hw3

501 hw2.1

Qian Chen

Contents

Executive Summary

To understand whether there are significant differences in gender and age factors with different medication efficiency, age and gender were divided into four categories, and the naive Bayes model was established, and the training set and test set were used to test. It is found that the accuracy of the training set and the test set are very close, indicating that the model has a good effect, but the accuracy is not very high. In particular, men were more accurate and women were less accurate,older men (over 40) had the highest accuracy. It indicated that these drugs had obvious differences in gender and age factors, and the effect was more significant in men.

Full Report

Data Cleansing

The data in the first 18 columns of the original data were selected and classified according to gender and age. Taking 40 years old as the cut-off point, the data were roughly divided into four categories, namely, young men, old men, young women and old women, which were saved as group factor variables. Finally, the TRAINING set and test set are divided by a ratio of 4:1 and named training and TESTING. There are many missing values in the original data. Since the naiveBayes function allows missing values when doing naiveBayes analysis, it will not calculate the true terms, so the missing terms are not processed.

```
data=read.csv("R_cleaned_data.csv")

data$group=1*((data$Age=="13-19"|data$Age=="20-40")&data$Gender=="Female")+2*(!(data$Age=="13-19"|data$A

data=data[,c(3:18,26)]
data$group=as.factor(data$group)
```

Descriptive Statistics

By setting random seeds and making descriptive statistics on the training set, information such as minimum value, first quantile, median value, mean value, third quantile, maximum value and number of missing values of each variable can be seen. The training set has a total of 13416 rows, of which the first group has 2911 rows, the second group has 7126 rows, the third group has 1005 rows, and the fourth group has 2374 rows. The second group accounted for more in the discovery group.

```
set.seed(88)
sampling=sample(1:nrow(data),nrow(data)*0.8,replace = FALSE)
TRAINING=data[sampling,]
TESTING=data[-sampling,]
nrow(TRAINING)
```

```
## [1] 13416
```

```
summary(TRAINING)
```

```
##  AMITRIPTYLINE     BUPROPION       CITALOPRAM    DESVENLAFAXINE
##  Min.   :0.000   Min.   :0.000   Min.   :0.047   Min.   :0.046  
##  1st Qu.:0.008   1st Qu.:0.024   1st Qu.:0.361   1st Qu.:0.203  
##  Median :0.018   Median :0.040   Median :0.475   Median :0.511  
##  Mean   :0.029   Mean   :0.068   Mean   :0.478   Mean   :0.442  
##  3rd Qu.:0.037   3rd Qu.:0.058   3rd Qu.:0.617   3rd Qu.:0.635  
##  Max.   :0.319   Max.   :0.698   Max.   :0.912   Max.   :0.823  
##  NA’s   :12778   NA’s   :11310   NA’s   :11455   NA’s   :13236  
##     DOXEPIN        DULOXETINE     ESCITALOPRAM     FLUOXETINE   
##  Min.   :0.025   Min.   :0.038   Min.   :0.019   Min.   :0.029  
##  1st Qu.:0.119   1st Qu.:0.275   1st Qu.:0.116   1st Qu.:0.311  
##  Median :0.164   Median :0.378   Median :0.162   Median :0.450  
##  Mean   :0.212   Mean   :0.389   Mean   :0.209   Mean   :0.446  
##  3rd Qu.:0.290   3rd Qu.:0.505   3rd Qu.:0.273   3rd Qu.:0.592  
##  Max.   :0.726   Max.   :0.779   Max.   :0.948   Max.   :0.909  
##  NA’s   :13299   NA’s   :12370   NA’s   :10974   NA’s   :11417  
##   MIRTAZAPINE    NORTRIPTYLINE     PAROXETINE      ROPINIROLE   
##  Min.   :0.029   Min.   :0.010   Min.   :0.065   Min.   :0.065  
##  1st Qu.:0.109   1st Qu.:0.045   1st Qu.:0.324   1st Qu.:0.235  
##  Median :0.139   Median :0.065   Median :0.415   Median :0.382  
##  Mean   :0.182   Mean   :0.099   Mean   :0.443   Mean   :0.352  
##  3rd Qu.:0.221   3rd Qu.:0.127   3rd Qu.:0.560   3rd Qu.:0.449  
##  Max.   :0.704   Max.   :0.745   Max.   :0.942   Max.   :0.716  
##  NA’s   :13001   NA’s   :13196   NA’s   :12363   NA’s   :13260  
##    SERTRALINE      TRAZODONE      VENLAFAXINE        OTHER       group  
##  Min.   :0.045   Min.   :0.000   Min.   :0.051   Min.   :0.019   1:2911
##  1st Qu.:0.357   1st Qu.:0.008   1st Qu.:0.309   1st Qu.:0.182   2:7126  
##  Median :0.484   Median :0.015   Median :0.515   Median :0.333   3:1005  
##  Mean   :0.482   Mean   :0.035   Mean   :0.475   Mean   :0.346   4:2374  
##  3rd Qu.:0.625   3rd Qu.:0.026   3rd Qu.:0.636   3rd Qu.:0.487           
##  Max.   :0.960   Max.   :0.586   Max.   :0.908   Max.   :0.885           
##  NA’s   :10671   NA’s   :12225   NA’s   :12084   NA’s   :7156
```

