**CSI5387: Data Mining and Concept Learning, Winter 2012**

**Assignment 1**

**Due Date: Tuesday February 7, 2012**

Here is a list of themes that will be explored by this assignment:

* **Learning Paradigms:** Decision Trees, Multi-Layer Perceptrons (MLP).
* **Techniques:** Parameter Optimization
* **Evaluation:**
* **Evaluation Metrics:** Accuracy, AUC
* **Error Estimation:** 10-fold CV, 10x10-fold CV, leave-one-out, bootstrapping, permutation test
* **Statistical Testing:** 2-matched samples t-test, McNemar’s test

The purpose of this assignment is to compare the performance of Decision Trees and multi-Layer Perceptrons (MLP) on a single domain, the [UCI Sonar Domain](http://archive.ics.uci.edu/ml/datasets/Connectionist+Bench+(Sonar,+Mines+vs.+Rocks)), in a systematic fashion. There are five parts to this assignment:

1. Use the [WEKA](http://www.cs.waikato.ac.nz/ml/weka/) toolbox or any other series of tools that you prefer to run C4.5 and MLP on the Sonar data.[[1]](#footnote-1) Both algorithms can have various parameter settings (this is particularly true of MLPs, for which you need to find the appropriate number of hidden units given your domain, and possibly a good learning rate and momentum). Experiment with various parameter options for both classifiers until you settle on the one that gives you the best results. Report your results according to both the Accuracy and AUC measures.
2. If you used WEKA in Question 1, the chances are that the re-sampling method you used was 10x10-fold Cross-validation (CV). In this section of the assignment, I am asking you to experiment with other re-sampling methods. Namely, use 10-fold CV, 10x10-fold CV, Leave-one-out, Bootstrapping and the permutation test. All these re-sampling methods and their implementation in R are discussed in the textbook entitled [Evaluating Learning Algorithms](http://www.amazon.ca/Evaluating-Learning-Algorithms-Classification-Perspective/dp/0521196000). Argue for or against each of these testing regimens in your specific case. [Please, note that if some of the re-sampling schemes are too time-consuming (this could be the case, in particular, for the MLPs used with leave-one-out, bootstrapping or the permutation test), feel free to drop them and discuss the problem you’ve encountered with them.]
3. Since you are working with a single domain, two statistical tests are appropriate to consider: the t-test and McNemar’s test. When using the t-test, if you found statistical significance in the difference in performance between the two classifiers, then verify that this difference is practically significant by computing the effect size, using Cohen’s d statistics. Again, discussions of these tests and their implementation in R are discussed in the textbook entitled [Evaluating Learning Algorithms](http://www.amazon.ca/Evaluating-Learning-Algorithms-Classification-Perspective/dp/0521196000).
4. Was the strategy of looking for good parameters in 1) and then comparing the two classifiers’ performance in 2) and 3) a good one? Argue your case. If it was not a good strategy, then explain what a better strategy might be and how practical it would be to use. Discuss the overall issue.
5. Do you have any additional comments on the way in which the evaluation was structured in this homework? What could be done better? In what way would it be better (easier, would yield more reliable results, etc…)?

1. Though you may prefer WEKA which is easy to use, it might be a good idea to get used to working in R altogether since many of the evaluation tools you will need to use have been implemented in R. Note that there is a WEKA package in R called [RWeka](http://cran.r-project.org/web/packages/RWeka/index.html). There also are other implementations of various learning algorithms in R. Download a good description of these available packages [here](http://cran.r-project.org/web/views/MachineLearning.html). [↑](#footnote-ref-1)