PROJECT 4

DUE: 11:59PM, APRIL 12

1 Introduction

In this project, you will need to train a logistic regression model to classify the movie reviews into positive and negative reviews. We are going to use the following dataset.

Internet Movie Database (IMDb). This dataset was collected by Maas et al. [1]. The movie review dataset consists of $50,000$ popular movie reviews that are labeled as either positive or negative; here, positive means that a movie was rated with more than six stars on IMDb, and negative means that a movie was rated with fewer than five stars on IMDb. You can download it from `http://ai.stanford.edu/~amaas/data/sentiment/`.

After downloading the dataset, you can proceed to build your model. To help you build the model, I have listed the key steps you need to do during the training process:

1. **Clean text data.** During this step, you may want to strip all unwanted characters from review texts.

2. **Process documents into tokens.** Some words may have different forms. You may want to group them into one word.

3. **Transform words into feature vectors.** Transform each document into a vector where each dimension represents the frequency of a word (bag-of-words model).

4. **Access word relevancy.** When we are analyzing text data, we often encounter words that occur across multiple documents from both classes. Those frequently occurring words typically don't contain useful or discriminatory information. In this step, you may want to downweight those frequently occurring words in the feature vectors.

5. **Build the logistic model.** After finishing all the abovementioned steps, you are now ready to build the model using the manipulated vectors. Remember to report your 5-fold cross-validation error.

You are highly suggested to read Pg. 233 - 245 of Python Machine Learning book, where you can find most of the codes you need.

2 Submission

You will need to submit one **pdf** file generated by the notebook. Fail to do so will make your final grade deducted. Make sure all codes are run before generating the pdf file. In the report, you should specify your model details when necessary. Try to write your code clearly so that someone else reading the code can understand it without significant effort (i.e. structure it and put enough documentation). The final grade is based on the clarity of your report.

3 Collaboration

Note that this is an **independent** project, which means you are not allowed to make a group. However, discussion is allowed. If you have discussed with someone or got any help from others, you need to clearly specify their names in acknowledgement.

References

[1] Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *ACL*, pages 142–150, 2011.