CS771:Machine learning: tools, techniques, applications
Assignment #1: Decision trees and random forest

Due on: 30-8-2013, 5pm 22-8-2013
MM:160

1. In this assignment you have to use the cardiac arrythmia data set available from: http://archive.ics.uci.edu/ml/machine-learning-databases/arrhythmia/

   (a) Build a DT classfier using 'grow maximum possible and then prune 'algorithm to construct the tree. Experiment with at least two different impurity functions.

   (b) Build a random forest classifier and find the right number of trees to include in the forest by a binary search approach to K, the number of DTs in the forest. The randomization should use bagged learning sets and a random selection of m features to find the split.

   (c) Compare the performance of the classifiers in 1a and 1b by doing 5-fold validation to find an estimate of the error.

   (d) The data set has missing data. Experiment with two ways to handle missing data during tree construction.

   (e) Find the three most influential features for classfication using the random forest classifier in 1b.

   (f) Compute an estimate of the error using the out-of-bag data for the classifier in 1b.

   (g) Find a lower bound on the random forest classifier error by estimating the strength s and the average correlation $\overline{\rho}$ and using the formula for the error bound.

   The data set has two files: arrythmia.data which contains the data and arrythmia.names which describes the data. This is a typical medical diagnosis problem where we expect the DT and random forest approach to work well.

   You can use the Weka library or any other decision tree and random forest construction library/toolkit (barring Matlab or any priced software).

   [40,40, 10, 20, 20, 10, 20=160]