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 training error of 0.80.

To run the model the Naive Bayes 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 had a training error of …

To run the discriminative model…

**Instance Based Method**

The instance based model is a k nearest neighbors model. We used 10-fold cross validation method to get the best number of neighbors (K = 449) and the distance (spearman). Use train\_words.mat to fit the model. The output of and get the accuracy 0.2829.

**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\_words.mat and train\_color.mat from the train\_set folder and outputs the 10-fold cross validation error and the model.