MIR Assignment 3. Fall 2006 (8 pts)

This assignment explores the use of the Weka software for machine learning. It consists of 4 questions (two points) each. They are relatively straightforward and therefore the assignment shouldn't take longer than 1 hour. An optional 2 point extra credit question is also included.

You will need the following datasets in the WEKA .arff format:

- `http://www.cs.uvic.ca/~gtzan/mir_course/Music_Sing_LPCC_.arff`

- `http://www.cs.uvic.ca/~gtzan/mir_course/Music_Sing_MFCC_.arff`

To answer the question you will need to download and install Weka from: `http://www.cs.waikato.ac.nz/ml/weka/`. For all questions you will be using the Explorer interface.

1. Load the Music_Sing_MFCC_.arff dataset. Click on the Visualize All option on the right hand side. If you were forced to use only 1 feature for classification which of the following two features would you choose: *Std_Mem40_MFCC_12* or *Mean_Mem40_MFCC_9*. Explain your answer briefly. What do the numbers (218, 418) mean in the last graph labeled output ?

2. Click on the Classify tab window. Select the ZeroR classifier and 10-fold cross-validation. The classification accuracy is 65.7233. Explain how this number is calculated (hint: look at the output graph of the previous question)

3. Select the bayes category NaiveBayesSimple classifier and 10-fold cross-validation. The classification accuracy is 87.1069. Based on the confusion matrix in the botoom what percentage of class_0 feature vectors are misclassified as class_1 ? Select use Training Set. The classification accuracy is 88.2. Explain why it is different than the cross-validation approach

4. Select the functions category SMO classifier and 10-fold cross-validation. What is the classification accuracy ? Which class (class_0 or class_1) is easier to classify ?

5. EXTRA CREDIT: Compare the two feature datasets: Music_Sing_LPCC_.arff and. Music_Sing_MFCC_.arff dataset in terms of classification accuracy using the NaiveBayesSimple classifier and the SMO classifier. Are the feature sets equivalent for this task ? Comment on the differences in accuracy between the datasets and classifiers.