CS771, CS771A:Machine learning: tools, techniques, applications
Assignment #2: Classifier ensembles: Random forests

Due on: 1-2-2015, 23.00 23-1-2015
MM: 180

This assignment is on random forests. The data set you have to use is the letter recognition data set: https://archive.ics.uci.edu/ml/datasets/Letter+Recognition.

a) As discussed in class build a random forest classifier using bagging and random splitting using m random features to decide the split at each node. Find the number of trees at which the error levels off by using a binary search between 2 trees and 400-500 trees. Plot 5-fold cross-validated error rates against the number of trees in the forest and report the number at which error levels off.

b) Compute and report the *out-of-bag* error for the forest with the least error in part a).

c) Experiment with different values of m (say 1, 2, 4, 8) and report 5 fold cross-validated error rates in the form of a table fixing the number of trees at 1.25 times the number obtained in part a) (where the error levels off).

d) Study the effect of the size of the randomly sampled data set from $\mathcal{L}$ while constructing a tree. Start by sampling 10% of the points from $\mathcal{L}$ for constructing a tree and go up to 80% in increments of 10% find the number of trees, say T, in the forest at which the error levels off in each case. Find the the 5 fold cross-validated error for a forest with T trees and plot it against the size of the sampled data set expressed as a percentage of $\mathcal{L}$. Comment on the results and say whether bagging is justified as a randomization method to select samples?

[60,20,50,50]