Movie Review Sentiment Analysis

Ahmad Al-Dhalaan

11/27/2020

Introduction

In this report, we build a binary classification model on a dataset consisting of 50,000 IMDB movie reviews to predict the sentiment of a movie review. The sentiments are 0 (negative) and 1 (positive) as given by the dataset.

Implementation

We generate 5 sets of training/test splits, training a set on a 997 term vocabulary set and testing each split seperatley. See the included "Appendix - Vocabulary Construction" file for details on the vocabulary construction process. We create a document-term matrix for the vocabulary set as predictor variables and the binary sentiment as the response to train a binary ridge logistic regression model. We use cross-validation to determine the λ_{min} that minimizes the area under the curve (AUC) and refit the model with that λ_{min} value.

To test our model, we construct a document-term matrix for the test reviews and predict the probability of each review being positive. An example is below, showing the predictions on the first five reviews of the first split:

Table 1: Probability of a positive review for the first five review of the first split

id	probability
598	0.6134437
12048	0.2071972
40908	0.7657705
33283	0.9169127
40600	0.0003928

Results

We use a t2.large AWS instance with 8GB of RAM to run the five folds. The running time for the five folds and their AUCs are below. All folds achieve a AUC above the 0.96 benchmark and are trained in just over a minute. We note that although split 1 is used to construct the vocabulary, it in fact has the smallest AUC. This could be due to the vocabulary appearing more in the test set of the remaining splits, thereby improving predictions.

Table 2: AUC for train/test splits

Split	Time(s)	AUC
1	73.828	0.9608
2	74.188	0.9639
3	76.665	0.9671
4	74.563	0.9670
5	79.828	0.9668

We have shown that a vocabulary size of 997 terms is sufficient in predicting a binary sentiment classification of the IMDB review dataset. The average AUC for the five splits is 0.965. With sufficient memory such as a t2.large AWS cluster, training and testing is a relatively quick process averaging 75.81 seconds.

Interpretability

While we are confident of our prediction accuracy and overall performance, discussion of the interpretability of our model is of top priority. Specifically, the vocabulary chosen is highly determined by our tuning. While removing stop words and highly frequent and infrequent terms are standard in constructing a vocabulary set, other choices were less so. Given the 1000 word limit for our vocabulary, we had to select the most significant terms. In a situation where we are asked to interpret out model, these terms would face the most scrutiny.

Analyzing our 997 terms, we note that 424 terms are considered "positive" terms (determined by their positive t-statistic value). These are terms, such as "great", "excellent" and "best". There are 571 terms that are considered "negative", such as "bad", "worst" and "waste". We can see that negative terms are slightly more predominant in our vocabulary set at 57.3% of the set. This slight imbalance might mean that our model is more "inclined" to predict a negative sentiment, given the vocabulary input. Indeed, looking at the confusion matrix for the first split, for example, we see a slightly higher sensitivity at 91.1% compared to the specificity at 88.8%. This indicates that the model is making more false negatives than false positives. The model and by extension the vocabulary set might be biased by missclassifying positive reviews as negative.

	Reference	
Prediction	0	1
0	11091	1436
1	1079	11394

Indeed, examing the negative vocabulary set, we see terms that might not necessarily indicate a negative sentiment. These include terms such as "oh", "director" and "actor". We do not see such ambigious terms in the positive set.

However, on a practical level, it is better that our model missclassifies positive reviews as negative than vice versa. A production studio would be more interested in capturing negative sentiment in order to resolve it for future movie releases. Therefore, not missing negative reviews is much more important. Our model misses only 8.9% of negative reviews (for split 1) and that should give a production company confidence that most negative reviews will be captured in the future.