Introduction to Weka

Objectives of this Class

1. Learn to Use the Weka software tool;

YOU MUST bring a USB drive to the tutorial on Friday, Sept. 9 in 046 Colburn.

Weka

The Weka workbench is a set of tools for preprocessing data, experimenting with data-mining/machine-learning algorithms, and comparing the performance of different methods. Weka also provides a Java class library that enables one to use the Weka filters and classifiers in their own programs and modify them as desired.

1. Although there are other Weka interfaces for advanced users, we will use the Explorer interface for most of our work.
2. Go to http://www.cs.waikato.ac.nz/ml/weka and download Weka to your laptop. You should download the stable version for the 3rd edition of the textbook (weka-3-6-14). Make sure you can invoke Weka — do not use the "with console" option. You should get a screen that displays a bird and offers a choice of four graphical user interfaces. Click on **Explorer**.
3. To invoke Weka in the 046 Colburn lab, do the following:
   (a) Press any key and then login using your UDEL username and password.
   (b) Once you are logged in, double click on AppCast in the upper lefthand corner of the screen.
   (c) Then scroll down to find Weka (it is near the bottom of the page).
   Click on the weka-3-6-14 entry and click on *Launch*.
   (d) You should get a screen that displays a bird and offers a choice of four graphical user interfaces. Click on **Explorer**.
4. If you are working on a machine in 046 Colburn, you will need to store files on a personal USB drive.
   (a) Insert your USB drive. *This PC* is on the left side of the screen. Double click on it to get to your USB drive.
   (b) To remove your USB drive, go to the icons on the bottom right of the screen. Click on the leftmost icon to safely eject your USB drive.
5. To log off of the machines in 046 Colburn, right click on the icon in the lower left corner, and then sign off or restart.

ARFF data files

An **ARFF file** (Attribute-Relation File Format) is a standard way of representing machine learning data sets as flat files (no relationships among instances). Weka works with ARFF files.

- Lines beginning with % are comments.
- Lines beginning with @ define the relation and then its attributes.
  1. The relation is defined first using the command **@relation**:
     - @relation <relation-name>
  2. Then each attribute is defined using the command **@attribute**:
     - @attribute <attribute-name> {<set-of-comma-separated-possible-attribute-values>}
     - @attribute <attribute-name> numeric
  3. Then the data values are given:
     - @data
     - set of data instances, given as comma-separated values, with each data instance on a separate line

Using Weka

1. **Weka Preprocessor: Loading and Examining Data**

   (a) Often data is in an Excel spreadsheet, which can be converted to a CSV file (comma-separated file) which can in turn be converted to an ARFF file in Weka.

   (b) Open the Excel spreadsheet *Mushroom-data-625.xls* which can be found in the **Data-sets** directory on the class web site at www.cis.udel.edu/∼carberry/CISC-483-683. Open this file and save it to a folder on your USB drive. (On the machines in 046 Colburn, *This PC* is on the left side of the screen. Click on it to get to USB drive.) If you are working on your own laptop, you might want to put it in a subfolder of the *data* folder that is created as a subfolder of the *weka-3-6-14* folder that was created when you downloaded Weka. (Notice that the Weka *data* folder already contains some data files that we will use during the course.)

   (c) Open the file *Mushroom-data-625.xls*. Let us examine the structure of *Mushroom-data-625.xls*:
   - The first row gives the attribute names.
   - Each subsequent row represents an instance, with the value for each attribute given in the respective column.

   (d) To convert to a csv file, click on "Save-as", select "CSV" as the file type, and save (in this case as *Mushroom-data-625*). (Be sure to save it on your USB drive if you are working on a machine in 046 Colburn.) This saves the file as a comma-separated csv file, though if one opens the file, it still looks like an Excel spreadsheet. Note that you now have both a csv file and an Excel file named *Mushroom-data-625*.

(e) To convert the csv file to ARFF:

i. Invoke the Weka Explorer GUI

ii. Select "Open File" under "Preprocess", move to the folder in which you stored the CSV file *Mushroom-data-625*, set the file type to CSV, and select the file *Mushroom-data-625* as the file to open. You should be opening the file that you saved as a csv file.

- Weka automatically changes the file to ARFF format.

(f) The csv file does not specify which attribute is the class attribute. You can specify the class attribute by clicking on "Class" in the middle of the right side and selecting the attribute that should serve as the class. Select the attribute **Status** as the class attribute — it can take on the value *e* for edible or *p* for poisonous.

(g) By clicking on one of the attributes on the left, you will see a histogram that shows how often each of the two values of the class occurs for each value of the selected attribute.

- Note that if you select the class attribute itself (in this case **Status**), the histogram shows how often each of the classes occurs in the data. The table on the right above the histogram enables you to identify what the histogram colors mean — for example the first row of the table shows 63 instances of Status=*p* which correlates with the first column in the histogram; note that the first column of the histogram is labelled as having 63 instances.

(h) On the left side, select *gill-color* as the attribute (but keep *Status* as the Class value). On the right you see a histogram showing the distribution of values of the *Status* attribute for the different possible values of the *gill-color* attribute.

(i) **Questions-1:**

i. How many instances are there in the data file?

ii. How many different values can *gill-color* take on?

iii. Which values of the *gill-color* attribute result in only edible mushrooms?

2. **Weka Preprocessor: Saving and Viewing ARFF Files**

   (a) Save the file in ARFF format. To do this, click on "Save" in the Weka Explorer; in the window that appears, set the file type to "Arff data files", give the file a name (in this case, set the name as *Mushroom-data-625*), and then click on "Save". Weka has stored the file in ARFF format so that you can use it again in the future. Note that the Weka data files stored in the data subfolder of the Weka folder are stored in ARFF format.

