Weka: Preprocessing

Lab 01 (in-class): 10 NOV 2016, 10:00-12:00, TURING

Learning objectives

In this assignment you are going to:

- install Weka; do some troubleshooting, if needed;
- get familiar with the arff format;
    - create a small dataset from scratch;
    - analyze and explore standard datasets.

Preliminaries: Weka Installation

The first thing to do is to download and install Weka.
Go to <http://www.cs.waikato.ac.nz/ml/weka/downloading.html>.

Download and install Weka 3.6. This is a stable version that corresponds to what is described in the third edition of the data mining book.

Occasionally you may find that Weka crashes with an out-of-memory exception, particularly for larger data sets or with extensive preprocessing of data. This is related to the default heap size allocation for the Java Virtual Machine (JVM). Read here if you experience this problem: https://weka.wikispaces.com/OutOfMemoryException (solution varies according to you operating system).

IMPORTANT: If you get errors while running any software, it is common practice to ***google*** the text of the error message. Solutions are often available on the web, sometimes with comprehensive explanations. If you find a plausible solution on the web, always check the date (the solution might be obsolete), the version of weka the solution refers to, and the underlying operating system.

Weka Interfaces: The Explorer

Weka is, in general, easy to use and well documented. Weka interfaces are user-friendly and intuitive. Weka is also equipped with contextual help, which is very useful when you wish to know more about parameters. Tooltips (bubbles with text) tell you what to do in order display contextual help.

Weka has three graphical user interfaces: the Explorer, the Experimenter and the Knowledge Flow Interface. Weka is also provided with a Command Line Interface (CLI) and with a Java API (if interested in the API, watch this introductory tutorial:
< https://www.youtube.com/watch?v=q3Gf6kqaJWA >.
In this lab, we are going to use the **Explorer**, which gives access to all its facilities using menu selection and form filling.

(It is worth knowing that all the standard weka sample datasets are available online here: < http://storm.cis.fordham.edu/~gweiss/data-mining/datasets.html >. They are also included in the Weka distribution you installed on your computer).

Tasks 1 - Create an arff file from scratch: the Course Ratings dataset

Look at the dataset in Fig. 1 This dataset is described in Daumé (2015: 11[i]) and shown in the Appendix of the Daumé's book). We will call this dataset "Course Ratings", for our convenience.

Rating	Easy?	AI?	Sys?	Thy?	Morning?
+2	y	y	n	y	n
+2	y	y	n	y	n
+2	n	y	n	n	n
+2	n	n	n	y	n
+2	n	y	y	n	y
+1	y	y	n	n	n
+1	y	y	n	y	n
+1	n	y	n	y	n
0	n	n	n	n	y
0	y	n	n	y	y
0	n	y	n	y	n
0	y	y	y	y	y
-1	y	y	y	n	y
-1	n	n	y	y	n
-1	n	n	y	n	y
-1	y	n	y	n	y
-2	n	n	y	y	n
-2	n	y	y	n	y
-2	y	n	y	n	n
-2	y	n	y	n	y

Fig. 1: The Course Ratings dataset (Daumé, Appendix)

Preparing the data

Open an ASCII editor of your choice and create an arff file from the dataset in Fig 1. The arff format is described in Witten et al. (2011: Ch 2; Ch 11: 407-410) and in the Weka wiki: < http://weka.wikispaces.com/ARFF -->"ARFF (book version)".
Save your file as "**course_ratings.arff**" in your working folder on your computer.

Loading the data into the Explorer

Launch the Weka Explorer. Load the dataset (Preprocess tab → Open file).

Warming up Questions:

1. Can you load the file without any problem? If not, go back to the editor and try to understand and fix the problem(s).
   ATT! By default, the class is the last attribute in an ARFF file, but you can declare another position. Watch this video to know more:

https://www.youtube.com/watch?v=bhxqV3GK-K8

2. Once you have loaded the file, observe the interface: What's the name of the relation? How many instances have you got? How many attributes? Which one is the class?
3. What else is worth noting in this visualization of the dataset?
4. What happens if the class attribute is left on the first column? Compare the visualizations of the 2 datasets (the one with class on the first column, before any manipulation, and the one with the class on the last column) as they appear on the right side of the Preprocessing panel: can you notice any differences?

Once you have become familiar with the visualization, try the following actions: *deletion, inversion, undoing, editing*. You can delete an attribute by clicking its checkbox and using the **Remove** button. **All** selects all the attributes, **None** selects none, and **Invert** inverts the current selection. You can undo a change by clicking the **Undo** button. The **Edit** button brings up an editor that allows you to inspect the data, search for particular values and edit them, and delete instances and attributes. **Right-clicking** on values and column headers brings up corresponding context menus.

Task 2 - Nominal data: the Weather (nominal) dataset

Download the *weather nominal* dataset:
http://stp.lingfil.uu.se/~santinim/ml/2016/Datasets/weather.nominal.arff

Load the dataset into the Explorer. Click the **Edit** button from the row of buttons at the top of the **Preprocess** panel. This opens a new window called *Viewer*, which lists all instances of the weather nominal data.

Q1. How many instances are listed in the dataset? How many attributes? What's the attributes names? What type of attributes are these? which one is the class?

Click on the name of an attribute in the left subpanel to see information about the **Selected attribute** on the right, such as its values and how many times an instance in the dataset has a particular value. This information is also shown in the form of a *histogram* in the lower part of the **Selected attribute** subpanel.

