COL 341: Assignment 5

Max Marks: 100
Due Date: 18 Nov 2019

Notes:

- This assignment has one part - Naive Bayes.
- Submit whole code with a detailed write-up.
- The report should be a brief description explaining what you did. Include any observations and/or plots required by the question in the report.
- You should use Python for all your programming solutions.
- Your assignments will be auto-graded, make sure you test your programs before submitting. We will use your code to train the model on training data and predict on test set.
- You should submit work of your own. You should cite the source, if you choose to use any external resources. In case of plagiarism serious actions will be taken.

1. Text Classification using Naive Bayes

   In this problem, we will use the Naive Bayes algorithm for text classification. The dataset for this problem is a subset of IMDB movie review data. First entry of each row is a review followed by a sentiment. The sentiment attached with the review has two class labels: positive and negative. Given a review, task is to predict the sentiment of the review. Take positive label as 1 as negative label as 0

   (a) (50 points) Implement the Naive Bayes algorithm to classify each of the articles into one of the given categories.The output should contain the predicted label of the review (positive or negative) in terms of 1 and 0 in a text file format. Report the accuracy over the training as well as the test set. Notes:

   I. Use Unigrams as features.
   II. Make sure to use the Laplace smoothing in Naive Bayes to avoid any zero probabilities. Use c=1.
   III. You should implement your algorithm using logarithms to avoid underflow issues.
   IV. You should implement Naive Bayes from the first principles and not use any existing python modules.

(b) (25 points) The dataset provided is in the raw format.This includes words such as of, the, and etc. (called stopwords). These words may not be relevant for classification. In fact, their presence can sometimes hurt the performance of the classifier by introducing noise in the data. Similarly, the raw data treats different forms of the same word separately, e.g., eating and eat would be treated as separate words. Merging such variations into a single word is called stemming.

Perform stopword removal and stemming before using naive bayes. You can use libraries (nltk for python) to perform stopword removal and stemming.

(c) (25 points) Feature engineering is an essential component of Machine Learning. It refers to the process of manipulating existing features/constructing new features in order to help improve the overall accuracy of the prediction task. For example, instead of using each word as a feature, you may treat bi-grams (two consecutive words) as a feature. Come up with at least two alternative features and learn a new model based on those features. Explain your best performing model in the report.