Data Mining Assignment 1

Identify a problem from your own experience that you think would be amenable to data mining. For that problem describe:

Problem: Spam Filter

The junk folder in our email inbox. It is the place where emails that have been identified as spam by the algorithm. Spam emails are at best an annoying part of modern-day techniques, and at worst, an example of people phishing for the personal data.

1. What the data is.

Data is the essential ingredients before we can develop any meaningful algorithm. There are two types of data present namely, **ham** (non-spam) and **spam** data.

- Ham Data: This looks like a normal email reply to another person.

- Spam Data: One of the spam training data does look like one of those spam advertisement email in our junk folder.

2. What type of benefit you might hope to get from data mining.

This mainly focuses on the classification of spam E-mails using data mining techniques and Spam-Email data set collected. The purpose is not only to filter messages into spam and not spam, but still to divide spam messages into similar groups and to analyze them, in order to define the social networks of spammers.

3. What type of data mining (classification, clustering, etc.) you think would be relevant.

Various data mining techniques have been applied on email data. In this type of machine learning problem, I think K-Means clustering techniques or Naïve Bayes would be more relevant to be an effective way of identifying spam. The way that it works is by looking at the different sections of the email (header, sender, and content). The data is then grouped together. These groups can then be classified to identify which are spam. Including clustering in the classification process improves the accuracy of the filter to 97%. This is excellent news for people who want to be sure they're not missing out on your favorite newsletters and offers.

4. Name one type of data mining that you think would not be relevant, and describe briefly why not.

Linear Regression and KNN technique would not be relevant as it is numerical and usually used to measure future data and make certain predictions. In the last five years, people have started using stochastic gradient methods to avoid the noninvertible (overfitting) matrix problem. Switching to logistic regression with stochastic gradient methods helped a lot, and can account for correlations between words. So, in view of these circumstances, Naive Bayes is pretty impressive.