BUSN 41201 HW3

Group 11: Yu-Ting Weng, Mengdi Hao, Elena Li, Minji Park, Sarah Lee

```
library(knitr)

data<-read.table("Review_subset.csv",header=TRUE)
dim(data)
```

```
words<-read.table("words.csv")
words<-words[,1]
length(words)
```

```
doc_word<-read.table("word_freq.csv")
names(doc_word)<-c("Review ID","Word ID","Times Word" )
```

Question 1

```
Y<-as.numeric(data$Score==5)

library(gamlr)
source("naref.R")

data$Prod_Category<-as.factor(data$Prod_Category)
class(data$Prod_Category)
```

```
data$Prod_Category<-naref(data$Prod_Category)
products<-data.frame(data$Prod_Category)
x_cat<-sparse.model.matrix(~., data=products)[,-1]
colnames(x_cat)<-levels(data$Prod_Category)[-1]
lasso1<- gamlr(x_cat, y=Y, standardize=FALSE,family="binomial", lambda.min.ratio=1e-3)
```

1) What is the in-sample R2 for the AICc slice of the LASSO path?

: **0.1048737**

```
null_model <- glm(Y ~ 1, family = 'binomial')
null_deviance <- null_model$deviance
```

```
aicc_values <- AICc(lasso1)
```

```
min_aicc_index <- which.min(aicc_values)
best_lambda <- lasso1$lambda[min_aicc_index]
best_lambda
```

```
##        seg91
## 1.464148e-05
```

```
best_deviance <- lasso1$deviance[min_aicc_index]
R_2 <- 1 - (best_deviance/null_deviance)
R_2
```

```
##     seg91
## 0.1048737
```

2) Why did we use standardize FALSE?

The variable Prod_Category is categorical. Standardizing is generally applied to continuous variables to bring them onto the same scale. Once converted into dummy or indicator variables for modeling (as happens with model.matrix or sparse.model.matrix), **categorical variables do not require standardization because they represent binary presence or absence (encoded as 0 or 1).**

Question 2

```
data$Prod_Category<-naref(data$Prod_Category)
spm<-sparseMatrix(i=doc_word[,1],
                  j=doc_word[,2],
                  x=doc_word[,3],
                  dimnames=list(id=1:nrow(data),
                                words=words))
dim(spm)
```

```
x_cat2<-cbind(x_cat,spm)

lasso2 <- gamlr(x_cat2, y=Y,lambda.min.ratio=1e-3,family="binomial")
```

1) Use AICc to select lambda

```
aicc_values <- AICc(lasso2)
min_aicc_index <- which.min(aicc_values)
best_lambda <- lasso1$lambda[min_aicc_index]
best_lambda
```

```
##        seg89
## 1.683414e-05
```

2) How many words were selected as predictive of a 5 star review?

: **1022**

```
n_x_cat <- dim(x_cat)[-1]
```

```
coefficients <- coef(lasso2, lambda = best_lambda)
# consider only words (not product category)
coefficients_spm <- coefficients[-1][(n_x_cat + 1):(length(coefficients)-1)]
non_zero_spm <- sum(coefficients_spm != 0)
non_zero_spm
```

```
## [1] 1022
```

3) Which 10 words have the most positive effect on odds of a 5 star review?

: **worried, plus, excellently, find, grains, hound, sliced, discount, youd, doggies**

```
words_spm <- rownames(coefficients)[-2][(n_x_cat + 1): (length(coefficients)-1)]
df <- data.frame(words = words_spm, coefficients = coefficients_spm)
head(df)
```

```
##        words coefficients
## 1      about    0.5515879
## 2   absolute   -1.1025174
## 3 absolutely    1.7864341
```

```
## 4 absorbable    4.1204837
## 5     action    3.5581502
## 6   actually    1.1506747
```

```
df_sorted <- df[order(-df$coefficients), ]
head(df_sorted$words, 10)
```

```
##  [1] "worried"     "plus"        "excellently" "find"        "grains"
##  [6] "hound"       "sliced"      "discount"    "youd"        "doggies"
```

4) What is the interpretations of the coefficient for the word 'discount'?

```
discount_coefficient <- coefficients["discount", ]
discount_coefficient
```

```
## [1] 6.961539
```

Since the coefficient is positive(6.96), **it indicates that the presence or higher frequency of the word "discount" in a product review is associated with an increased likelihood of the review being five stars.** This could suggest that customers who mention "discounts" in their reviews may be more satisfied with their purchase, possibly because they feel they received good value for money.

Question 3

```
set.seed(123)
cv.fit <- cv.gamlr(x_cat2, y=Y,
                   lambda.min.ratio=1e-3,
                   family="binomial",
                   verb=TRUE)
```

1) How many coefficients are nonzero then?

: **987**

```
non_zero_min <- sum(coef(cv.fit, select = 'min')[-1] != 0)
non_zero_min
```

```
## [1] 987
```

2) How many are nonzero under the 1se rule?

: **830**

```
non_zero_1se <- sum(coef(cv.fit, select = '1se')[-1] != 0)
non_zero_1se
```

```
## [1] 830
```