**CSCI316 Big Data Mining Implementation and Techniques**

**Laboratory 5**

**Objective**

* Implement Naïve Bayes classifiers *from scratch*
* Implement common evaluation metrics

(Note: “Implementation from scratch” means “not relying on any pre-implemented machine learning libraries”, but libraries such as NumPy, Pandas and SciPy can be used.)

**Naïve Bayes classifier**

Review the Naïve Bayes classifier theory and implementation technique in Lecture 5. Develop a Naïve Bayes classifier as an email filter in Python. Namely, the classifier predicts whether emails are ordinary or adverts.

**Dataset**: Files “wordsList” and “classList” (available in the datasets folder of this assignment on Moodle)

The wordsList file contains 72 pre-processed emails. Each line is a list of words extracted from each email. The classList file contains the class labels that indicating whether the emails are ordinary or adverts (0 for ordinary and 1 for adverts).

Requirements

* Use stratified sampling to randomly select 66 out of 72 lines for training and the remaining 6 lines for test. Return the classification probabilities of these 6 records.
* The Naïve Bayesian classifier must account for multiple occurrences of words and implements techniques to overcome the numerical underflows and zero counts.
* Compute the TP/TN/FN/FP and plot the ROC for your classifier.