

COMP4702/COMP7703 - Machine Learning

Prac 1 – Introduction to Machine Learning and WEKA

Aims:

- To reinforce some of the basic concepts of machine learning from lectures.
- To gain familiarity with the WEKA software package.
- To produce some assessable work for this subject.

Procedure:

Weka is a freely available machine learning software package that is very well-known and widely used. The Weka homepage is:

<http://www.cs.waikato.ac.nz/~ml/weka/index.html>

The software is written in java and full source code and code documentation is provided. There is also an accompanying book (available in the library):

<http://www.cs.waikato.ac.nz/~ml/weka/book.html>

While the book is quite good, we won't be following it in the course and it isn't really a users' guide to the software anyway. The attachments to this prac are from this book however.

Weka contains implementations of many machine learning algorithms, and has many other features such as the ability to run large batches of experiments, data visualization and preprocessing and so on. We will start by using some of the basic components in Weka - try not to be distracted by all the features that we won't be using!

- Run Weka and select **Explorer** from the **GUI Chooser** window.
- Scan through the Explorer User Guide document (course materials webpage) and compare with what you see on the screen, to get a feel for the software. Don't worry about understanding all points yet.

Weather dataset

- To start, open one of the datasets that comes with Weka (Open file, data folder, weather.arff). A hardcopy table of the data is available as a handout (at the prac or lectures). The **Preprocess** tab displays information about the data (list of attributes, number of instances, etc).
- Go to the Visualize tab. What you see is a “draftman's display” – i.e. scatterplots of the data with respect to all combinations of pairs of attributes. Some of these aren't very useful – e.g. a plot of an attribute with itself (humidity-humidity): 1-D data in a 2-D space. Another example is looking at “windy” and “play” – these attributes are binary, so many datapoints “sit on top of each other” in the scatterplot. A better plot is “temperature” v's “humidity” – click on this scatterplot to get a closer look at this plot.
 - **Q1:** for a dataset with n attributes, how many unique scatterplots are in a draftsman's display?

- Now, we will use a very simple classification algorithm to try and predict if we should play some game or not given the other attributes in the data. Go to the **Classify** tab and **choose** the **OneR** classifier. This is a one-level decision tree: for more information refer to the handout. For now we will simply evaluate the classifier on the training set (i.e. all the data). Note the “(Nom) play” drop-down indicates our current output variable (you could easily choose one of the other attributes here). Click Start to train the classifier, and study the output that is produced in the main window. Everything up to and including Incorrectly Classified Instances should make sense.
 - **Q2:** explain in words the classification rule produced by OneR on the data. How many examples does it classify correctly?
- To visualize the results, right-click on the item generated in the “Result list”, and select “Visualize classifier errors”. Change the variables plotted on the X and Y axes for different views of the data. Click on a particular data point to see the correct and predicted output values for that point. Remember, OneR has only used one attribute in its decision.
- Try out a more powerful decision tree on this dataset (**J48** implements C4 decision trees). Don’t attempt to change any algorithm parameters.
 - **Q3:** explain in words the classification rule produced by J48 on the data. How many examples does it classify correctly?

Iris Dataset

One of the most widely used datasets is the Iris dataset, which originates from a famous statistician – R. A. Fisher:

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annual Eugenics* 7 (part II): 179-188. Reprinted in *Contributions to Mathematical Statistics*, 1950. New York: John Wiley.

The dataset contains fifty examples of each of three types of plant (Iris setosa, Iris versicolor, Iris virginica), with four numerical measurements for each observation: sepal length, sepal width, petal length and petal width. This data is included as one of the Weka example datasets.

- Load the iris data and have a look at it via the Visualize tab. It should be clear that there is structure in the data which has a close relationship with the different plant species (classes).
- Try out the OneR and J48 classifiers on this data, but this time use a test set to evaluate the generalization performance of your classifiers. In the classify tab, Weka allows you to select a percentage of the dataset to set aside for testing.
 - **Q4:** Run five experiments with each classifier, using these values for percentage split: 10, 30, 50, 70, 90. Record the number and percentage of correctly and incorrectly classified instances. What can you conclude (if anything!) from these results?

Final comment: Weka has an interesting feature – a **UserClassifier** which allows you to interactively build up a classification tree (i.e. a hypothesis) by selecting regions of the space with the mouse. Experiment with this on the Iris data (select UserClassifier and click start). You can for example draw rectangles similar to the simple example in Chapter 2 of the text.