report

Tianqi Wu

12/11/2018

Overview of Model

The dataset consists of IMDB movie reviews, where each review is labelled as positive or negative. We want to predict the sentiment of a movie review with our binary classfication model. The dataset has 50000 observations and we split them into three sets of training and testing data. The model used is ridge regression from R package 'glmnet'. The input of the model is document_term_matrix of the testing data and output is the estimated probability of having sentiment = 1. The minimum AUC over three test data is 0.9621.

Customized vocabulary

First, I used R package 'text2vec' to build vocabulary and prune the vocabulary to remove some infrequent and frequent terms based on the training data of third split. Then, I construct DT matrix for the training data and I ended up having a short-and-fat design matrix: the number of features is 30436. Also, I used a customized list of stop-words. Finally, I did a two-sample-t-test to select the top 2000 words as the vocabulary and save it as myVocab.txt.

Technical details

The preprocessing also includes cleanning the html tag. After reading the myVocab.txt, we can create vocab for training and testing data. When creating iterators in order to create vocabularies, I used following parameters for itoken function: preprocess_function = tolower, tokenizer = word_tokenizer. I used ridge regression on the selected 2000 words and made prediction with lambda.min.The parameter used for cv.glmnet is:family='binomial',type.measure = "auc", nfolds = 10, alpha=0.

Model validation:

Te performance on the three test datasets is illustrated below:

Split1	Split2	Split3	Vocab_Size
0.9656	0.9643	0.9621	2000

The model uses ridge penalty to keep correlated features selected together as a group and Elastic-net penalty could be tried. Neither lemmatisation nor stemming is applied and the vocabulary may be redundent.Some terms may be meaningless too. We could further reduce the vocab size. Also, some polarized terms may be missed during the process and we could further investigate. The model sometimes fails to correctly classify the sentiment of reviews mixed with positive and negative terms. We may look into those misclassfied ones too. The third split has the highest AUC and the first split is lowest. For future steps, we could also try TF_IDF transformation and then apply the screening. It takes around 1 minute to run and the computer system is Macbook Pro 3.1GHz, 8GB memory.