Project 4

Xiang Gao

The goal of this project is to build a sentiment classifier over the movie reviews of IMDB. The model will be trained on the text of reviews(ipnut) and predict the sentiment(output).

Preprocessing

Instead of load data using read.table, I used fread function to get rid of the exceptions when dealing with tsv file. Then I used regular expression to remove punctuations and numbers.

Since we are allowed to use all data to generate vocabulary, I build a dictionary using text2vec package, then create a word and bigram vector and docment term matrix. Then apply two sample t-test to two sentiment groups. The feature with higher t-statistics will be selected. High magnitude of t-statistics means for certain feature, the mean of positive sentiment group is close to negative group, thus we can acquire features with more predictive power.

Since we already have the desired vocabulary, a word vector will be built based on that. Then the training and testing data matrix will be created using the counts of words and bigrams in the vocabulary. I use 10 folds cross validated of ridge regression to find the lambda value that produces the minimum value of auc. Then a ridge regression model is built based on that. Then I use predict() to acquire the probability of class equals to 1.

Ridge regression can't zero out coefficient. So there is no variable selections.Another downside is Interpretability, unimportant variables are hard to explain. The metric of this project is AUC, which is the area under the ROC curve, At different threshold (for example 0.01,0.02,...1.00), the true positive rate and false positive rate are computed, then ROC combines these two rates. My next step is trying gradient boosting classifier and neural network.

Below shows the auc results of model performance for 3 splits.

```
     [,1]      [,2]      [,3]
```

[1,] 0.9629918 0.9628988 0.9636469

Vocab size 3000