Homework 1, due September 2nd, 11:59pm

August 25, 2020

Use a programming language or package where decision trees and random forests can be trained and applied. Examples include Matlab, Python (scikit-learn package), or R.

1. Using the training and test sets specified in the syllabus, perform the following tasks:

   a) On the `madelon` dataset, train decision trees of maximum depth 1, 2, ... up to 12, for a total of 12 decision trees. If your package does not allow the max depth as a parameter, train trees with $2^1, 2^2, ..., 2^{12}$ nodes, again a total of 12 trees. Use the trained trees to predict the class labels on the training and test sets, and obtain the training and test misclassification errors. Plot on the same graph the training and test misclassification errors vs tree depth (or log2 of nodes) as two separate curves. Report in a table the minimum test error and the tree depth (number of nodes or splits) for which the minimum was attained. (3 points)

   b) Repeat point a) on the `wilt` dataset, with maximum tree depths d from 1 to 10 (i.e. $2^1, ..., 2^{10}$ nodes). (2 points)

   c) On the `madelon` dataset, for each of $k \in \{3, 10, 30, 100, 300\}$ train a random forest with k trees where the split attribute at each node is chosen from a random subset of $\sim \sqrt{500}$ features. Use the trained trees to predict the class labels on the training and test sets, and obtain the training and test misclassification errors. Plot on the same graph the training and test errors vs number of trees k as two separate curves. Report the training and test misclassification errors in a table. (3 points)

   d) Repeat point c) on the `madelon` dataset where the split attribute at each node is chosen from a random subset of $\sim \ln(500)$ features. (1 point)

   e) Repeat point c) on the `madelon` dataset where the split attribute at each node is chosen from all 500 features. (1 point)