CS534 — Implementation Assignment 3 — Due 11:59PM November 12th, 2016

General instructions.

1. The following languages are acceptable: Java, C/C++, Matlab, Python and R.
2. You can work in a team of up to 3 people. Each team will only need to submit one copy of the source code and report. You need to explicitly state each member's contribution in percentages, i.e., for each group member provide a number xx% - percentage of the total project she/he is responsible for. Note that all team members are expected to contribute equally to the assignment. The individuals whose contribution is significantly lower than expectation will receive penalty to the assignment grade.
3. Your source code and report will be submitted through the TEACH site

`https://secure.engr.oregonstate.edu:8000/teach.php?type=want_auth`

Please clearly indicate your team members' information.
4. Be sure to answer all the questions in your report. You will be graded based on your code as well as the report. In particular, **the clarity and quality of the report will be worth 10 pts**. So please write your report in clear and concise manner. Clearly label your figures, legends, and tables.
5. In your report, the results should always be accompanied by discussions of the results. Do the results follow your expectation? Any surprises? What kind of explanation can you provide?

Decision trees and Random forest (total points: 50 + 10 pts)

In this assignment you will implement (1) the decision tree learning algorithm; and (2) construction of the random forest (using feature Bagging and bootstrapped sampling) with the modified decision tree learning algorithm from part (1) as the base learner. You will test your implementation on the IRIS data sets, which is a tree-class classification problem, with 4 continuous features. You will train your classifiers using the iris-train.csv file and test on the iris-test.scv file. First 4 columns provide feature values and the last column provides a class label:

(a) sepal length in cm

(b) sepal width in cm

(c) petal length in cm

(d) petal width in cm

(e) class: Iris Setosa (class 0), Iris Versicolour (class 1), Iris Virginica (class 2).

In particular, you need to do the following:

1. (15 pts) Implement a decision tree learning algorithm.

(a) Implement a decision tree learning algorithm, using the number of instances at leave node as a stopping condition. That is, if the number of examples at a node is less than k, one must stop and turn it into a leave node. Note that, since the features are continuous, for every node of the tree, you should compute threshold θ that gives you the best information gain.

(b) Please report the information gain of each threshold and each feature (recall that you only need to compute information gain when class label changes) for the root node.

(c) Report (plot) training and testing errors (i.e., the percentage of correctly classified examples) versus parameter k.

(d) What effect does k has on training and testing accuracy of the tree?

2. (35 pts) Implement random forest by constructing an ensemble of decision trees. In particular you should do the following:

(a) Modify tree learning algorithm (from part 1) such that, at each candidate split in the learning process, it selects a random subset of the features (please select 2 random features out of 4). This process is sometimes called "feature bagging".

(b) Then, using modified learning algorithm, build a tree on a randomly (with replacement) drawn subset of your training data (the size of each of such randomly drawn subset is equal to the size of your original training data). This technique is called "bootstrap aggregating" or "bagging".

(c) Repeat this procedure L times (thus, you will have L decision trees). Please, consider the following values of L: 5, 10, 15, 20, 25 and 30. Note that prediction for any samples can be made by outputting the class that is the mode of the classes predicted by each of L trees (i.e., majority vote).

(d) For every value of L (5, 10, 15, 20, 25 and 30) plot the training and testing errors (i.e., the percentage of correctly classified examples) versus values of the parameter k. Note that for bagging, due to its stochastic nature, different random runs may lead to different results. To increase the robustness of the results, please show the average error rates over 10 random runs. Your plots should be clearly labeled with easy-to-read legends.

(e) What trend do you observe in terms of the accuracy on the training and testing data respectively as we increase the number of trees in the ensemble? How random feature selection affects the accuracy of classification? What effect does k has on training and testing accuracy?

Note, you need to submit your source code of your implementations and your report.
Do not forget to include group members and indicate project contribution for each member in percentages.