Jackson Centeno / W1280823

Dr. Anastasiu

COEN140 Program 1

10/26/2020

As of the writing of this report, I hold rank 20 and best accuracy of 66%.

I personally did look at the leaderboard as I was interested in what results everyone achieved and if I could beat the best, as well as assess what acceptable results of accuracy are.

To perform dimensionality reduction, I first pre-process the data through the algorithm of c-mer's, which is used to create features from the data. It does so by slicing subsets of the original sample, while preserving the order, and makes each of these a feature. Next with this pre-processed data, I build a dictionary of a model based on the training data. To build this csr_matrix, we first identify all the non-zeros in the data, and start adding values and assigning indexes as we find new features. Next we normalize the data in regards to each row, such that we the computation for the prediction will no longer need the L2 norms, making the computation faster and also normalizing the data helps distribute the weights of the features. This sparsity is also addressed through the use of csr_matrices which leverage the identity of asymmetric attributes. Finally, we classify the sample using our established training data. To do so, we use lda as our method to fit our model.

My methodology to solve this for the best parameters was mostly used through cross validation instead of resubmitting to clp over and over. To do this, we split the training data (because the testing data has no class labels) into subsets such that we can test the accuracy. Sklearn was the library we used for our lda and cross validation. We split the training into 10 components

and went through it three times, then took the average of these accuracies to start to feel a range of our model.

I experimented with different types of cmers for preprocessing as well as the different components for dimensionality reduction. After iterations of testing, I found the best results for my model in using these.

There are no special instructions for my code, both train.dat and test.dat should be in the same directory as the script is the only requirement for this program.