Alvin Chou

11-676

Assignment 3

In this assignment, a random forest algorithm is added on top of the existing decision tree structure. The setup of the process is as follows:

* The core data consists of 4 features and 50,000 instances. Each instance is a 1-minute interval. Each feature is the change in bid price of the previous 1-minute interval of one of the four unique stock tickers used.
* When the price change is 0 or positive, a label of 1 is used. Otherwise, the label 0 is used.
* Each individual instance is stored in the <Record> format, with features as ArrayList<double>, a timestamp as <string>, and a label in <int>.
* When building the decision tree, each node consists of the data table in ArrayList<Record> format
* The information gain on each feature on each node is calculated and the instances are split on the median value of the feature with the highest information gain.
* The tree continues to build and features are reused until the instances are completely split.
* When building the forest, 1000 iterations are used.
* For each iteration, only 2/3 of the randomly selected instances are used as well as sqrt(features), in this case, 2.

From the observations, individual decision trees achieve an accuracy range between 50-54%. However, as part of the random forest, the forest as a whole did not achieve a higher accuracy. This could be due to the lack of features (2 in each tree) used in building the trees. In addition, it is observed that runtime for the random forest increases exponentially since for every new tree added, a test set of data is used to evaluate the forest, running the test set through every existing tree. For improvements, addition features will be added to future random forests as well as using greater time intervals (e.g. 5-minute or 15-minute).