Sharon Kasturiarachi

MIS 64060 Fundamentals of Machine Learning

Dr. Rouzbeh Razavi

Assignment 2- Answers

Kasturiarachi-Assignment 3

(a)

```
> str(UniversalBank)
spec_tbl_df [5,000 x 14] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ ID                : num [1:5000] 1 2 3 4 5 6 7 8 9 10 ...
 $ Age               : num [1:5000] 25 45 39 35 35 37 53 50 35 34 ...
 $ Experience        : num [1:5000] 1 19 15 9 8 13 27 24 10 9 ...
 $ Income            : num [1:5000] 49 34 11 100 45 29 72 22 81 180 ...
 $ ZIP Code          : num [1:5000] 91107 90089 94720 94112 91330 ...
 $ Family            : num [1:5000] 4 3 1 1 4 4 2 1 3 1 ...
 $ CCAvg             : num [1:5000] 1.6 1.5 1 2.7 1 0.4 1.5 0.3 0.6 8.9 ...
 $ Education         : num [1:5000] 1 1 1 2 2 2 2 3 2 3 ...
 $ Mortgage          : num [1:5000] 0 0 0 0 0 155 0 0 104 0 ...
 $ Personal Loan     : num [1:5000] 0 0 0 0 0 0 0 0 0 1 ...
 $ Securities Account: num [1:5000] 1 1 0 0 0 0 0 0 0 0 ...
 $ CD Account        : num [1:5000] 0 0 0 0 0 0 0 0 0 0 ...
 $ Online            : num [1:5000] 0 0 0 0 0 1 1 0 1 0 ...
 $ CreditCard        : num [1:5000] 0 0 0 0 1 0 0 1 0 0 ...
 - attr(*, "spec")=
  .. cols(
> DF=Bank
> # Before converting
> summary(DF)
       ID            Age          Experience       Income          ZIP Code
 Min.   :   1   Min.   :23.00   Min.   :-3.0   Min.   :  8.00   Min.   : 9307
 1st Qu.:1251   1st Qu.:35.00   1st Qu.:10.0   1st Qu.: 39.00   1st Qu.:91911
 Median :2500   Median :45.00   Median :20.0   Median : 64.00   Median :93437
 Mean   :2500   Mean   :45.34   Mean   :20.1   Mean   : 73.77   Mean   :93153
 3rd Qu.:3750   3rd Qu.:55.00   3rd Qu.:30.0   3rd Qu.: 98.00   3rd Qu.:94608
 Max.   :5000   Max.   :67.00   Max.   :43.0   Max.   :224.00   Max.   :96651
     Family          CCAvg          Education        Mortgage     Personal Loan
 Min.   :1.000   Min.   : 0.000   Min.   :1.000   Min.   :  0.0   Min.   :0.000
 1st Qu.:1.000   1st Qu.: 0.700   1st Qu.:1.000   1st Qu.:  0.0   1st Qu.:0.000
 Median :2.000   Median : 1.500   Median :2.000   Median :  0.0   Median :0.000
 Mean   :2.396   Mean   : 1.938   Mean   :1.881   Mean   : 56.5   Mean   :0.096
 3rd Qu.:3.000   3rd Qu.: 2.500   3rd Qu.:3.000   3rd Qu.:101.0   3rd Qu.:0.000
 Max.   :4.000   Max.   :10.000   Max.   :3.000   Max.   :635.0   Max.   :1.000
 Securities Account   CD Account         Online         CreditCard
 Min.   :0.0000     Min.   :0.0000   Min.   :0.0000   Min.   :0.000
 1st Qu.:0.0000     1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000
 Median :0.0000     Median :0.0000   Median :1.0000   Median :0.000
 Mean   :0.1044     Mean   :0.0604   Mean   :0.5968   Mean   :0.294
 3rd Qu.:0.0000     3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.000
 Max.   :1.0000     Max.   :1.0000   Max.   :1.0000   Max.   :1.000
> DF$CreditCard = as.factor(Bank$CreditCard)
```

After converting to a categorical variable and deleting column “Online”

