Homework 5, due February 9th, 11:59pm

This project could be worked on using sklearn or other Python packages. Alternatively, you could download and install the WEKA library from

`http://www.cs.waikato.ac.nz/ml/weka/`

The program is in Java, so it runs on any platform. Preferably download the kit that includes the Java VM. If you have a 64 bit machine, download the 64bit version since it can use more memory. For the experiments, you could use the Weka Explorer since it has a nice GUI.

1. Use the `satimage` dataset from Canvas to compare a number of learning algorithms. For Weka you might have to modify the files to make them compatible with Weka as follows:

- Add a header row containing the variable names (e.g. X1, X2, ... Y)
- Change the class labels from numeral (1,2,3,4...) to literal (e.g. C1, C2, C3...)

Train the following models on the training set and use the test set for testing. Report in a table the obtained misclassification errors on the training and test sets and the training times (in seconds) of all algorithms.

a) A decision tree classifier. (1 point).

b) A Random Forest with 100 trees and one with 300 trees. (1 point).

c) Logistic Regression. (1 point)

d) Naive Bayes. If there are multiple versions try all that can be applied ti the data and choose the one with the smallest test error. (1 point)

e) Adaboost with 30 weak classifiers that are decision trees (not stumps), and one with 100 trees. (1 point)

f) LogitBoost or GradientBoost with 30 regression stumps, and one with 100 stumps. (1 point)

g) LogitBoost or GradientBoost with 30 regression trees. (1 point)

h) An SVM classifier (named SMO in Weka) with RBF kernel. Try different combinations of the parameters C and γ to obtain the smallest test error. Report in a table all parameter combinations that your tried and the train and test errors you obtained. Note: You should be able to obtain one of the smallest test errors among all these methods. (2 points)

i) Based on the above results, which learning algorithm would you recommend to use on this dataset? Take into consideration the training and testing time, and the ease of parameter tuning. (1 point)