Movie Sentiment Analysis Report

Jiamin Geng (jiamin2)

I. Overview

The aim of this project is to predict the sentiment of the movie reviews from IMBD data. This dataset we used has 50,000 observations with 3 variables, id, sentiment and review. We built three models: Logistic model, Naïve Bayes and XG-Boost on three training datasets and compare their performance on the three test sets. It turned out that the Logistic model with lasso penalty performs the best over all three test datasets.

II. Creating Bag of Words

To extract information from text data, I used 'text2vec' package in R to create a vocabulary with 3000 words and use this vocabulary to create training and test matrix using the following steps:

i. Created vocabulary using up to 4 grams and set some common stop words to avoid meaningless words.
ii. Prune the vocabulary by setting the minimum counts of word in all documents to 5, maximum portion of appearance to 0.5 and minimum portion to 0.001. This give us a smaller vocabulary with about 28000 words.
iii. Apply two sample T-test on all the. Pick the first 3000 words based on the magnitude of the T-statistics. And write a document to save the final bag of words.
iv. Generate the test and train data Document Term Matrix with words in the final vocabulary.

III. Classifiers

i. Logistic model with Lasso penalty

First, I implemented a 10-folds CV to pick the lambda value and used the lambda.min result from the CV to build the classifier.
This model performs the best among all three model with all three test datasets achieved AUC values above 0.96.

ii. Naïve Bayes

I used e1071 package in R to do Naïve Bayes. I predicted the posterior probability of the results in test data using this classifier and computed the AUC value.
The performance of this model is not as good as the other two, with AUC of all three models around 0.80.

iii. XG-Boost

The third model is a boost model. I built the XG-Boost classifier, and set the learning rate to be 0.05, the max depth of each tree to be 4 and used 500 trees to train the model. This classifier has an average AUC around 0.94.

IV. Performance

The AUC value of the three datasets is shown as follows:

	Split 1	*Split 2*	*Split 3*
Logistic	0.9633	0.9636	0.9609
Naïve Bayes	0.8075	0.8159	0.8036
XGboost	0.9367	0.9379	0.9364

V. Model limitations

i. There are a lot of meaningless words and repetitive words in the vocabulary. Further cleaning of the vocabulary may help improve the predict power of all three models.

ii. The posterior probability given by Naive Bayes is too close to 0 or 1. Maybe a log transformation on the predicted probabilities will help improve the performance of the model.

iii. We can also try to prune the vocabulary based on the performance of each model to achieve better performance in each classifier.

iv. Further tuning of the parameters in XG-Boost model may let to better performance, however, since calculating this model takes too much time, I didn’t tune the parameters very well.