Build model

Using the training set data to do Naive Bayes classification training, the parameter data of each variable are obtained as follows:

```
#install.packages("klaR")
#install.packages("caret")
library(klaR)
library(MASS)
library(tidyverse)
library(e1071)
library(caret)
model <- naiveBayes(group~., data = TRAINING,laplace = 0)
summary(model)
```

```
##           Length Class  Mode     
## apriori    4     table  numeric  
## tables    16     -none- list     
## levels     4     -none- character
## isnumeric 16     -none- logical  
## call       4     -none- call
```

```
model[1:2]
```

```
## $apriori
## Y
##    1    2    3    4 
## 2911 7126 1005 2374 
## 
## $tables
## $tables$AMITRIPTYLINE
##    AMITRIPTYLINE
## Y          [,1]       [,2]
##   1 0.01593162 0.02297333
##   2 0.03021216 0.03727609
##   3 0.02461720 0.04393033
##   4 0.04232392 0.04906467
## 
## $tables$BUPROPION
##    BUPROPION
## Y          [,1]       [,2]
##   1 0.06345101 0.10452770
##   2 0.06648531 0.10737051
##   3 0.06247740 0.08988754
##   4 0.08043100 0.12637988
```

```
##
## $tables$CITALOPRAM
##    CITALOPRAM
## Y        [,1]      [,2]
##   1 0.3969153 0.1595273
##   2 0.5259181 0.1711053
##   3 0.3614962 0.1510808
##   4 0.5152602 0.1693922
##
## $tables$DESVENLAFAXINE
##    DESVENLAFAXINE
## Y        [,1]      [,2]
##   1 0.4005846 0.2049971
##   2 0.5041807 0.2307743
##   3 0.3084157 0.1576679
##   4 0.4014441 0.2260122
##
## $tables$DOXEPIN
##    DOXEPIN
## Y        [,1]       [,2]
##   1 0.1471834 0.06732850
##   2 0.2206078 0.14341429
##   3 0.1468870 0.06475589
##   4 0.2483045 0.13022295
##
## $tables$DULOXETINE
##    DULOXETINE
## Y        [,1]      [,2]
##   1 0.3234018 0.1445106
##   2 0.4190225 0.1614224
##   3 0.3072140 0.1328874
##   4 0.3764414 0.1458542
##
## $tables$ESCITALOPRAM
##    ESCITALOPRAM
## Y        [,1]      [,2]
##   1 0.1742376 0.1102565
##   2 0.2316238 0.1549544
##   3 0.1740031 0.1142118
##   4 0.2302593 0.1409038
##
## $tables$FLUOXETINE
##    FLUOXETINE
## Y        [,1]      [,2]
##   1 0.3662099 0.1569459
##   2 0.5127779 0.1783902
##   3 0.3197576 0.1528554
##   4 0.5143816 0.1661522
##
## $tables$MIRTAZAPINE
##    MIRTAZAPINE
## Y        [,1]       [,2]
##   1 0.1462789 0.09932235
##   2 0.1809437 0.12423292
##   3 0.1283701 0.07253908
##   4 0.2113635 0.12926088
##
## $tables$NORTRIPTYLINE
##    NORTRIPTYLINE
## Y         [,1]       [,2]
##   1 0.06591028 0.05544560
##   2 0.10449437 0.10138691
##   3 0.08325420 0.02843395
##   4 0.13444744 0.12716724
##
## $tables$PAROXETINE
##    PAROXETINE
## Y        [,1]      [,2]
##   1 0.3573463 0.1314918
##   2 0.4807117 0.1673142
##   3 0.3519902 0.1402756
##   4 0.4902271 0.1765165
##
## $tables$ROPINIROLE
##    ROPINIROLE
## Y        [,1]       [,2]
##   1 0.2405408 0.09187148
##   2 0.3419170 0.13846804
##   3 0.3310911 0.08443993
##   4 0.3916550 0.12692624
##
## $tables$SERTRALINE
##    SERTRALINE
## Y        [,1]      [,2]
##   1 0.4097992 0.1672594
##   2 0.5445272 0.1824576
##   3 0.3594761 0.1698660
##   4 0.5393300 0.1837558
##
## $tables$TRAZODONE
##    TRAZODONE
## Y         [,1]       [,2]
##   1 0.03924059 0.08121022
##   2 0.03430328 0.07965588
##   3 0.03787187 0.07892185
##   4 0.03244389 0.07165869
##
## $tables$VENLAFAXINE
##    VENLAFAXINE
## Y        [,1]      [,2]
##   1 0.4185478 0.1835313
##   2 0.5092101 0.1909295
##   3 0.3967992 0.1821960
##   4 0.5011050 0.1922947
##
## $tables$OTHER
##    OTHER
## Y        [,1]      [,2]
##   1 0.2474623 0.1543639
##   2 0.3839827 0.1824118
##   3 0.2207556 0.1505837
##   4 0.3719530 0.1805411
```