DF$Online_Category<-factor(Bank$Online, levels = c(0,1))

```
> summary(DF)
       ID            Age          Experience       Income          ZIP Code
 Min.   :   1   Min.   :23.00   Min.   :-3.0   Min.   :  8.00   Min.   : 9307
 1st Qu.:1251   1st Qu.:35.00   1st Qu.:10.0   1st Qu.: 39.00   1st Qu.:91911
 Median :2500   Median :45.00   Median :20.0   Median : 64.00   Median :93437
 Mean   :2500   Mean   :45.34   Mean   :20.1   Mean   : 73.77   Mean   :93153
 3rd Qu.:3750   3rd Qu.:55.00   3rd Qu.:30.0   3rd Qu.: 98.00   3rd Qu.:94608
 Max.   :5000   Max.   :67.00   Max.   :43.0   Max.   :224.00   Max.   :96651
     Family          CCAvg          Education        Mortgage     Personal Loan
 Min.   :1.000   Min.   : 0.000   Min.   :1.000   Min.   :  0.0   0:4520
 1st Qu.:1.000   1st Qu.: 0.700   1st Qu.:1.000   1st Qu.:  0.0   1: 480
 Median :2.000   Median : 1.500   Median :2.000   Median :  0.0
 Mean   :2.396   Mean   : 1.938   Mean   :1.881   Mean   : 56.5
 3rd Qu.:3.000   3rd Qu.: 2.500   3rd Qu.:3.000   3rd Qu.:101.0
 Max.   :4.000   Max.   :10.000   Max.   :3.000   Max.   :635.0
 Securities Account   CD Account     CreditCard Online_Category
 Min.   :0.0000     Min.   :0.0000   0:3530     0:2016
 1st Qu.:0.0000     1st Qu.:0.0000   1:1470     1:2984
 Median :0.0000     Median :0.0000
 Mean   :0.1044     Mean   :0.0604
 3rd Qu.:0.0000     3rd Qu.:0.0000
 Max.   :1.0000     Max.   :1.0000
> |
```

```
spec_tbl_df [5,000 x 14] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ ID                : num [1:5000] 1 2 3 4 5 6 7 8 9 10 ...
 $ Age               : num [1:5000] 25 45 39 35 35 37 53 50 35 34 ...
 $ Experience        : num [1:5000] 1 19 15 9 8 13 27 24 10 9 ...
 $ Income            : num [1:5000] 49 34 11 100 45 29 72 22 81 180 ...
 $ ZIP Code          : num [1:5000] 91107 90089 94720 94112 91330 ...
 $ Family            : num [1:5000] 4 3 1 1 4 4 2 1 3 1 ...
 $ CCAvg             : num [1:5000] 1.6 1.5 1 2.7 1 0.4 1.5 0.3 0.6 8.9 ...
 $ Education         : num [1:5000] 1 1 1 2 2 2 2 3 2 3 ...
 $ Mortgage          : num [1:5000] 0 0 0 0 0 155 0 0 104 0 ...
 $ Personal Loan     : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 2 ...
 $ Securities Account: num [1:5000] 1 1 0 0 0 0 0 0 0 0 ...
 $ CD Account        : num [1:5000] 0 0 0 0 0 0 0 0 0 0 ...
 $ CreditCard        : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 2 1 1 ...
 $ Online_Category   : Factor w/ 2 levels "0","1": 1 1 1 1 1 2 2 1 2 1 ...
 - attr(*, "spec")=
```

```
> mytable <-xtabs(~ CreditCard+`Personal Loan`+Online_Category, data = Train.df)
> ftable(mytable)
                          Online_Category    0    1
CreditCard Personal Loan
0          0                               785 1145
           1                                65  122
1          0                               317  475
           1                                34   57
>
```

(b)

```
> prop.table(mytable)
, , Online_Category = 0

           Personal Loan
CreditCard          0          1
         0 0.26166667 0.02166667
         1 0.10566667 0.01133333

, , Online_Category = 1

           Personal Loan
CreditCard          0          1
         0 0.38166667 0.04066667
         1 0.15833333 0.01900000

>
```

By row proportions

```
> round(prop.table(mytable, 1),3)
, , Online_Category = 0

           Personal Loan
CreditCard     0     1
         0 0.371 0.031
         1 0.359 0.039

, , Online_Category = 1

           Personal Loan
CreditCard     0     1
         0 0.541 0.058
         1 0.538 0.065
```

There are a total of (475+ 57) = 532 records where online = 1 and cc = 1.

57 of them accept the loan. Therefore, the conditional probability is 57/532 = 0.1071

(c) 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

```
> table(`Personal Loan`=Train.df$`Personal Loan`, Online_Category=Train.df$Online_Category)
             Online_Category
Personal Loan    0    1
            0 1102 1620
            1   99  179
> table(`Personal Loan`=Train.df$`Personal Loan`, CreditCard=Train.df$CreditCard)
             CreditCard
Personal Loan    0    1
            0 1930  792
            1  187   91
```

(d) Compute the following quantities [P(A | B) means "the probability of A given B"]:

i. P(CC = 1 | Loan = 1) (the proportion of credit card holders among the loan acceptors)

```
> table(`Personal Loan`=Train.
             CreditCard
Personal Loan    0    1
            0 1930  792
            1  187   91
> |
```

	Credit Card		Total
Personal Loan	0	1	
0	1930	792	2722
1	187	91	278
	2117	883	3000

