CIS520 Report: ML Crackers

Francine Leech, Ziyin Qu, Chen Xiang

December 12, 2016

1 Introduction

2 Preliminary Methods

- eveyrthing tyat didn't work - why it didnt work - image - Ziyin - PCA/GMM - Other ensemble methods methods

3 Main Methods

3.1 Naive Bayes

Accuracy

The second main

3.2 GentleBoost

The second main method we utilized was an ensemble method. We used GentleBoost, a weak learning that was built by MATLAB under the fitensemble function. The method combines many weak learners into one high quality ensemble predictor. We chose this ensemble methods over the others offered by MATLAB, because it is preforms well with binary classification trees with many predictors (ensemble citation).

The input of the model was the $word_train$ data. We used a 10-fold cross validation method to observed how the model preformed, specified the use 300 learners, and the type of learner as 'tree'. The average cross validation error was 0.21. The algorithm classified joy and sadness well.

The method could have improved if we increased the number of learners, however it would have taken a very long time to train because the data is large. Initially we tried the method with the default number of learners, 100 trees, and found that the cross validation accuracy only improved slightly. This slight improvement with triple number of learners reveals that the data has some intricacies or patterns that the ensemble method cannot learn.

3.3 Support Vector Machine

Support vector machines (SVMs) proved to the most promising method to classify the data. We used the MATLAB function, fitcsvm, to train an SVM model for binary classification on the on the $word_train$ data. We had a experimented on with this method a lot.

We tried a simple SVM by specifying a linear kernel. The cross validation error was 0.80. After experimenting with a variety of kernels, we found that the linear preformed the best. fitsvm allows you to make an assumption about the fraction of outliers in the data. While we could have gone through the raw tweets

and looked through the data, we decided to experiment with 5%, 10%, 20%, and 30%. The cross validation error 0.21 as we increased the outlier percentage.

Lastly we tried to optimize our SVM by using MATLABs built in method to optimize a cross-validated SVM using Bayes Optimization (citation). The method originates from The Elements of Statistical Learning, Hastie, Tibshirani, and Friedman (2009). Paraphasing from the MATLAB documentation, "the model begins with generating 10 base points for a "green" class, distributed as 2D independent normals with mean (1,0) and unite variance. It then generates 10 base points for a "class" that is also distributed as 2-D independent normals with mean (0,1) and unit variance. For each of the classes, it generate 100 random points by choosing a base point, b, of the respective color uniformly at random. It then generates an independent random point with 2-D normal distribution with mean b and variance I/5, where I is the 2-by-2 identity matrix. After 100 points for each of the colors has been generated, the point are classified using fitcsvm. The function bayesopt is used to optimize the parameters of the final SVM model with respect to cross validation."

4 Final Method

Ensembl: Sentiment Analysis 1 + Methods

Ensembl: Sentiment Analysis 2 + Methods dictionry wasn't as good

5 Discussion

6 Works Cited

References