RMIT University
School of Science
COSC2110/COSC2111 Data Mining
Laboratory Week 2

Aims of this lab

- Learn about the importance of randomizing training and test data.
- Learn how to apply a number of classifiers to a data set and interpret the results.

---

1. You will need to have access to the WEKA package, see week 1 lab sheet.
2. The data files for this lab can be found at
   `/KDrive/SEH/SCSIT/Students/Courses/COSC2111/DataMining/data`
3. Start Weka and load the file: `../data/arff/UCI/iris.arff`.
   Run the IBK classifier with 66% split.
   (a) From the output, find the number of examples in the training and test sets?
   (b) What is the test error rate?
4. Now load the file `../data/lab01-iris-train.arff`
   (a) Select the IBK classifier. From test options select 'Supplied test set' and for the test file use `../ data/arff/lab01-iris-test.arff`.
   (b) You will see that the training file has 100 examples and the test file has 50 examples.
   (c) What is the error rate of the IB1 classifier?
   (d) Explain the difference in error rates. You may want to inspect the files with an editor.
5. Repeat the above exercise with OneR. Is there any difference?
6. Load the file `../data/arff/UCI/splice.arff` into Weka.
7. Select the ZeroR classifier from Rules. Right click on "ZeroR" and follow the help windows to determine what this classifier does.
8. Run the classifier, what is the accuracy?
9. Run the IBK classifier with default parameters. What is the accuracy?
10. Run the classifier for increasing values of K. Does there seem to be an optimal value for K?. [You might want to plot Accuracy vs K.]
11. IBK has a number of parameters. Explore the effect of changing "distance weighting" and "nearestNeighbourSearchAlgorithm". Summarise the effect of each choice.
12. Now apply the OneR classifier to splice.arff with default parameters.

13. Explore the effect of changing bin size.
14. What do you conclude from the results of ZeroR, IBK and OneR.
15. Choose some other files from the `../data/arff/UCI/` directory and investigate the performance of these three classifiers.