CS571 NLP Project One Result Summary

Name: Jingzhi (John) Wang
Email: jwang67@emory.edu

Program Overview

The johnmallet program is consisted of three major classes, which included importData.java, training.java, and evaluation.java. In importData.java, the input files are sent into the pipelist and processed with various MALLET pre-defined processors to generate a feature vector. Then, the feature vector is split (90%:10%) into the training set and the testing set. A classifier is generated by training on the training set. At last, an evaluation of accuracy and precision is assessed by applying the classifier against the testing set.

Method Adopted

In johnmallet program, I recruited three commonly used methods to process the given Enron Spam corpus: Naive Bayesian Method, Maximum Entropy Method, and AdaBoost Method. Among these methods, Naive Bayesian method is the most computationally simple method. Maximum Entropy method requires a hill-climbing process to optimize the classifier. AdaBoost method assigns weights on weaker classifier trainer methods, such as Naïve Bayesian method, and adjusts by perturbation.

Accuracy Measurement

For accuracy measurement, all three methods achieved very high scores. Naive Bayesian method achieved accuracy rates of 98.892% for the training set and 98.991% for the testing set based on 10 observations. Maximum Entropy methods achieve even high accuracy rates of 99.947% for the training and 99.970% for the testing set. The AdaBoost method based on weaker Naive Bayesian method achieved a perfect accuracy of 100.0% for both of the training and the testing sets.

Precision Measurement

For precision measurement, again, all three methods achieved very high scores. Naïve Bayesian method achieved precision rate of 98.980% for the training set and 99.248% for the testing set based on 10 observations. Maximum Entropy methods showed improved precision rates of 99.916% for the training and 99.942% for the testing set. The AdaBoost method based on weaker Naive Bayesian method scored perfect precision of 100.0% for both of the training and the testing sets.

Time Consumption

For the time consumption, Naive Bayesian methods used the shortest time of 378 milliseconds. Maximum Entropy method spent 12259 milliseconds. Meanwhile, AdaBoost method used 23952 milliseconds. Average corpus loading and processing time is 22982 milliseconds.

Extra Comments

Training and testing on only ham or spam files resulted in uniformly perfect accuracy and precision of 100.0%. The hypothesis is that the number of files (the size of the corpus) is not big enough to generate substantial error rates.

Summary

Accuracy : AdaBoost > MaxEnt > NaiveBayesian
Precision : AdaBoost > MaxEnt > NaiveBayesian
Time Consumption : NaiveBayesian > MaxEnt > AdaBoost