

# Homework 5, due October 3rd, 11:59pm

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. In `runweka.ini` change the heap size to at least 1024mb otherwise you will run out of memory. For the experiments, you could use the Weka Explorer since it has a nice GUI.

1. Use the `covtype` dataset from Blackboard to compare a number of learning algorithms. Split the data into training and test sets as specified in the syllabus (training set contains first 11,340 +3,780 observations, test set contains the remaining 565,892 observations). You will have to modify the files to make them compatible with Weka as follows:

- Add a first 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 (J48). (1 point).
- b) A Random Forest with 100 trees and one with 300 trees. (1 point).
- c) Logistic Regression. (1 point)
- d) Naive Bayes. (1 point)
- e) Adaboost with 20 weak classifiers that are J48 decision trees, and one with 100 trees. (1 point)
- f) LogitBoost with 10 decision stumps, and one with 100 stumps. (1 point)
- g) LogitBoost with 100 stumps and weight trimming (pruning) at 95%. (1 point)
- h) LogitBoost with 25 M5P regression trees. (1 point)
- i) An SVM classifier (named SMO in Weka). Use an RBF kernel and try different parameters to obtain the smallest test error. Report the parameters that gave the smallest test error. Note: You should be able to obtain one of the smallest errors among all these methods. (1 point)
- j) Using the misclassification error table, draw a scatter plot of the test errors (Y) vs log training times (seconds, on X axis) of all the algorithms from above. (1 point)