Data in Weka

Weka prefers to load data in the ARFF format. ARFF is an acronym that stands for Attribute-Relation File Format. It is an extension of the CSV file format where a header is used that provides metadata about the data types in the columns.

For example, the first few lines of the classic iris flowers dataset in CSV format look as follows:

```
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
```

The same file in ARFF format looks like:

```
@RELATION iris

@ATTRIBUTE sepallength REAL
@ATTRIBUTE sepalwidth REAL
@ATTRIBUTE petallength REAL
@ATTRIBUTE petalwidth REAL
@ATTRIBUTE class {Iris-setosa,Iris-versicolor,Iris-virginica}

@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
```

You can see that directives start with the at symbol (@) and that there is one for the name of the dataset (e.g. @RELATION iris), there is a directive to define the name and datatype of each attribute (e.g. @ATTRIBUTE sepallength REAL) and there is a directive to indicate the start of the raw data (e.g. @DATA).

Each attribute can have a different type, for example:

- **Real** for numeric values like 1.2.
- **Integer** for numeric values without a fractional part like 5.
- **Nominal** for categorical data like “dog” and “cat”.
- **String** for lists of words, like this sentence.

On classification problems, the output variable must be nominal. For regression problems, the output variable must be real.

Lines in an ARFF file that start with a percentage symbol (%) indicate a comment.

Values in the raw data section that have a question mark symbol (?) indicate an unknown or missing value. The format supports numeric and categorical values as in the iris example above, but also supports dates and string values.

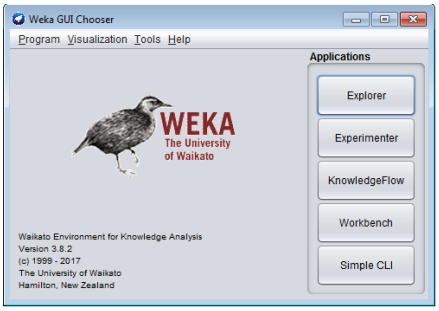
If you need some more examples of .arff files, remember that your Weka installation came with data files under the *data/* subdirectory. These default datasets distributed with Weka are in the ARFF format and have the .arff file extension.

Load CSV Files in the ARFF-Viewer

Your data is not likely to be in ARFF format. In fact, it is much more likely to be in Comma Separated Value (CSV) format. This is a simple format where data is laid out in a table of rows and columns and a comma is used to separate the values on a row. Quotes may also be used to surround values, especially if the data contains strings of text with spaces.

The CSV format is easily exported from Microsoft Excel, so once you can get your data into Excel, you can easily convert it to CSV format. Weka provides a handy tool to load CSV files and save them in ARFF. You only need to do this once with your dataset. Using the steps below you can convert your dataset from CSV format to ARFF format and use it with the Weka workbench.

1. Start the Weka chooser.



2. Open the ARFF-Viewer by clicking “Tools” in the menu and select “ArffViewer”.

3. You will be presented with an empty ARFF-Viewer window.

4. Open your CSV file in the ARFF-Viewer by clicking the “File” menu and select “Open”. Navigate to your current working directory. Change the “Files of Type:” filter to “CSV data files (*.csv)”. Select your file and click the “Open” button.

5. You should see a sample of your CSV file loaded into the ARFF-Viewer.

6. Save your dataset in ARFF format by clicking the “File” menu and selecting “Save as…”. Enter a filename with an .arff extension and click the “Save” button.

7. You can now load your saved .arff file directly into Weka.

8. Note, the ARFF-Viewer provides options for modifying your dataset before saving. For example you can change values, change the name of attributes and change their data types.

9. It is highly recommended that you specify the names of each attribute as this will help with analysis of your data later. Also, make sure that the data types of each attribute are correct.

To learn more about .arff and .csv file formats, check out the following links.

https://www.cs.waikato.ac.nz/ml/weka/arff.html
https://en.wikipedia.org/wiki/Comma-separated_values