Lab 7 – Machine learning

To get the lab approved, send your answers to: david.sundell@plantphys.umu.se

RapidMiner is a freely available tool supporting a large variety of machine learning algorithms: http://rapid-i.com

Start RapidMiner. If the program asks for a place to put its repository chose “New local repository” and type “d:”. Answer “No” to new updates.

RapidMiner represents each machine learning project as an operator chain (or tree) of algorithms. You can assemble different algorithms as you like and change parameters of each of them.

A number of templates are already available for common tasks such as cross validation using decision trees. Go to “File->Open Template”. Choose “Crossvalidation (Nominal, Decision Tree)”, “Next” and then “Finish”. An operator chain appears.

We will study the dataset in adenoca_markers.xls available on the course web site. For an introduction to the dataset and previous analysis, see Dennis-ClinCancerRes2005.pdf.

We can change any operation in a chain. Since our data is in Excel format, we need to exchange the “Retrieve” operator with an appropriate importer operator. Right-click the “Retrieve” operator, and click “Replace Operator->Import->Data->Read Excel”. Click on the “Read Excel” operator. In the window to the right, click “Import Configuration Wizard” and follow the steps to load the data. In Step 1, locate the file adenoca_markers.xls on your computer. Then click “Next” three times. In Step 4, mark the first column (ID) as “id” (should not be used for prediction) and the last column (Site) as “label” (the classes that we want to predict). Click Finish.

Run the operator chain by pressing the play-button. You will see the confusion matrix that is the result of the cross validation. How did your results compare to those reported in the paper?

Go back to the Operator chain. Double click the “Validation” operator. Explain what you see.

Right-click the “Decision Tree” operator and select “Breakpoint After”. When you now run the chain, the process will stop and show the decision tree in each iteration of the cross validation. Press the play-button to continue running. How many trees do you get? Why?

Can you obtain better performance by using another learner (by selecting under "Replace Operator->Modeling->Classification and regression")? Experiment! Notice that the adenoca_markers.xls data set contains a mix of discrete (e.g. + and -) and numerical data. It also contains multiple classes. Because of this, many machine learning methods cannot be applied to this dataset. RapidMiner will warn you if you select an inappropriate method.

To try other machine learning methods, such as SVMs, you can use one of the data sets supplied with RapidMiner. Replace the "Read Excel" operator with "Repository Access->Retrieve". Select e.g. the Samples/data/Weighting-dataset. This contains two classes and only numerical data. Experiment!