CS 6350: Project Milestone 1

Name: Tejas Ramesh Pawar — u1452462

March 7, 2024

- **What I did so far and Descriptive Statistics**

  After completion of the dummy submission, i.e. the project milestone 0. I downloaded the whole data and went through it on my local machine.

  I studied the whole dataset and its features. I did some preprocessing on the data, i.e. I went through the training, testing and evaluation files and calculated the dimensions of the dataframe, so that it would be easier for me to work on and employ classifiers that can get trained well within time and produce respectable results.

  The *bag-of-words* and *tfidf* feature sets have **10000** features, while the *glove* feature set has **300** features.

  - First, I tried working with ID3 algorithm. I used the code I wrote in HW1, and made some modifications for this data. But since, the number of features were too high in number, I decided not to go with decision tree classifier.
  - Next, I used the perceptron algorithm. I used the simple perceptron algorithm, where the update of *weights* and *biases* depend on the *learning rate* and the feature vector and label value.
  - I ran a 5 fold cross validation on each of the feature set and thus fiddled around with the *learning rate* and the number of epochs to train the perceptron depending on the training accuracy. After each change, I made a csv file in the submission format and uploaded on kaggle.
  - The submission I made on Kaggle has a learning rate of **0.01** and the number of epochs as **100**
  - I ran the same classifier on all the feature sets and chose the one with the highest test accuracy. Using the best *weights* and *biases*, I ran my classifier for the eval set.
  - The statistics of my run is as follows - The number of examples that predicted a label 1 are 3317 and the rest are 0. The score on kaggle was **0.68**.

- **What I am going to do next till the next Milestone?**

  I am going to work on the classifier and try to improve the accuracy. I will be using the miscelleneous data and improve the efficiency of my classifier.

  I will be using multiple feature set combinations and classifiers.

  I will be employing *k fold cross validations* on all the feature sets and choose k accordingly. This will help me choose the best hyper parameter for this data. I have only experimented with 5 fold cross validation till now, since, I am yet to choose the best hyperparameter after treating all the feature sets of the data.

  I will be making the next non-dummy submission before the next milestone.