Use the training set to see model efficiency

The confusion matrix was established to check the fitting efficiency of the training set, and it was found that the accuracy rate was 53.14%, and the accuracy rate of the second group (elderly men) was as high as 88.9%, and the accuracy rate of women was very low, indicating that these drugs had a significant effect on elderly men (over 40 years old).

```
pred <- predict(model,TRAINING)

cm=table(TRAINING$group,pred)
cm
```

```
##    pred
##        1    2    3    4
##   1  737 2104   45   25
##   2  716 6338   28   44
##   3  327  645   25    8
##   4  235 2099   11   29
```

```
confusionMatrix(cm)
```

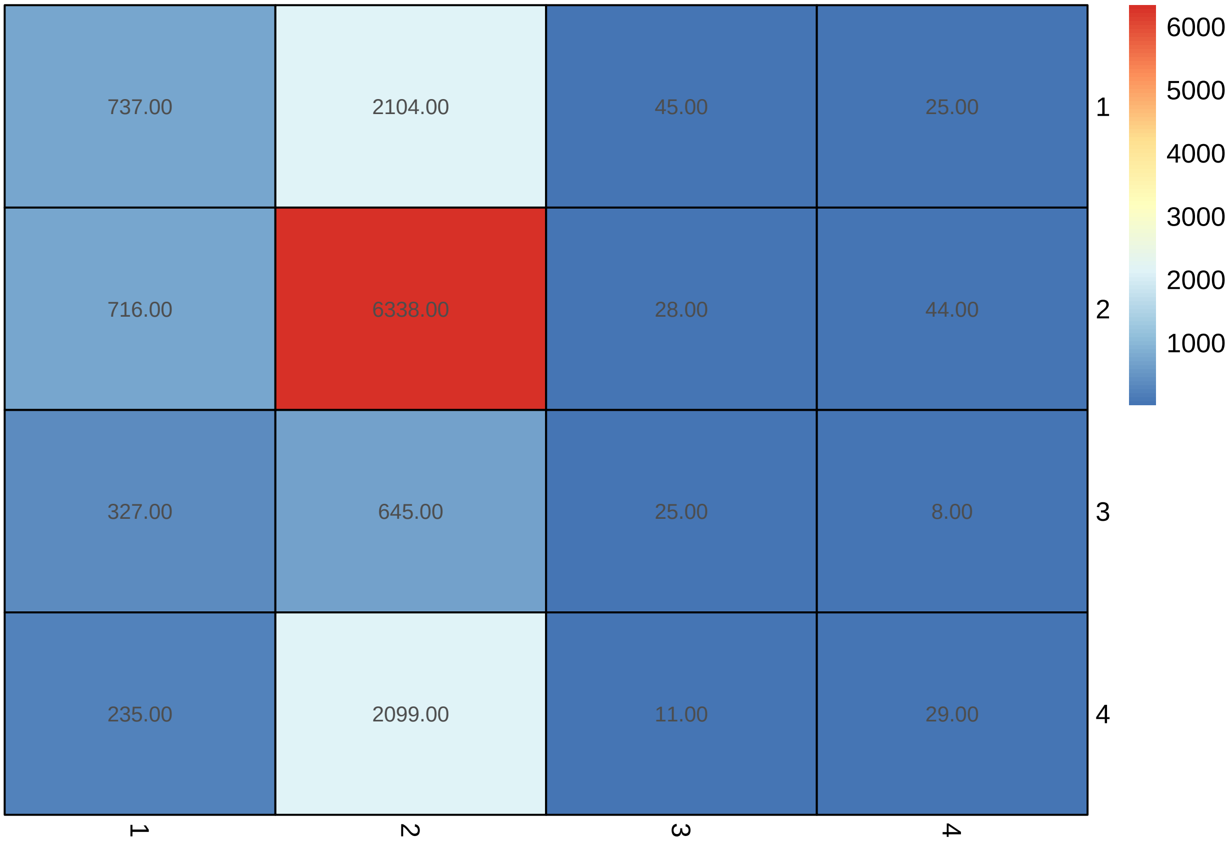
```
## Confusion Matrix and Statistics
##
##    pred
##        1    2    3    4
##   1  737 2104   45   25
##   2  716 6338   28   44
##   3  327  645   25    8
##   4  235 2099   11   29
##
## Overall Statistics
##
##                Accuracy : 0.5314
##                  95% CI : (0.5229, 0.5399)
##     No Information Rate : 0.8338
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1032
##
##  Mcnemar’s Test P-Value : <2e-16
##
## Statistics by Class:
##
##                      Class: 1 Class: 2 Class: 3 Class: 4
## Sensitivity           0.36576   0.5666 0.229358 0.273585
## Specificity           0.80931   0.6466 0.926355 0.823817
## Pos Pred Value        0.25318   0.8894 0.024876 0.012216
## Neg Pred Value        0.87834   0.2293 0.993232 0.993027
## Prevalence            0.15019   0.8338 0.008125 0.007901
## Detection Rate        0.05493   0.4724 0.001863 0.002162
## Detection Prevalence  0.21698   0.5312 0.074911 0.176953
## Balanced Accuracy     0.58754   0.6066 0.577856 0.548701
```

```
library(gmodels)

CrossTable(pred, TRAINING$group,
