BUAN 6341
APPLIED MACHINE LEARNING
Assignment 2
Due date: October 12, 11:59 pm

In this assignment, you are required to implement three learning algorithms:

1. Support Vector Machines (SVM)
2. Decision trees
3. Boosting

You must use R or Python for this assignment. You **can** use **any** publicly available R or Python library/package. You can even use H2O.

Tasks:

1. You are to use two data sets for this assignment.
    a. First data set: Use the data set from assignment 1 as the first data set. Convert it to a binary classification problem by thresholding the output to a class label. Use this transformed data set for all the tasks and experiments in this assignment
    b. Second data set: Get a data set suitable for classification from anywhere (either publicly available or your own). The data set should have a reasonable amount of features and instances. You need to explain why you think this data set and the corresponding classification problem is interesting.

    Divide your data sets in train and test sets.

2. Download and use any support vector machines package for your classification problems. Do it in such a way that you are able to easily change kernel functions. Experiment with at least two kernel functions (in addition to linear) of your choice. You can pick any kernels you like (shown in the class or not).

3. Download and use any decision trees for your classification problems. Experiment with pruning. You can use information gain or GINI index or any other metric to split on variables. Just be clear to explain why you used the metric that you used.

4. Implement (or download) a package to use a boosted version of your decision trees. Again, experiment with pruning.

5. Implement and use cross validation for both your datasets with the above algorithms.

Deliverables:
You are required to submit the following:

- Your code file(s)
- A readme file explaining how to run your code
- Report (must not exceed 10 pages total)
- Any supporting files (data set, etc. If the data set is too large, submit the url)

Your report should be both thorough and concise and contain at the very least the following:

- Description of your classification problem / data sets, and why you find them interesting.

- Error rates (train and test) for various algorithms on the two data sets. Plot various types of learning curves that you can think of (e.g. but not limited to – error rates vs. train data size, error rates vs. clock time to train/test, etc.).

- Performance comparisons (learning curves, confusion matrices, etc.) of various functions/parameters for the algorithms (e.g. kernels in SVM, pruning in decision trees, etc.) on both the data sets.

- Comparisons of the three learning algorithms using the two data sets.

Be creative and think of various experiments that you can come up with for this assignment. You need to give clear description of all your experiments and analysis. Why did you get the results that you did? Were these results good or bad because of these two specific data sets? Compare and contrast the different algorithms. What additional things can you do to get better results? How did cross validation help? How did you do pruning and when and why did you decide to stop? How did you pick various kernels, and how did they compare? Which algorithm performed the best and why? How did you define best? Think of as many questions as you can! This assignment will take time. So get started on it today!

Grading:
Total weightage: 12.5% of final grade

Breakdown:

Code: 0 points (Code should execute and produce the results presented in the report with minimum effort – We will run the code and if it doesn't run or has errors, points will be deducted from the report).

Report: 100 points
Points will be awarded based not only on how good your results are, but also on how well you describe them as well as underlying experimentation. Any plots without explanation = 0 points. Similarly, explanation without plots = 0 points. Keep in mind that you are graded on your analysis and description, as well as creativity. Have fun!