. P(CC = 1 | Loan = 1) (the proportion of credit card holders among the loan acceptors) ii. P(Online = 1 | Loan = 1) iii. P(Loan = 1) (the proportion of loan acceptors) iv. P(CC = 1 | Loan = 0) v. P(Online = 1 | Loan = 0) vi. P(Loan = 0)

```r
#(i) P(CC = 1 | Loan = 1)
P1=80/(80+208)
P1
```

```
[1] 0.2777778
```

```r
#(ii) P(Online = 1 | Loan = 1)
P2=179/(179+109)
P2
```

```
[1] 0.6215278
```

```r
#(iii) P(Loan = 1)
P3=288/(288+2712)
P3
```

```
[1] 0.096
```

```r
#(iv) P(CC = 1 | Loan = 0)
P4=779/(779+1933)
P4
```

```
[1] 0.2872419
```

```r
#(v) P(Online = 1 | Loan = 0)
P5=1599/(1599+1113)
P5
```

```
[1] 0.5896018
```

```r
#(vi) P(Loan = 0)
P6=2712/(288+2712)
P6
```

```
[1] 0.904
```

Question-E Use the quantities computed above to compute the naive Bayes probability P(Loan = 1 | CC = 1, Online = 1)

Calculating the naive Bayes probability for P(Loan = 1 | CC = 1, Online = 1).

```r
(P1*P2*P3)/((P1*P2*P3)+(P4*P5*P6))
```

```
[1] 0.09768187
```

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

The probability derived from the pivot table stands at 0.1005587, while the computed probability using the naive Bayes approach amounts to 0.1120411. Naive Bayes operates under the premise that attributes exhibit independence from one another. This implies that the probability obtained from the pivot table is the more reliable of the two.

Question-G Which of the entries in this table are needed for computing P(Loan = 1 | CC = 1, Online = 1)? Run naive Bayes on the data. Examine the 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).

```r
Naivebayes_model<-naiveBayes(Personal.Loan~CreditCard+Online,data = training_setPart)
testing<-data.frame(CreditCard=1,Online=1)
testing$CreditCard<-factor(testing$CreditCard)
testing$Online<-factor(testing$Online)
predict(Naivebayes_model,testing,type = 'raw')
```

```
             0         1
[1,] 0.8984709 0.1015291
```

The likelihood of the test data aligns with the probability computed in question E, which stands at 0.09768187. This implies that the Naive Bayes algorithm has arrived at an identical prediction as the calculated probability.