CS 584 : Machine Learning
Final Project Proposal

Haripriya Pandya (A20299643)
Sai Arjun Tanguturi (A20307238)

Problem Statement:
In this project we aim to analyze the Naive Bayes classifier as a spam filtering algorithm and compare variants of the algorithm for spam classification.

Approach:
Naive Bayes classifiers (NB) are one of the most common, yet accurate classifiers for commercial email service providers. However, there are various forms of NB classifiers available with variations incorporated in the original algorithm. In this project we aim to test them against a common dataset and compare their results. The different versions include,

1. Multi-variate Bernoulli NB
2. Multinomial NB with term frequency
3. Multinomial NB with Boolean attributes
4. Multi-variate Gaussian NB
5. Flexible Bayes

Naive Bayes classifiers are popular since they are very simple algorithm and implementation. They also provide a great time complexity of linear dimension and they deliver quite accurate predictions despite the simpler algorithm. However due to the simplicity of this algorithm, there are different approaches taken to improve the accuracy of the output. According to the observations made by

Multi-variate Bernoulli NB: Treats each email message as a set of tokens (unique),

Multinomial NB, TF attributes: Treats each email message as a bag of terms (it keeps track of repeating words and hence it relies on frequency of individual term)

Multinomial NB, Boolean attributes: Similar to TF attribute approach, but the features are boolean.
The authors mention that," it has been proven [7] that the multinomial nb withtf attributes is equivalent to a nb version with attributes modelled as following Poisson distributions in each category, assuming that the document length is independent of the category. Hence, the multinomial nb may perform better with Boolean attributes, if tf attributes in reality do not follow Poisson distributions."

Multi-variate Gauss NB: The multi-variate Bernoulli NB can be modified for real valued attributes, by assuming that each attribute follows a normal distribution in each category.

Flexible Bayes:
The Flexible Bayes algorithm relies upon computing the average of L normal distributions with different mean values. L is dependent upon the number of terms present in a certain category (ham/spam) in the training data.

An important consideration is time complexity, and the five Naive Bayes implementations listed above are O(m*N) during training, where N is the number of documents in the training dataset. We plan to implement these 5 different versions of NB, and determine which is the most efficient in terms of classification, as well as time complexity. We will compute the confusion matrices, and compare the classifiers through their precision, recall, f-measure, and accuracy.

Data:
For the purpose of this study we will use the Enron email dataset which contains emails that have already been classified as spam or not-spam(known in this dataset as ham). Specifically, we will use the version that has been processed(duplicates, virus emails, and messages containing non-Latin encodings have been removed) by The Athens University of Economics and Business. This dataset will give us a good idea of how these NB classifiers generalize to normal user email accounts (since this dataset is derived from the email accounts of former Enron employees).

References:

1) Metsis, Vangelis, Ion Androutsopoulos, and Georgios Paliouras. "Spam filtering with naive bayes-which naive bayes?." In *CEAS*, pp. 27-28. 2006.
2) Song, Yang, Aleksander Kołcz, and C. Lee Giles. "Better Naive Bayes classification for high-precision spam detection." *Software: Practice and Experience* 39, no. 11 (2009): 1003-1024.
3) ZHANG, Fu-zhi, Zhao-hui WU, and Fang YAO. "Research and improvement of spam filter technology based on Bayesian [J]." *Journal of Yanshan University* 1 (2009): 012.
4) Athens University of Economics and Business. (2006). *Enron-Spam Dataset* [Data files]. Retrieved from http://www.aueb.gr/users/ion/data/enron-spam/