CptS 315: Introduction to Data Mining
Empirical Analysis 1 (EA1)

You will use the Weka: `http://www.cs.waikato.ac.nz/ml/weka/downloading.html` software. You can use the Graphical Interface to answer all the questions below – It is easier. Weka employs the ARFF (https://www.cs.waikato.ac.nz/ml/weka/arff.html) format for datasets. All the specific details provided below are for Weka.

You can also use Scikit-learn `http://scikit-learn.org/stable/` software if you are more comfortable with Python.

- Bagging (weka.classifiers.meta.Bagging). You will use decision tree (weka.classifiers.trees.j48) as the base supervised learner. Try trees of different depth (1, 2, 3, 5, 10) and different sizes of bag or ensemble, i.e., number of trees (10, 20, 40, 60, 80, 100). Compute the training accuracy and testing accuracy for different combinations of tree depth and number of trees; and plot them. List your observations.
- SVM Classification learner (weka.classifiers.functions.supportVector). You will run the SVM classifier on the training data to answer the following questions.

  (a) Using a linear kernel (-t 0 option), train the SVM on the training data for different values of C parameter(-c option): $10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}$. Compute the training accuracy, and testing accuracy for the SVM obtained with different values of the C parameter. Plot the training accuracy and testing accuracy as a function of C (C value on x-axis and Accuracy on y-axis) – one curve each for training, validation, and testing data. List your observations.

  (b) Repeat the experiment (a) with polynomial kernel (-t 1 -d option) of degree 2, 3, and 4. Compare the training and testing accuracies for different kernels (linear, polynomial kernel of degree 2, polynomial kernel of degree 3, and polynomial kernel of degree 4). List your observations.

1. Please use the “voting” dataset provided for this question in ARFF format. Please use the last 100 examples for testing and the remaining examples for training.
2. Please do the same thing using the “ionosphere” dataset. Please use the last 25 percent examples for testing and the remaining examples for training.