CIS520: Implementation of Models

Francine Leech, Ziyin Qu, Chen Xiang

The four we implemented were a generative model, discriminative model, and an instance based method with the text data in words\_train.mat. We used a semi-supervised dimensionality reduction method with a GMM model to predict the color data in train\_color.mat. We used a 10-fold cross validation to calculate the average cross validation errors of the models.

**Generative Model**

Our generative model was a Naive Bayes Model. We used the Matlab function, fitcnb() and set the distribution within the model as multinomial distribution. The model had a 10-fold cross validation error of 0.1027.

To run the the Naïve Bayes model, open NB\_model.m and run the function. The function takes in train\_words.mat from the train\_set folder and outputs the 10-fold cross validation error and the model.

**Discriminative Model**

The discriminative model we fit was a cross-validated SVM classifier. The model classifies the tweet as 0 or 1 using fitcsvm(). The model has a 10-fold cross validation error of

To run the model open, … and run the file. The input of the model is train\_words.mat from the train\_set folder. The output of the function is the model the 10-fold cross validation error.

**Instance Based Method**

The instance based model is a K Nearest Neighbors model. We used the MATLAB function, fitcknn(), with K = 449 number of neighbors and the spearman distance. We use 10-fold cross validation to determine these specifications. The 10-fold cross validation error is 0.2798.

To run the model, open knnWords.m and run the file. The input of the model is train\_words.mat from the train\_set folder. The output is the model and the 10-fold cross validation error.

**Semi-Supervised Dimensionality Reduction**

We reduced the dimensionality of the train\_color.mat data with PCA using 30 principal components, a regularization value of 0.0001, and specified a diagonal covariance matrix with the pca() function. We ran a Gaussian mixture model to predict the reduced training data of the colors with two clusters with fitgmdist(). We picked two clusters to represent the two outputs we are trying to predict The 10-fold cross validation error was 0.4489.

To run the model open PCA\_GMM.m and run the function. The function takes in train\_words.mat and train\_color.mat from the train\_set folder and outputs the 10-fold cross validation error and the model.

Overall we found that the generative model had the lowest error, followed by the