Homework 2, due September 12th, 11:59pm

August 30, 2018

1. Use a programming language or package where random forests can be trained and applied. Examples include Python (scikit-learn package), R and Matlab. Using the training and test sets specified in the syllabus, perform the following tasks:

a) 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. (4 points)

b) Repeat point a) on the `madelon` dataset where the split attribute at each node is chosen from a random subset of $\sim \ln(500)$ features. (2 points)

c) Repeat point a) on the `madelon` dataset where the split attribute at each node is chosen from all 500 features. (2 points)