           prop.chisq = FALSE, prop.t = FALSE, prop.r = FALSE,
           dnn = c('predicted', 'actual'))
```

```
## 
## 
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Col Total |
## |-------------------------|
## 
## 
## Total Observations in Table:  13416
## 
## 
##              | actual
##    predicted |         1 |         2 |         3 |         4 | Row Total |
## -------------|-----------|-----------|-----------|-----------|-----------|
##            1 |       737 |       716 |       327 |       235 |      2015 |
##              |     0.253 |     0.100 |     0.325 |     0.099 |           |
## -------------|-----------|-----------|-----------|-----------|-----------|
##            2 |      2104 |      6338 |       645 |      2099 |     11186 |
##              |     0.723 |     0.889 |     0.642 |     0.884 |           |
## -------------|-----------|-----------|-----------|-----------|-----------|
##            3 |        45 |        28 |        25 |        11 |       109 |
##              |     0.015 |     0.004 |     0.025 |     0.005 |           |
## -------------|-----------|-----------|-----------|-----------|-----------|
##            4 |        25 |        44 |         8 |        29 |       106 |
##              |     0.009 |     0.006 |     0.008 |     0.012 |           |
## -------------|-----------|-----------|-----------|-----------|-----------|
## Column Total |      2911 |      7126 |      1005 |      2374 |     13416 |
##              |     0.217 |     0.531 |     0.075 |     0.177 |           |
## -------------|-----------|-----------|-----------|-----------|-----------|
## 
## 
```

```
library(pheatmap)
pheatmap(cm,
  cluster_cols = F, cluster_rows = F, scale = "none",
  treeheight_col = 0, treeheight_row = 0,
  display_numbers = T,
  border_color = "black")
```



