Abstract

Many different machine learning techniques have been applied to the problem of classifying emails as spam. Among the most popular of these techniques are Naïve-Bayes Networks, SVMs, and KNN classification. In my project, I will forgo the traditional approaches and attempt to classify email with deep learning techniques. In initial experiments on deep learning for email classification (Sallab & Rashwan, 2012), a Deep Belief Network (DBN) averaged 87.24% classification accuracy, while SVM and KNN approaches averaged 70% and 63.86%, respectively. These findings suggest that deep learning can be a powerful approach to spam detection.

In classifying emails, the goal will be to maximize true positives while minimizing false positives, as false positives are basically unacceptable.

I will perform my classification using Caffe deep learning software. I have chosen Caffe at the professor's recommendation, because it is fast, free, modern, and well-documented. It will allow me to implement a deep learning network by writing code in either Python or C++, two languages with which I am familiar. If Caffe turns out not to work for whatever reason, I will instead use the DeepLearnToolbox or deepmat, which are two Matlab libraries for deep learning.

Datasets

Email classification is a topic of such interest to machine learning researchers that there exist many datasets of spam and nonspam emails. Though I have considered several, I
intend to use the Enron Dataset. This collection of emails is particularly interesting because of its size. The full dataset contains 50,000 spam messages and 43,000 non-spam messages. It includes all the emails sent and received by 150 individuals over many years. This is the same dataset used in many papers on email classification, meaning I will be able to compare the outcomes of my deep learning approach to the outcomes of other machine learning approaches performed on the same dataset. This will be a huge help in judging the quality of my model.

Bibliography

