CS5262 Project Proposal

Yizhou Guo

- Topic and Dataset

The project will focus on classifying email spam. The data set I am planning to use is Enron-Spam dataset from Athens University: http://nlp.cs.aueb.gr/software_and_datasets/Enron-Spam/index.html It is raw data of emails and thus requires more pre-processing to apply machine learning methods compares to the Spambase dataset available in UCI ML catalog, but it is more intense in information and I hope to develop the project further into my thesis paper.

- Methods

Currently, I am planning to focus on SVM to classify spam and Neural Networks to predict the probability of an email being spam. However, methods like clustering and multiple linear regression may be explored as well, and NLP methods will be applied in the future.

Due to the nature of the dataset being raw, tokenize will be needed. A common way is to use count of keywords, and I am planning to include some additional parameters, such as percentage of punctuation, count of words outside of frequently used English words, as well.

- Performance Evaluation

Like the author of UCI Spambase dataset said, we want to have spam emails detected and avoid classifying non-spam emails as spam at the same time. Therefore, performance will be evaluated by three metrics: detection rate, false detection rate, and learning efficiency (running time).

- Timeline

By February 28th: Process, tokenize, and visualize data.
By March 7th: Apply simple methods to learn the data using small datasets.
By March 24th: Complete SVM method.
By April 7th: Complete Neural Networks method.
By April 14th: Prepare for presentation.
By April 26th: Final write up.

- Related references

Machine learning for email spam filtering: review, approaches and open research problems
https://www.sciencedirect.com/science/article/pii/S2405844018353404?ref=pdf_download&fr=RR-2&rr=8527f073e8c11d6e
Spam Filtering with Naive Bayes – Which Naive Bayes?
https://www2.aueb.gr/users/ion/docs/ceas2006_paper.pdf