Cpts575 Hw5

Mengxiao Zhang

Part 1

a. Estimate the probability.

The predict function is :

$$\hat{p} = \frac{e^{-7+0.1X_1+X_2-0.04X_3}}{1+e^{-7+0.1X_1+X_2-0.04X_3}}$$

$$\hat{p}(X_1 = 32, X_2 = 3.0, X_3 = 12) = 0.2175502$$

So, the probability of gets an A in the class is 21.75502%

b. How many hours would the student need to study.

According to the question a, we have the predict function and we need to let the predict equal to at least 0.5.

$$\hat{p}(X_1 = a, x_2 = 3.0, x_3 = 12) = 0.5$$

$$\Rightarrow \frac{e^{-7+0.1a+3.0-0.04*12}}{1+e^{-7+0.1a_1+3.0-0.04*12}} = 0.5$$

$$\Rightarrow e^{-7+0.1a+3.0-0.04*12} = 1$$

$$\Rightarrow -7 + 0.1a + 3.0 - 0.04 * 12 = 0$$

$$\Rightarrow a = 44.8$$

So, the student is part(a) needs to study 44.8 hours to have 50% chance of getting an A.

c. How many hours would a student with 3.0 GPA and a PSQI score of 3 need to have a 50% chance of getting an A.

$$\hat{p}(X_1 = a, x_2 = 3.0, x_3 = 3) = 0.5$$

$$\Rightarrow \frac{e^{-7+0.1a+3.0-0.04*3}}{1+e^{-7+0.1a_1+3.0-0.04*3}} = 0.5$$

$$\Rightarrow -7 + 0.1a + 3.0 - 0.04 * 3 = 0$$

$$\Rightarrow a = 41.2$$

So, the students with 3.0 GPA and PSQI score of 3 need to study 41.2 hours to have 50% chance of getting an A.

Part 2

a. Data Collection.

```
library(RJSONIO)
sections = c("world", "science", "business", "technology", "sport", "environment")
mydata = setNames(data.frame(matrix(ncol = 4, nrow = 0)),
                  c("id", "title", "body", "section"))
n = 0
dataset = list()
for (page in 1:10){
    for (s in 1:length(sections)){
        url = paste("http://content.guardianapis.com/search?section=", sections[s], "&lang=en&show-fiel
        data = fromJSON(url)$response$results
        for (i in 1:120){
            if (data[[i]]$type == "article"){
                mydata[nrow(mydata)+1,] = c(data[[i]]$id, data[[i]]$webTitle, data[[i]]$fields, data[[i]
            }
        }
    }
}
mydata[1,]$title
```

```
## [1] "Hong Kong's reluctant police officer: 'It's not for us to deliver punishment'"
```