Test the model with test sets

Using the test set to test, it is found that the accuracy is 52.06%, which is very close to the accuracy of the training set. Other results are consistent with the data of the training set, indicating that the model is accurate and the fitting result is good.

```
pred <- predict(model,TESTING)

cm=table(TESTING$group,pred)
cm
```

```
##    pred
##        1    2    3    4
##   1  184  512   13   10
##   2  172 1542   11   13
##   3   74  160   11    1
##   4   93  544    5    9
```

```
confusionMatrix(cm)
```

```
## Confusion Matrix and Statistics
##
##    pred
##        1    2    3    4
##   1  184  512   13   10
##   2  172 1542   11   13
##   3   74  160   11    1
##   4   93  544    5    9
## 
## Overall Statistics
## 
##                Accuracy : 0.5206          
##                  95% CI : (0.5035, 0.5376)
##     No Information Rate : 0.8223          
##     P-Value [Acc > NIR] : 1               
## 
##                   Kappa : 0.1083          
## 
##  Mcnemar’s Test P-Value : <2e-16          
## 
## Statistics by Class:
## 
##                      Class: 1 Class: 2 Class: 3 Class: 4
## Sensitivity           0.35182   0.5591  0.27500 0.272727
## Specificity           0.81102   0.6711  0.92909 0.806685
## Pos Pred Value        0.25591   0.8872  0.04472 0.013825
## Neg Pred Value        0.87135   0.2475  0.99067 0.991121
## Prevalence            0.15593   0.8223  0.01193 0.009839
## Detection Rate        0.05486   0.4597  0.00328 0.002683
## Detection Prevalence  0.21437   0.5182  0.07335 0.194097
## Balanced Accuracy     0.58142   0.6151  0.60204 0.539706
```

```
table(TESTING$group)
```

