BIOM/SYSC5405 – Pattern Classification and Experiment Design

Assignment 4— Due 11:25am Mon 13 Mar 2017

Please submit a single **PDF** file with all your answers, discussion, plots, etc. **on CULearn**.

Question 1: Weka

Download and install Weka to get a working Weka environment on your computer. Download the sample dataset assigData4.csv. This file has 1200 positive and 6000 negative samples. Each sample 215 features (assume the first feature is "feature 1" below). The final value on each line of the file is the class (1 = positive, 0 = negative) of that sample. These data represent features extracted from genomic windows that do (positive) and do not (negative) correspond to microRNA. Use Weka to do the following:

a) Data visualization:
   i. Load the data (Weka has a CSV importer where you can specify that there is no header line and that the 'last' attribute should be considered as nominal).
   ii. Plot the distribution of feature 15 for the two classes on a single histogram.
   iii. Plot a scatterplot illustrating the correlation between features 4 and 8, colouring the data by class.

b) Preprocessing: Weka implements filter type (i.e. not wrapper type) feature selection methods in two parts: a means to evaluate a set of features and a means to search for and build sets of features.
   i. Describe the *CFSSubsetEval* evaluator method (~50 words, don't just copy)
   ii. Name one supervised and one unsupervised evaluator method.
   iii. What is the difference between the *Ranker* search method and the *GreedyStepWise* search method?
   iv. Run one combination of a feature set evaluator and search method.
      i. Briefly describe which ones you selected and what parameters you used.
      ii. Summarize the results: how many features were selected and which features selected.

c) Classification1: using a naïve Bayes classifier:
   i. What parameters must be set by the user (briefly describe their meaning)
   ii. When creating a hold-out test set, what is stratified sampling and how is it applicable here?
   iii. For the original feature set (215 features): Conduct a hold-out test (70% train; 30% test; stratified sample). Provide the confusion matrix, the accuracy, the precision, the sensitivity, and the specificity. Generate a ROC curve and a precision-recall curve.
   iv. Repeat iii using one of your optimal feature sets from b) above.
   v. Which feature set led to the best performance? (discuss difference in observed performance metrics)

d) Classification2: using a K-NN classifier:
   i. What is the K parameter? What other parameters must be selected? (briefly describe their meaning)
   ii. Using 5-fold cross-validation, perform a parameter sweep of K={1,3,5,13}.
   iii. Which value of K works best when optimizing overall accuracy?
   iv. What value of K works best when optimizing the precision/PPV?
   v. Plot the precision-recall curve for the value of K that optimizes precision.