b. Data cleaning.

```
for (i in 1:nrow(mydata)){
    Encoding(mydata[i,][["body"]]) <- "UTF-8"
    Encoding(mydata[i,][["title"]]) <- "UTF-8?gsub"
    mydata[i,]$body <- gsub("<.*?>", "", mydata[i,][["body"]])
    mydata[i,]$body <- gsub("[^[:alnum:][:space:]]",
"", mydata[i,][["body"]])
    mydata[i,]$body <- tolower(mydata[i,][["body"]])
}
strwrap(mydata[432,]$body, width=80)
```

```
##  [1] "the human thumb being opposable is a blessed thing we can hold a pen send a"
##  [2] "text or play thumb war as the mood takes us todays digital natives will have"
##  [3] "learned to type before they crawl and are capable of bashing out a hey you up"
##  [4] "text faster than the brain can process that it is a bad decision but not"
##  [5] "everyone is as lightningfast all of us know a onefinger typer whether it is the"
##  [6] "coworker who takes an eternity to reply to an email or the beloved grandparent"
##  [7] "who pauses every few seconds for the most part we have typed faster on computer"
##  [8] "keyboards than phone screens until now a study of more than 37000 volunteers"
##  [9] "from 160 countries has found that people can type almost as quickly on a screen"
## [10] "as they can on a keyboard those who used two thumbs were able to type on"
## [11] "average 38 words a minute making them just 25 slower than the average computer"
## [12] "keyboard user we took to the streets to put the study to the test and find out"
## [13] "how fast the general public could type the same threesentence 38word phrase"
## [14] "lifted of course from a guardian article on the uk economy move over mavis"
## [15] "beacon the keyboard is dead long live the phone screen jacky scottcombes 72"
## [16] "retired lincolnshire jacky scottcombes photograph graeme robertsonthe guardian"
## [17] "on a phone i type with one finger i can type but not on a phone2 minutes 33"
## [18] "seconds elspeth gower 22 yoga teacher edinburgh elspeth gower photograph graeme"
## [19] "robertsonthe guardian im pretty fast maybe34 seconds james hopethompson 52"
```

```
## [20] "hospitality worker liverpool james hopethompson photograph graeme robertsonthe"
## [21] "guardian i dont think im a fast typer im definitely a slow typer1 min 28"
## [22] "seconds yasmin bashir 25 medical student southall yasmin bashir photograph"
## [23] "graeme robertsonthe guardian im an ok typer average57 seconds carol grant 49"
## [24] "local government officer london carol grant photograph graeme robertsonthe"
## [25] "guardian i can type 150 words a minute on a keyboard i was trained to type but"
## [26] "on a phone nah1 minute 8 seconds austin champion 20 student london austin"
## [27] "champion photograph graeme robertsonthe guardian im dyslexic so i have to think"
## [28] "about everything quite a lot so i dont know if ill win but ill give it a go40"
## [29] "seconds"
```

c. Tokenization.

```
library(tm)
```

```
## Loading required package: NLP
```

```
TermMatrix = DocumentTermMatrix(Corpus(VectorSource(mydata$body)),
                                control=list(removeNumbers=TRUE,
                                             stopwords=TRUE,
                                             stemming=TRUE))
as.matrix(TermMatrix[32, which(as.matrix(TermMatrix[32, ]) != 0)])
```

```
##     Terms
## Docs actéon age agent alongsid alreadi also among angel anonym anoth
##   32       3   1     1        1       1    1     1     1      1     1
##     Terms
## Docs around art artist ask attent attract attribut auction away began bid
##   32      1   3      1   1      1       1        1      12    1     2   2
##     Terms
## Docs bidder biggest bin brandish bring broke buyer byzantin came case
##   32      1       1   1        1     1     1     1        2    2    1
##     Terms
## Docs cenni characteris child christ cimabu clearanc coent collect come
##   32     1           1     1      3     10        1     1       1    5
##     Terms
## Docs compiègn confirm consid contact content cook crept crowd crucial
##   32         1       1      1       1       2    1     1     1       1
##     Terms
## Docs crucifixion decad decid decor depict describ didnt diptych discoveri
##   32           1     1     1     1      1       1     2       3         1
##     Terms
## Docs dispos dominiqu due dump earli eight eighth element empti end enter
##   32      1        1   1    2     2     1      1       1     1   1     1
##     Terms
## Docs evalu ever everyth expens expert explor famili fee fell fetch final
##   32     1    2       1      2      5      1      2   1    1     2     1
##     Terms
## Docs flagel florentin forefath foreign form found four french frick
##   32      1         1        1       1    1     1    1      1     1
##     Terms
## Docs furnitur galleri gather giotto give hall hammer hand hang head
##   32        2       2      1      1    1    1      2    1    1    1
##     Terms
```

```
## Docs highest histori hotplat hous hung icon idea ident imagin immedi
##   32       1       1       1   12    2    1    1     1      1      1
##     Terms
## Docs includ incorpor influenc infrar initi italian job june just kitchen
##   32      1        1        1      1     1       5   1    1    1       4
##     Terms
## Docs known larger last leonardo local london long look make market master
##   32     2      1    2        1     1      1    1    1    2      2      2
##     Terms
## Docs masterpiec measur mediev middl might million miracul mock moment
##   32          1      1      3     1     2       1       1    1      2
##     Terms
## Docs month move movement much museum nailbit nation naturalist near never
##   32     1    2        1    1      1       1      2          1    1     2
##     Terms
## Docs new none north notic now object old one openplan outsid paint painter
##   32   1    1     1     1   1      1   3   1        1      1    12       3
##     Terms
## Docs panel pari parisien part passion peopl pepo perspect philomèn piec
##   32     1    3        1    1       1     1    1        1         1    1
##     Terms
## Docs pioneer popular price primitiv qualiti rank raphael rare readi
##   32       1       1     2        1       1    1       1    2     1
##     Terms
## Docs reflectographi relat remain rembrandt renaiss repres reveal rise room
##   32              1     1      2         1       3      1      1    1    1
##     Terms
## Docs ruben russia said sale scene schedul see seem sell sen sign silent
##   32     1      1    5    3     3       1   1    1    1   4    1      1
##     Terms
## Docs simpli sold someth soon space spot statement still style suggest
##   32      1    3      1    1     1    1         1     1     1       1
##     Terms
## Docs summer sunday surpris taught telephon thcenturi thought
##   32      1      1       1      1        1         1       3
##     Terms
## Docs threedimension tini told top turquin two uniqu unsign use valu view
##   32              1    1    2   1       1   3     1      1   1    2    1
##     Terms
## Docs vinci virgin wall week western wide will wolf woman wood work year
##   32     1      1    1    1       1    1    1    2     5    1    6    1
##     Terms
## Docs york
##   32    1
```