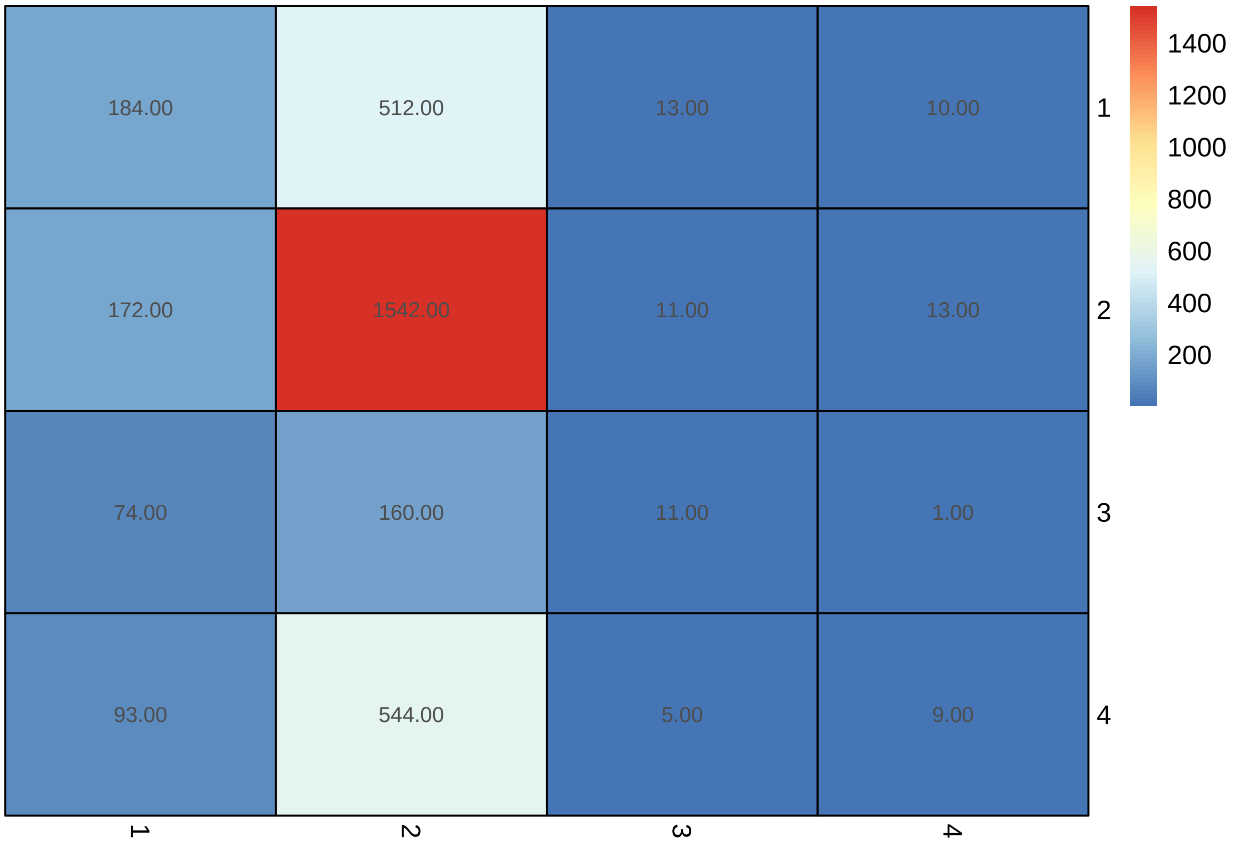
```
## 
##    1    2    3    4 
##  719 1738  246  651
```

```
library(gmodels)

CrossTable(pred, TESTING$group,
           prop.chisq = FALSE, prop.t = FALSE, prop.r = FALSE,
           dnn = c('predicted', 'actual'))
```

```
## 
## 
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Col Total |
## |-------------------------|
## 
## 
## Total Observations in Table:  3354 
## 
## 
##              | actual 
##    predicted |         1 |         2 |         3 |         4 | Row Total | 
## -------------|-----------|-----------|-----------|-----------|-----------|
##            1 |       184 |       172 |        74 |        93 |       523 |
##              |     0.256 |     0.099 |     0.301 |     0.143 |           |
## -------------|-----------|-----------|-----------|-----------|-----------|
##            2 |       512 |      1542 |       160 |       544 |      2758 |
##              |     0.712 |     0.887 |     0.650 |     0.836 |           |
## -------------|-----------|-----------|-----------|-----------|-----------|
##            3 |        13 |        11 |        11 |         5 |        40 |
##              |     0.018 |     0.006 |     0.045 |     0.008 |           |
## -------------|-----------|-----------|-----------|-----------|-----------|
##            4 |        10 |        13 |         1 |         9 |        33 |
##              |     0.014 |     0.007 |     0.004 |     0.014 |           |
## -------------|-----------|-----------|-----------|-----------|-----------|
## Column Total |       719 |      1738 |       246 |       651 |      3354 |
##              |     0.214 |     0.518 |     0.073 |     0.194 |           |
## -------------|-----------|-----------|-----------|-----------|-----------|
##
##
```

```
library(pheatmap)
pheatmap(cm,
  cluster_cols = F, cluster_rows = F, scale = "none",
  treeheight_col = 0, treeheight_row = 0,
  display_numbers = T,
  border_color = "black")
```



By mapping heat, confusion matrices can be better visualized.

Summarize

According to the above results, it can be found that both the training set and the test set have the highest accuracy of about 88% for the second group, namely, the older male group. It was followed by young men, meaning younger than 40, with an accuracy rate of about 25 percent. The accuracy rate for women is very low, all of which are less than 5%, indicating that these drugs have no significant change in women, so it can be said that the effect on women is not great.