Apply Support Vector Machine to Text Classification

# Yang Zhang (11529139)

**Abstract**

This report analyzed the feasibility of classifying texts using Support Vector Machine. In this report, I will explain the reasons why SVM is suitable to solve the text categorization problem by exploring the properties of text classification problem. In the experiment part, empirical data are provided to support the theoretical ideas.

**Introduction**

Text classification or document categorization is a common problem in information science and computer science. The task is to assign a document to one or more classes or categories. The media not only limits to plain text, but also is expendable to images, sound tracks, etc. With the rapid growth of online information, text classification is increasingly important and widely used in search engine and media databases. People are already benefited from this technology, for example, they can read news by their interests without manually filtering. It will be advantageous to train a robust classifier from samples.

Naïve Bayes is a popular algorithm that commonly used for solving text classification problem. However, Naïve Bayes has its own limitation that cannot be easily overcome and there are noticeable benefits of applying support vector machines in this problem. (I will discuss this detailly in next part of the report). The approach of proving the feasibility of SVM is to compare the performance between Naïve Bayes and SVM on the same set of classification tasks. The result will be explored in the experiment part.

**Problem Setup and Theoretical Supports**

The main task of text classification is to assign a document with one or more predefined tags. In the matter of solving the problem by machine learning, the objective is to automatically learn classifiers from provided training examples. This is a kind of supervised learning problem, which SVM can be fitted in with general speculation. In fact, if thinking deeper, SVM should work well in text classification problem. To support the hypothesis, let’s see how merits of SVM integrate into the properties of general text classification.

Usually, the feature vector of text set is very large. The size could be more than dozens of thousands. While, such high dimensional input space is not a problem to SVM. Due to it not depends on the number of features, SVM is able to deal with such a feature vector with many features.