d. Classification.

```
library(e1071)
library(caret)
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following object is masked from 'package:NLP':
##
##     annotate
```

```
TermMatrix = removeSparseTerms(TermMatrix, 0.99)

CorMatrix = cor(as.matrix(TermMatrix))
CorTerms = sort(findCorrelation(CorMatrix, cutoff = 0.9))
TermMatrix = TermMatrix[, -c(CorTerms)]

TM_train = TermMatrix[1:floor(nrow(TermMatrix)*0.8),]
TM_test = TermMatrix[(floor(nrow(TermMatrix)*0.8)+1):nrow(TermMatrix),]
data_train = mydata[1:floor(nrow(mydata)*0.8), ]
data_test = mydata[(floor(nrow(mydata)*0.8)+1):nrow(mydata), ]

MyBayes = naiveBayes(as.matrix(TM_train), as.factor(data_train$section))
Prediction = predict(MyBayes, as.matrix(TM_test))
confusionMatrix(Prediction, as.factor(data_test$section))
```

```
## Confusion Matrix and Statistics
##
##              Reference
## Prediction    business environment science sport technology world
##   business         193          38       5     0         42    16
##   environment       12         157      10     0          7    33
##   science            2          28     188     1         18    18
##   sport              1           8      15   209         12     8
##   technology         6           3      17     2        145     5
##   world              6           6       4     4         14   156
##
## Overall Statistics
##
##                Accuracy : 0.7545
##                  95% CI : (0.731, 0.7769)
##     No Information Rate : 0.1728
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.7057
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
## Statistics by Class:
##
##                      Class: business Class: environment Class: science
## Sensitivity                   0.8773             0.6542         0.7866
## Specificity                   0.9136             0.9460         0.9417
## Pos Pred Value                0.6565             0.7169         0.7373
## Neg Pred Value                0.9753             0.9291         0.9550
## Prevalence                    0.1584             0.1728         0.1721
## Detection Rate                0.1389             0.1130         0.1353
## Detection Prevalence          0.2117             0.1577         0.1836
## Balanced Accuracy             0.8954             0.8001         0.8642
```

```
##                      Class: sport Class: technology Class: world
## Sensitivity                0.9676            0.6092       0.6610
## Specificity                0.9625            0.9713       0.9705
## Pos Pred Value             0.8261            0.8146       0.8211
## Neg Pred Value             0.9938            0.9232       0.9333
## Prevalence                 0.1555            0.1713       0.1699
## Detection Rate             0.1505            0.1044       0.1123
## Detection Prevalence       0.1821            0.1281       0.1368
## Balanced Accuracy          0.9650            0.7903       0.8158
```