Introduction

In machine learning, a typical scenario has a set of features from which we wish to know about some data which can be either quantitative or categorical. This is done using training data from which we build a learner, which is then used for prediction of unkown or unseen data. This is an example of supervised learning. On the other hand in unsupervised learning, we only have the features and no measurement of the outcome; in this case the task is to identify some organization which may or may not be present in the data set.

Example: Email spam

Problem - Classification

Data set – 4601 emails categorized as spam or email.

Features – Relative frequency of 57 of the most commonly occurring words or punctuation marks. (Relative frequency is frequency of a given word in a given email divided by the total number of occurrences of that word in 4601 emails)

The learning method must choose which features to use (all of them is also a possibility), and how those features categorize the emails. Note that in this problem two types of error can occur:

- Categorizing emails as spam
- Categorizing spam as emails

The former of the two is a more serious error than the latter thus we would like our learner to have less error in the former category.