BUS 41201 Homework 3 Assignment

Group 11

2024-04-06

```
# Read in Data
setwd("C:/Users/user/Desktop/Big Data/HW/week3")
data<-read.table("Review_subset.csv",header=TRUE)
dim(data)
```

```
## [1] 13319     9
```

```
# 13319 reviews
# ProductID: Amazon ASIN product code
# UserID: id of the reviewer
# Score: numeric from 1 to 5
# Time: date of the review
# Summary: text review
# nrev: number of reviews by this user
# Length: length of the review (number of words)
# READ WORDS
words<-read.table("words.csv")
words<-words[,1]
length(words)
```

```
## [1] 1125
```

```
#1125 unique words
# READ text-word pairings file
doc_word<-read.table("word_freq.csv")
names(doc_word)<-c("Review ID","Word ID","Times Word" )
# Review ID: row of the file Review_subset
# Word ID: index of the word
# Times Word: number of times this word occurred in the text
```

QUESTION 1

We want to build a predictor of customer ratings from product reviews and product attributes. For these questions, you will fit a LASSO path of logistic regression using a binary outcome:

For 5 stars: $Y = 1$

Less than 5 stars: $Y = 0$

Fit a LASSO model with only product categories. The start code prepares a sparse design matrix of 142 product categories. What is the in-sample R2 for the AICc slice of the LASSO path? Why did we use standardize FALSE? (1 point)

```
# Let's define the binary outcome
# Y=1 if the rating was 5 stars
# Y=0 otherwise
Y<-as.numeric(data$Score==5)
# (a) Use only product category as a predictor
library(gamlr)
```

```
## Warning: package 'gamlr' was built under R version 4.2.3
```

```
## Loading required package: Matrix
```

```
## Warning: package 'Matrix' was built under R version 4.2.3
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
## Loading required package: Matrix
source("naref.R")

# Cast the product category as a factor
data$Prod_Category<-as.factor(data$Prod_Category)
class(data$Prod_Category)
```

```
## [1] "factor"
```

```
# Since product category is a factor, we want to relevel it for the LASSO.
# We want each coefficient to be an intercept for each factor level rather than a contrast.
# Check the extra slides at the end of the lecture.
# look inside naref.R. This function relevels the factors for us.
```

```
data$Prod_Category<-naref(data$Prod_Category)
# Create a design matrix using only products
products<-data.frame(data$Prod_Category)
x_cat<-sparse.model.matrix(~., data=products)[,-1]
# Sparse matrix, storing 0's as .'s
# Remember that we removed intercept so that each category
# is standalone, not a contrast relative to the baseline category
colnames(x_cat)<-levels(data$Prod_Category)[-1]
# let's call the columns of the sparse design matrix as the product categories
# Let's fit the LASSO with just the product categories
lasso1<- gamlr(x_cat, y=Y, standardize=FALSE,family="binomial",
lambda.min.ratio=1e-3)
```

```
optimal_index <- which.min(AICc(lasso1))
optimal_index
```

```
## seg91
##    91
```

```
# in-sample R2 for the AICc slice of the LASSO path?
summary(lasso1)$r2[which.min(AICc(lasso1))]
```

```
##
## binomial gamlr with 142 inputs and 100 segments.
```

```
## [1] 0.1048737
```

When standardize=FALSE is set in the LASSO model, it means that the predictors (the product categories) retain their original scale before fitting the model. With categorical predictors like product categories, the scale doesn't have a clear numerical interpretation, so standardization may not be meaningful.

Question 2

Fit a LASSO model with both product categories and the review content (i.e. the frequency of occurrence of words). Use AICc to select lambda. How many words were selected as predictive of a 5 star review? Which 10 words have the most positive effect on odds of a 5 star review? What is the interpretation of the coefficient for the word 'discount'? (3 points)

```
# Fit a LASSO with all 142 product categories and 1125 words
spm<-sparseMatrix(i=doc_word[,1],
j=doc_word[,2],
x=doc_word[,3],
dimnames=list(id=1:nrow(data),
words=words))
dim(spm) # 13319 reviews using 1125 words
```

```
## [1] 13319  1125
```

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

```
## Warning in gamlr(x_cat2, y = Y, lambda.min.ratio = 0.001, family = "binomial"):
## numerically perfect fit for some observations.
```

```
# AICc selected coef
log(lasso2$lambda[which.min(AICc(lasso2))])
```

```
##     seg89
## -8.334091
```

```
lasso2beta <- coef(lasso2)

# How many words were selected as predictive of a 5 star review?
num_predictive_words <- sum(lasso2beta[144:length(lasso2beta)]!=0)
cat('Number of words selected as predictive of a 5-star review: ', num_predictive_words, '\n')
```

```
## Number of words selected as predictive of a 5-star review:  1022
```

```
# Which 10 words have the most positive effect on odds of a 5 star review?
best_lambda <- lasso2$lambda[which.min(AICc(lasso2))]
coefs <- coef(lasso2, lambda = best_lambda)[144:length(lasso2beta)]

# Sort coefficients and indices by coefficient value
sorted_indices <- order(coefs, decreasing = TRUE)

# Extracting the top 10 words
top_10_words <- words[sorted_indices[1:10]]
cat('Top 10 words with the most positive effect on the odds of a 5-star review:')
```

```
## Top 10 words with the most positive effect on the odds of a 5-star review:
```

```
print(top_10_words)
```

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

```
# What is the interpretations of the coefficient for the word 'discount'?
coef(lasso2, lambda = best_lambda)["discount",]
```

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

The coefficient for the word 'discount' is approximately 6.96. This positive coefficient indicates that the presence of the word 'discount' in a review is associated with an increased likelihood of a 5-star rating.

Question 3

Continue with the model from Question 2. Run cross-validation to obtain the best lambda value that minimizes OOS deviance. How many coefficients are nonzero then? How many are nonzero under the 1se rule? (1 point)

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

```
## Warning in gamlr(x, y, ...): numerically perfect fit for some observations.
```

```
## fold 1,2,3,4,5,done.
```

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

```
## [1] 952
```

```
## CV 1se selection (the default)
cv.1se <- coef(cv.fit)[-1]
sum(cv.1se!=0) ## usually selects all zeros (just the intercept)
```

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