**Naive Bayes**

Pranaya Bhat (1001408965)

**About**

Naive Bayes classifiersare linear classifiers that are known for being simple yet very efficient. The probabilistic model of naive Bayes classifiers is based on Bayes’ theorem, and the adjective naive comes from the assumption that the features in a dataset are mutually independent. In practice, the independence assumption is often violated, but naive Bayes classifiers still tend to perform very well under this unrealistic assumption

**Bayes’ rule**

posterior probability=conditional probability \* prior probability / evidence

**Implementation**

I have used Bernoulli model in this assignment to implement Naïve Bayes. The data is divided into two equal sets namely training and testing sets. In the Pre-processing step, I am forming vocabulary using bag of words approach. The vocabulary excludes all the words that are listed in the stoplist.txt file and is then sorted in the alphabetical order.

Then I am converting the training data into a set of features, where I convert each article into a feature vector of size of M+1. For the first M slots, a 1 in the ith slot represents that the ith word in the vocabulary is present in the message and 0 represents the absence. The (M+1)th slot correspond to the classified class label where ‘A’ represents that the article is from class arxiv similarly ‘J’ and ‘P’ for jdm and pos respectively.

In the classification step the Bayes classifier reads the training data and learns the classifier used by classifier. In the testing phase, I convert the articles in the testing data into a feature vector of size M where everything remains the same as training data, except the (M+1) slot which is not there for the testing phase that indicates the class label.

The calculated Accuracy has been displayed.