CS910 Exercise Sheet 1: Getting Started With Weka

The aim of these exercises are to gain you some basic familiarity with the Weka tool as a way to explore data. First, we need to get started with Weka. Note: the marks are only indicative for the relative weights of the questions.

Opening Weka

Linux Systems at Computer Science. Log-on to a machine using your CS user name and password (E.g. the machines in CS0.01).

If you have any problems logging on, follow the advice at `http://www2.warwick.ac.uk/fac/sci/dcs/intranet/user_guide/`

Alternatively, you can log in remotely to joshua from a machine with an X windows server or VNC server, see `http://www2.warwick.ac.uk/fac/sci/dcs/intranet/user_guide/remote-login/`

Open up the program by typing `weka &` from the console.

Personal computer (Windows, Mac, Linux). Follow the instructions to dowload and install Weka from `http://www.cs.waikato.ac.nz/ml/weka/downloading.html`

Opening a data set

We will use the "Explorer" Graphical User Interface (GUI) to explore a data set, so click on 'Explorer'. The explorer window should now open up.

We will work with the "Adult" data set, which contains census data on a number of individuals.

A copy is available from the course pages under the Exercises tab, at `https://www2.warwick.ac.uk/fac/sci/dcs/teaching/material/cs910/exercises/adult.arff` in ARFF format (Weka's native file format).

You can either download a local copy of this data set, and access it via the "open file" dialogue, or by selecting "open URL" and pasting in the URL above.

Exploring the data

Weka should now show information about the data. The relation is called 'adult', there are 48842 instances (examples), and there are 15 attributes.

By default, the last attribute, "class", is selected as the target (class) attribute, and the interactions of the other attributes is visualized at a bar graph in the bottom right.

1. Select the "age" attribute, and read off the minimum, maximum and mean values for age. [15]

2. Tracking your mouse over the plot of the age attribute will give the count for each bar, and the range of input values it covers. Why do some bars stick out, about twice as far as their neighbours? What does this say about the data? [15]

Investigate the other attributes, and look at the breakdown of values in each.

3. Which attributes are nominal (categoric)? [15]

4. What is the average number of hours worked per week? [10]

5. Which is the most popular occupation in the data set? Which is the most common native country? [15]

The selector in the bottom right allows you to change the "class attribute", to see how other pairs of variables relate. Select "sex" as the class attribute.

6. Look at the interaction between 'sex' and 'relationship'. Some attribute values are 'pure' — only associated with one class attribute value. Why is this? [15]

7. Study the relationship between education and education-num. What do you learn about these two attributes? [15]