Q2. What is the function of the first column in the *Viewer* window? What is the class value of instance number 8 in the weather data?

Weka filters:
Let's remove an attribute from the dataset.
The appropriate filter is called Remove; its full name is weka.filters.unsupervised.attribute.Remove Examine this name carefully. Filters are organized into a hierarchical structure of which the root is weka. Those in the unsupervised category don't require a class attribute to be set; those in the supervised category do. Filters are further divided into ones that operate primarily on attributes (the attribute category) and ones that operate primarily on instances (the instance category). Click the Choose button in the Preprocess panel to open a hierarchical menu from which you select a filter by following the path corresponding to its full name. Use

the path given in the full name above to select the Remove filter. The text "Remove" will appear in the field next to the **Choose** button. Click on the field containing this text. The Generic Object Editor window, which is used throughout Weka to set parameter values for all of the tools, opens. In this case it contains a short explanation of the **Remove** filter, click **More** to get a fuller description Enter 3 into the attributeIndices field and click the OK button. The window with the filter options closes. Now click the **Apply** button on the right, which runs the data through the filter. The filter removes the attribute with index 3 from the dataset, and you can see that this has happened. This change does not affect the dataset in the file; it only applies to the data held in memory. The changed dataset can be saved to a new ARFF file by pressing the **Save** button and entering a file name. The action of the filter can be undone by pressing the **Undo** button. Again, this applies to the version of the data held in memory.

What we have described illustrates how filters are applied to data. However, in the particular case of Remove, there is a simpler way of achieving the same effect. Instead of invoking a filter, attributes can be selected using the small boxes in the Attributes subpanel and removed using the Remove button that appears at the bottom, below the list of attributes.

Q3: (Tricky! try and find out yourselves!) Use the filter weka. unsupervised.instance.RemoveWithValues to remove all instances in which the humidity attribute has the value high. To do this, first make the field next to the Choose button show the text RemoveWithValues. Then click on it to get the Generic Object Editor window, and figure out how to change the filter settings appropriately. *Undo the change to the dataset that you just performed, and verify that the data has reverted to its original state*.

Task 3 - Numeric Data: the Iris dataset

Download the *iris* dataset:
http://stp.lingfil.uu.se/~santinim/ml/2016/Datasets/iris.arff

Load the dataset into the Explorer. Click the **Edit** button from the row of buttons at the top of the **Preprocess** panel. This opens a new window called *Viewer*, which lists all instances of the iris data.

Q4. Describe the dataset. What is the content of the comments? How many instances does this dataset have? How many attributes? What is the range of possible values of the attribute petallength? How many numeric and how many nominal attributes does this dataset have? What's the name of class? What's the distribution of the classes (balanced or biased towards one class?

The Visualize Panel
Now take a look at Weka's data visualization facilities. These work best with numeric data. Click the Visualize tab to bring up the Visualize panel. Click the first plot in the second row to open up a window showing an enlarged plot using the selected axes. Instances are shown as little crosses, the color of which depends on the instance's class. The x-axis shows the **sepallength** attribute, and the y-axis shows **petalwidth**. Double-

clicking on one of the crosses opens up an **Instance Info** window, which lists the values of all attributes for the selected instance. Close the Instance Info window again. The selection fields at the top of the window containing the scatter plot determine which attributes are used for the x- and y-axes. Change the x-axis to petalwidth and the y-axis to petallength. The field showing Color: class (**Num**) can be used to change the color coding.

Each of the barlike plots to the right of the scatter plot window represents a single attribute. In each bar, instances are placed at the appropriate horizontal position and scattered randomly in the vertical direction. Clicking a bar uses that attribute for the x-axis of the scatter plot. Right-clicking a bar does the same for the y-axis. Use these bars to change the x- and y-axes back to sepallength and petalwidth. The **Jitter slider** displaces the cross for each instance randomly from its true position, and can reveal situations where instances lie on top of one another. (*in weka "jitter"=uncover obscured points; generally speaking, you can interpret this word in the sense of "displacement" or "movement"; jitter has mainly a technical sense related to "signals". Read the wikipidia article, if you wish to know more about the technical sense*).
Experiment a little by moving the Jitter slider. The **Select Instance** button and the **Reset**, **Clear**, and **Save** buttons let you modify the dataset. Certain instances can be selected and the others removed. Try the Rectangle option: Select an area by left-clicking and dragging the mouse. The Reset button changes into a Submit button. Click it, and all instances outside the rectangle are deleted. You could use **Save** to save the modified dataset to a file. **Reset** restores the original dataset.

Task 4 - Explore more by yourself: the Weather (numeric) dataset

Download the *weather* (numeric) dataset:
http://stp.lingfil.uu.se/~santinim/ml/2016/Datasets/weather.arff

Load the dataset into the Explorer. Get to know the new dataset: replicate all the actions in Task 3. What happens? Describe your explorations.

Additional references

Weka Tutorial 01: ARFF 101 (Data Preprocessing):
https://www.youtube.com/watch?v=gd5HwYYOz2U

---the end---

[i] "In order to learn, I will give you training data. This data consists of a set of user/course examples, paired with the correct answer for these examples (did the given user enjoy the given course?). From this, you must construct your questions. For concreteness, there is a small data set in Table ?? in the Appendix of this book. This training data consists of 20 course rating examples, with course ratings and answers to questions that you might ask about this pair. We will interpret ratings of 0, + 1 and + 2 as "liked" and ratings of -2 and -1 as "hated.""
(Daumé, 2015:11).