

## 08338 ACW Labs

This lab sheet follows on from the first ACW lab - the page and point numbering reflects this.

### General Advice

Don't get hung up on the names I give to any of the worksheets in these lab sheets – derive your own system that works for you. As a rule of thumb, if you want to modify data in a worksheet, copy the worksheet, give it a suitable name and make one kind of edit. If you are going to make another class of edit, repeat this copy and rename process.

It is also worth having an empty data template (with just the attribute names, and the top row frozen) for the worksheets where you are pasting data from other sheets. Again you can copy the template, rename and then paste the data in it.

### Overview

This sheet starts from Item 10 from the previous sheet and proceeds to cover new items. You are expected to have completed at least step 8 from the previous week before this sheet. Make sure you complete the steps in Sheet 1 before continuing with this sheet. Note in Lab sheet1 Step 9 was hints and step 10 was advanced work.

### ACW Labs: Step by Step: Final Data Description and Classifier Performance

11. All the following steps assume that you have clean data and it is in the same case (e.g. lowercase text); and ready to use in weka. Ensure that **FinalDataDescription** (describing just the clean data) is as complete as possible (see lecture DMDS-6-ACW2015-16 for initial help)
  - a. Use J48 on your clean data where nominal values are in the same case. Save the output from this in a suitably named worksheet.
  - b. This data acts as the baseline clean data for classifier performance and advanced data repair techniques (in later labs).
  - c. Check in Excel and weka that this data is consistent and complete.
12. In Excel start to complete the Classifier Performance worksheet. This will be the basis for ACW Stage3.
  - a. Ensure your Performance worksheet includes at least the following
    - i. Classifier for the name of the Classifier being used
    - ii. Data for the name of the worksheet containing the data used
    - iii. RMSE for Root mean squared error for Classifier (given in weka)
    - iv. Accuracy for Classifier Accuracy (given in weka)
    - v. Columns for TP, FP, TN, FN
    - vi. Sum (of TP, FP, TN, FN). Check that this agrees with number of records used by the classifier.
    - vii. Sensitivity for Classifier Sensitivity (you need to calculate this using given formula)
    - viii. Specificity for Classifier Specificity (you need to calculate this using given formula)
  - b. Enter the values for the j48.
13. Now choose another Classifier (from the Classifier → Bayes menu). Run on the same data and save the results. Then add its performance figure to the Excel Table.
14. Use one of the Neural Nets (from the Classifier → Function menu) on the same data. I recommend you use one of three Neural Nets (MultiLayerPerceptron, SMO or RBFNetwork). The lecture notes (to be supplied later) mostly support SMO, RBF and MultiLayerPerceptron.
15. Now choose yet another Classifier (not a Tree or Neural Net or Bayes). Run on the same data and save the results. Then add its performance figure to the Excel Table. , I recommend you actually use a few different

## 08338 ACW Lab Advice – Sheet2

classifiers (JRip, Ridor, NNGe and PART). Then decide which is best for your data (and steps in following labs). This classifier should produce Rulesets of use in stage 5 of the ACW.

16. You can use the Performance Table to identify the best classifier
17. **Advanced.** You can try other classifiers to see if you get a better performance.
18. **Advanced.** You can look to improve the performance of the classifiers by optimising the classifier parameters for any given data or by improving the data (later labs). This is how you gain full marks for stage3 of the ACW. If you feel confident you try this yourself, or ask in the labs or wait until after the lectures on specific classifiers. If you start to modify Classifier Parameters you need to note this in the Performance Spreadsheet.

Questions via lab, lectures Forum or email