Empirical/Programming Assignment 1

Due date: Wednesday, September 17 (electronically by 12:00 noon and in hardcopy in class)

1 Introduction

This is a simple assignment whose main goal is to introduce you to the weka system and make sure you are comfortable using it. You will test the sensitivity of the \textit{k nearest neighbors} algorithm and the \textit{decision tree learning algorithm} to existence of irrelevant features.

In this assignment we will use the EEG Eye State Data Set available from the UCI repository\(^1\) which we have subsampled and preprocessed for the assignment. In particular the original dataset has 14 features for each example. We have made multiple versions of the data with 14, 24, 44, 54, 64, 74, 84, 94 features respectively by adding random irrelevant features beyond the original 14 features. Each dataset is split into a training portion and test portion for use in experiments.

2 Your Tasks

2.1 Sensitivity to Irrelevant Features

Write a script to run the default versions of the IBk and J48 in weka on all datasets and plot the test set accuracy as a number of features (where use the options -t and -T to specify train and test files respectively). What can you conclude regarding algorithm sensitivity in this case?

2.2 Learning Curves

In this part you should investigate the sensitivity as a function of training set size. We will focus on the original (14 features) and the 54 features datasets. For this assignment please use the following simple pseudocode to produce the learning curves:

1. Let the initial data file be \textit{train.arff} and \textit{test.arff}.

2. Repeat 10 times:
   (a) for \(i \in \{50, 100, \ldots\}\)
      i. Let \textit{traini.arff} be a random subset of \textit{train.arff} with \(i\) examples (e.g. \textit{train50.arff}).
      ii. Run the learning algorithm on the corresponding train/test combination and record the accuracy on test data.

3. For each algorithm and each training set size calculate the average and standard deviation in performance over the 10 trials.

Then plot performance (accuracy and standard deviations) as a function of training set size. This makes 4 curves, for two algorithms times 2 datasets, which you should put together in the same plot. What can you conclude regarding algorithm sensitivity in this case?

\(^1\)See http://archive.ics.uci.edu/ml/datasets/EEG+Eye+State
3 Data

The data files are accessible through the course web page.

4 Submitting your assignment

• You should submit the following items both electronically and in hardcopy:
  (1) All your code for data processing and the experiments (please write clear code and document it as needed),
  (2) A brief report with the plots from the two parts and your observations from these plots.

• Please submit a hardcopy in class.

• Please submit electronically using provide by noon of the same day: Put all the files from the previous item into a zip or tar archive, for example call it myfile.zip. Then submit using provide comp135 a1 myfile.zip.