CSM6420/SEM6420 – Machine Learning for Intelligent Systems

Second Assignment - Walkthrough using WEKA, April 2016

1 Using the WEKA Explorer

Load the dataset *train_risk.arff*. Use only the training set and cross validation (CV) on this dataset to evaluate the classifiers. In order to reduce computation time you should use 5-fold cross validation (the default setting is 10).

1. To explore the dataset use the visualization functions in WEKA explorer. Note that WEKA allows you to see details of an individual data point: Open the dialog that shows the scatter plot for two attributes and click on the respective data point. Another useful feature in this dialog is the *Jitter* slider bar: Jittering the data points is helpful if many data points lie very close together or on top of each other in the scatter plot. Write down any salient observations. You could also apply dimensionality reduction techniques on this dataset using WEKA's attribute selection function.

2. Next, train and evaluate different classifiers on the dataset using 5-fold CV. Try Naive Bayes (NaiveBayes) and a Decision Tree (J48) at least. Compare the results, e.g. accuracy and area under the ROC curves (AUC), for the two different classifiers. What do you notice? What is a reasonable baseline against which to compare the classification performance? Relate the classification performance to your observations in part (1). Keep the result buffer for the Naive Bayes classifier, you will need it in question (4).

   HINT: Simply re-start a classifier if you receive a message "not enough memory".

3. Based on what you found out about the data in the previous questions, preprocess the training set as you see fit.

   HINT: The filter function could be useful for this purpose (there are both supervised and unsupervised instance/attribute filter in the Preprocessing tab).

4. Retrain and evaluate the classifiers that you have been using in (2) with 5-fold CV. What is their performance on the modified dataset? Compare the parameters of the model learned by the Naive Bayes classifier for the two datasets. Is there a salient difference?

5. You can build the final classifier using explorer. To make the prediction on the test cases, you need to change the Output predictions options in the More options dialog to 'csv' (the default setting is 'Null'). Upload the test data, and right-click on the respective result item for classifier building, then selecting "Re-evaluate model on current test set" will output the predictions on test set as well. Note: the statistics will be useless due to missing class labels in the test set, so just ignore them.

   HINT: Make sure that the test set used for model prediction contains all the features that you used to build the classifier: and the test data should be transformed or normalised, if appropriate, in the same way as the training data.

2 Using the WEKA Experimenter

The Experimenter has several advantages: Firstly it provides more information about the results than the Explorer. Secondly it is very convenient to run an experiment in which you compare different datasets, different classifiers, or different parameters for a classifier. Finally, it seems to be more stable than the Explorer.

Below you will find brief instructions on how to perform experiments with the simple Experiment Configuration Mode in the Experimenter. There is also an advanced mode which can be more convenient (but which is also more complicated to use). You can find more information about the Experimenter on the WEKA website or in the Experimenter tutorial:

http://sourceforge.net/projects/weka/files/documentation/3.5.x/ExperimenterTutorial-3-5-8.pdf

2.1 Setting up an Experiment

1. After starting the Experimenter go to the *Setup* tab. Make sure that the *Experiment Configuration Mode* is set to *Simple*.
2. Create a new experiment by clicking the *New* button.
3. Choose the dataset(s) that you would like to run your classifier on: In the *Datasets* box click *Add new...* and select a data file. The selected data file will then be shown in the list below the button. You can repeat this if you want to add multiple datasets.
4. Choose the experiment type: In the *Experiment Type* box select *Cross-validation*. Set the number of folds to 5. Make sure *Classification* is selected.
5. In Iteration *Control* set the *Number of repetitions* to 5.
6. Finally, select the classifiers that you would like to train and evaluate on your dataset(s): In the *Algorithms* box click *Add new...* In the dialog that opens choose the appropriate classifier from the list and set its parameter values. Then hit *OK*. The classifier name and its parameters will be added to the list. Repeat this for all classifiers / classifier parameterisations that you would like to compare.
7. When you are done configuring your experiment you can save the configuration. You can also specify a file to write the results to (but you do not have to).
8. To start your experiment go to the *Run* tab and hit *Start*.
9. In many cases, the WEKA Explorer allows you to modify the random seed that will be used. Just using the default seed is fine. If you do change the seed you need to report the seed you have chosen. You will have to work out some details in applying WEKA Explorer, including definitions of certain terminologies (such as "random seed" as referred above).

2.2 Examining the Results

When the experiment has finished go to the *Analyse* tab.

1. In the *Source* box click on the *Experiment* button to load the results of the experiment that has just finished running. In the *Configure* test box keep the default settings but check the box *Show std. deviations*. If this option is switched on, the standard deviation of the PC across the CV folds and runs will be displayed.
2. Start the analysis by clicking the *Perform test* button. The results will be shown in the *Test* output text field. This text field will contain a table that has one column for each classifier, and one row for each dataset. For each classifier / dataset pair this table contains the average PC (across CV folds and runs) as well as the standard deviation (in parentheses).
3. Under this table there is an extra row containing an entry with a format (../../..). This summarises the statistical significance of pair-wise comparisons of schemes using the corrected resampled T-Test. The column that has the (v/ /*) format is the baseline for the comparisons. The rest of the columns have an entry with three numbers. (1/0/2) indicates that the respective column had 1 significant win, 0 ties and 2 significant losses with regard to the baseline scheme. Additionally the v and * symbols will be displayed inside the table, next to the PC for the row that won or lost respectively.
4. You can configure the presentation of the results in various ways, using the options in the *Configure test* box. You can choose to see the classifiers as rows and the datasets as columns by clicking on *Select* next to *Row* and choosing *Schemes* in the list that appears and then clicking on *Select* next to *Column* and choosing *Dataset*. This way you can see if any of the datasets has significant wins/losses with regard to a baseline dataset. You can change the baseline by clicking on *Select base...* next to *Test base* and selecting the scheme that you want to use as baseline.
5. Each time you change the options in the *Configure test* box, you need to click the *Perform test* button to see the results.