# CSCI/ARTI 8950 Machine Learning

## Assignment Number 5: Due Thursday 3/20/2008 **(in class)**

You have two options for this assignment. You should choose and do only one. No extra credit will be given for doing both.

1. **Option 1 [100 points]** Obtain at least **100** text documents belonging to at least two classes and apply the Naive Bayes algorithm for text classification to the documents. You should use ten-fold cross-validation and report the results. To obtain the documents you may use news groups, some web queries or any method of your choice as long as you clearly describe it. You can use the code provided by Tom Mitchell which can be downloaded from: **http://www.cs.cmu.edu/afs/cs/project/theo-11/www/naive-bayes.html** or, if you prefer, you can use a suitable Naive Bayes code of your choice or write your own. If you use a code other than Tom Mitchell's or write your own code, you should provide me with clear documentation so I can grade your assignment in a reasonable amount of time.

2. **Option 2 [100 points]** For this assignment you need to create or use a suitable **Naive Bayes** package. I recommend the WEKA package but there are many other packages available for download on the web or you can write your own code. You should choose a data set from the UCI repository (preferably the same one you used for the decision tree and/or the neural network homeworks so you can compare the performance) for this assignment (please don't forget to mention which one you used!). You should note that the algorithm described in the book applies only to discrete attributes. However, the extension to continuous attributes is not difficult (it is included in the WEKA package). You should train a Naive Bayes classifier and use ten-fold cross-validation and report the results. You should turn in your code if you wrote your own or just the name and web location of the package you used as well as the settings and any code modifications.