Using Weka for Classification

Preparing a data file

Prepare a data file in CSV format. It should have the names of the features, which Weka calls attributes, on the first line, with the names separated by commas. The name of the class label should be the last attribute name.

Then for each instance in the training set, there should be a line of all the feature values, separated by commas (in the same order as the feature names, naturally) and the value of the class label should be last on the line.

Starting Weka and Preprocessing data

Start Weka (on a Mac, click on the jar file), which opens a small window titled Weka GUI Chooser. Select the Explorer button. A second window will open with Weka commands. Note that in the top row of tabs, we are in the Preprocess tab.

Open the .csv file that you just saved from Excel. You will need to select the .csv File Format in the Open window.
If there is any problem loading the file, for example, if there are conflicts between attribute names, it will give you an error message.

Now there may be some filtering that we need to apply to the attributes. First, if there is an ID field, we don’t want to use it for analysis. If there is an ID field, or any other field that you want to remove, check the square boxes in front of the attribute names. Then click the Remove button at the bottom of the Attributes window.

When Weka imports data from a .csv file, it makes all the cells with a number to be Weka type Numeric. But if the numbers represent a small number of possible choices, for example, 0 and 1, we should convert the attribute to the Weka type Nominal. Note the index numbers of any attributes that you need to convert from Numeric to Nominal.

Right below, the Open File button is the Filter window. Click Choose and follow the menu choices to get filters -> unsupervised -> attribute -> NumericToNominal. When you select this filter, it appears in the filter window with the default settings: NumericToNominal –R first-last, meaning convert all the attributes from the first to the last.
Instead, right click on the filter window itself and click Show Properties. A small window titled weka.gui.GenericObjectEditor will come up. In the attribute indices box, replace the attribute indices “first-last” with the numbers that you want to remove, for example, “5, 8, 21” and click o.k. If the filter box now looks o.k., click the apply on the right.

Viewing the attributes

If you select any attribute, you can view its properties and something about its values. The values summary is in the Selected Attribute pane and the distribution of the class label among the values of that attribute are shown in the bottom right pane.

Classification Analysis

To run the analysis, click the Classify button in the top row.

First, we Choose which classification algorithm we want to run. I recommend using SMO, which is one of the Support Vector Machines packages in Weka, on text classification problems. Click SMO and it will appear in the Classifier window with
its default settings. We will use its default settings, so there is no need to change them.

Next, we can choose either Cross-validation or Percentage Split. With Percentage Split, Weka will use choose a random set of examples in the percentage, e.g. 66%, to train the model, and then will use the remaining 33% to test if the model is correct. If we use Cross-validation, it does this for the number of folds across the data, thus averaging out any inconsistencies caused by random selection. Let’s try percentage split at first, just to save time, but remember that cross-validation gives more reliable results.

Click Start on the left about half-way down. While weka is training and testing the model, the weka icon on the lower right dances, and when it’s done, it sits down. The results show in the classifier output window.

For this example, taken from a donor spread sheet, we observer that the overall classifier accuracy is 54%.

We also have a Confusion Matrix at the bottom. In this particular example, it shows the following entries in the matrix:

Number supposed to be 0 correctly classified as 0: 848
Number supposed to be 0 incorrectly classified as 1: 712
Number supposed to be 1 incorrectly classified as 0: 696
Number supposed to be 1 correctly classified as 1: 864

Other Notes:
If you run other experiments, they will pile up in your Result List.
If you right-click on any item, you can save it.
If you do too many experiments without deleting results from the list, weka will run out of memory and crash, so delete some from time to time.

Experiments with Attributes

Even if you have loaded all the attributes into weka, you can experiment with using different numbers of them. In order to select attributes, go back to the Preprocess tab, and Remove the ones you don't want.

Note that if you want to get back attributes that you removed, there is an **Undo button** at the top.

More Information

The Weka documentation page is at:  
For a very introductory video tutorial, there is a link to Brandon Weinburg's video Weka tutorial:  [http://www.youtube.com/watch?v=m7kpIBGEdkI](http://www.youtube.com/watch?v=m7kpIBGEdkI)