P(CC = 1 | Loan = 1) = 91/278 = 0.327

```
> round(prop.table(mytable2),3)
             CreditCard
Personal Loan     0     1
            0 0.643 0.264
            1 0.062 0.030
> |
```

By row proportions

```
> round(prop.table(mytable2, 1),3)
             CreditCard
Personal Loan     0     1
            0 0.709 0.291
            1 0.673 0.327
> |
```

ii. P(Online = 1 | Loan = 1)

```
              Online_Category
Personal Loan    0    1
            0 1102 1620
            1   99  179
```

	Online_Category		Total
Personal Loan	0	1	
0	1102	1620	2722
1	99	179	278
	1201	1799	3000

P(Online = 1 | Loan = 1) = 179/278 = 0.644

```
> round(prop.table(mytable1), 3)
              Online_Category
Personal Loan     0     1
            0 0.367 0.540
            1 0.033 0.060
```

By row proportions

```
> round(prop.table(mytable1, 1),3)
              Online_Category
Personal Loan     0     1
            0 0.405 0.595
            1 0.356 0.644
> |
```

iii. P(Loan = 1) (the proportion of loan acceptors) when Personal Loan = 1 is 278 out of 30000

= 278/30000 = 0.093

iv. P(CC = 1 | Loan = 0)

```
              CreditCard
Personal Loan    0    1
            0 1930  792
            1  187   91
```

	Credit Card		Total
Personal Loan	0	1	
0	1930	792	2722
1	187	91	278
	2117	883	3000

P(CC = 1 | Loan = 0) = 792/2722 = 0.291

```
> round(prop.table(mytable2),3)
             CreditCard
Personal Loan     0     1
            0 0.643 0.264
            1 0.062 0.030
>
```

By row proportions

```
> round(prop.table(mytable2, 1),3)
             CreditCard
Personal Loan     0     1
            0 0.709 0.291
            1 0.673 0.327
>
```

v. P(Online = 1 | Loan = 0)

	Online_Category		Total
Personal Loan	0	1	
0	1102	1620	2722
1	99	179	278
	1201	1799	3000

P(Online = 1 | Loan = 0) = 1620/2722 = 0.595

```
             Online_Category
Personal Loan    0    1
            0 1102 1620
            1   99  179
```

By row proportions

```
> round(prop.table(mytable1, 1),3)
             Online_Category
Personal Loan     0     1
            0 0.405 0.595
            1 0.356 0.644
```

vi. P(Loan = 0) when Personal Loan = 0 is 2722 out of 30000

= 2722/3000 = 0.907

(e) Use the quantities computed above to compute the naive Bayes probability

P(Loan = 1 | CC= 1, Online = 1)

naive Bayes probability = (0.327)*(0.644)*(0.093)/ (0.327)*(0.644)*(0.093)+ (0.291)*(0.595)*(0.907)

=(0.327)*(0.644)*(0.093)/ (0.327)*(0.644)*(0.093)+ (0.291)*(0.595)*(0.907)

=0.0196/(0.0196 +0.1570) = 0.0196/0.1766 = 0.1109

= 0.111

(f) Compare this value with the one obtained from the pivot table in (B). Which is a more

accurate estimate?

From (b) 57/532 = 0.1071 from naiveByes = 0.111

The value using the naive Bayes is more accurate estimate than the value obtained from the pivot table in (b). This is because Naïve Base assumes conditional independence of the predictor variables even if the variables are correlated, it finds the probability of the belonging class without limiting the calculation to the records that have the same predictor values.

(g)

P(Loan = 1|CC =1; Online =1)

```
> library(e1071)
> nb.model<-naiveBayes(Online_Category~CreditCard+`Personal Loan`, data = Train.df)
> To_Predict=data.frame(CreditCard='1',`Personal Loan`='1' )
> predict(nb.model, To_Predict,type = 'raw')
              0         1
[1,] 0.3975085 0.6024915
>
```

```
> naiveBayes

Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
         0          1
0.90733333 0.09266667

Conditional probabilities:
   CreditCard
Y           0         1
  0 0.7090375 0.2909625
  1 0.6726619 0.3273381

   Online_Category
Y           0         1
  0 0.4048494 0.5951506
  1 0.3561151 0.6438849
```

```
> round(prop.table(mytable, 1),3)
, , Online_Category = 0

          Personal Loan
CreditCard     0     1
         0 0.371 0.031
         1 0.359 0.039

, , Online_Category = 1

          Personal Loan
CreditCard     0     1
         0 0.541 0.058
         1 0.538 0.065
```

NaiveBayes Probability=0.065/(0.538+0.065)= 0.065/0.603= 0.1077 same as in (b)