   (b) **Questions-2:**

      i. Use a text editor to view the ARFF file representing the mushroom data. What does the first line tell you?

      ii. What do the next 23 lines tell you?

      iii. What do the rest of the lines tell you?

3. **Weka Preprocessor: Editing ARFF Files in Weka**

   (a) In the Weka Explorer, you can edit the data file by clicking on *Edit*; you can save the edited file in Weka (not one of your folders) by clicking on *Save*. Click on *Edit* in the Preprocessor and examine what appears.

   (b) You can change the value of an attribute. For example, try left-clicking on one of the attribute values for an instance — note that you are given a choice from among the valid values that the attribute can assume.

   (c) You can make changes to an attribute or even remove an attribute. For example, right-click on one of the attribute names at the top of the window, such as *Cap-shape*. Note that you are given a number of choices — clicking on *delete-attribute* will delete the entire column for this attribute.

   (d) Click on *Cancel* to return to the Explorer without making any changes to the data being examined. (If you inadvertently saved a revised data file, you can always go back to your folder and reload the original file, since the changes are only being saved in Weka and are not reflected in the data files on your PC.)

4. **The Classifier**: Now we want to try using a classifier.

   (a) Click on *Classify*.

   (b) Click on *Choose* under *Classifier* and then click on the symbol to the left of *Trees* and then click on *J48*. This will place *J48* as the name of the classification method shown to the right of *Choose*. (J48 is the Weka name for a decision tree classifier based on C4.5, a well-known decision tree algorithm.)

   (c) Click on the method that is listed next to *Choose* — it should be *J48*. Click on *More* to get information about the method that will be used. In the future, you will find this information helpful. Then close the window, click on *Cancel*, and return to the Explorer.

   (d) Immediately below *More options* on the left side of the Explorer screen, there is a line that gives an attribute name preceded by "(Nom)"; this attribute is probably not the class attribute that you want. Click on the arrow to the right, and select *Status* as the attribute to be used as the class attribute.

   (e) Click on *Use training set* under *Test options* on the left side of the Explorer window. This means that the classifier will be built and tested on the same set of data.

   (f) Click on *Start* to build the classifier. The results appear on the right side of the Explorer screen. Below where it says *"J48 pruned tree"*, you will see a textual description of the tree. Look at this and get an idea of what the tree looks like.

(g) To actually visualize the tree, do the following. Right click on the last line on the left side of the screen under *Result list*, and select *Visualize tree*. A new window will appear with a graphical view of the decision tree that correlates with the textual description of the tree.

(h) **Questions-3:** Scroll through the screen to answer the following questions.

i. Draw the decision tree that was developed.

ii. How many instances were classified correctly? How many were classified incorrectly?

(i) Now we want to remove the column for the attribute *Odor* and see what happens if *Odor* is not available as an attribute. There are three ways that you could do this.

i. The first two ways have already been discussed. What are they? (If you don't remember, ask the instructor.)

ii. The third way is to go back to the Preprocessor by clicking on *Preprocess* at the top of the Explorer window, then click on the square box next to an attribute name and then click on *Remove*. Do this to remove the attribute *Odor*.

iii. Now also remove the attributes *gill-size*, *stalk-root*, and *habitat*.

(j) Invoke the same classifier on this revised data — make sure that you have reset the class attribute as *Status*. Visualize the resulting decision tree. Enlarge the window displaying the tree; then right click on empty space in the window and select *Fit to Screen* from the menu that appears. For each leaf node, the class value at that leaf node is given along wih one or two numbers. If there is only one number, that tells how many instances reached that leaf node and were classified correctly; if there are two numbers, the first tells how many of the instances that reached this leaf node were classified correctly and the second number tells how many were classified incorrectly.

(k) **Questions-4:**

i. How many instances were classified correctly? How many were classified incorrectly?

ii. What attribute is at the root of the decision tree?

iii. Consider the following path in the decision tree: *cap-color*=w, *gill-spacing*=c, *population*=a. What is the class value assigned to instances that follow this path?

iv. How many paths in the decision tree lead to a leaf node where some instances are classified incorrectly?

5. **The Classifier: Noise in the Data**

(a) Go back to your spreadsheet data and copy the first data instance and insert it as five new rows at the beginning of the spreadsheet. Then change the value of *Status* for these five new rows to *r* instead of *p*. Leave the sixth row unchanged. Then change the *Status* value for the next ten data instances to *r*. (Notice that the Status attribute now has 3 possible values, *p*, *e*, and *r*.) Convert the revised file to cvs format, save it as *Mushroom-data-625-revised*, load it into Weka, save the file in ARFF format, and then run the classifier on it. (Be sure to set *Status* as the class attribute.) Examine the results.

(b) At the bottom of *Classifier output* is a matrix that is called a *confusion matrix* which we shall refer to as *C*. If there are n possible classes, then the confusion matrix has n rows and n columns, one for each possible class. The entry C_i, j gives the number of data items whose correct class is i and which were classified by J48 as class j

(c) **Questions-5:**

i. How many instances are incorrectly classified? Why did this happen?

ii. What does the diagonal of the confusion matrix tell you?

iii. Which class did the classifier always get wrong?

6. **The Classifier: Dividing Data into Training and Test Sets**

   (a) Instead of training and testing on the same data set, we can ask Weka to hold out part of the data set (ie., not use it to train our classifier) and use it instead as a test set. To save part of the data set for testing, click on the radio button to the left of *Percentage split* under *Test options*, and enter the number 50 as the %. Run the Classifier and look at the results. (Remember that the dataset you are using, *Mushroom-data-625-revised*, has 630 instances in it.)

   (b) Now do this twice more, once with 25 and once with 5 as the % for the split. Look at the results and answer the following questions:

   (c) **Questions-6:**

      i. How many instances were used for training when there is a 50% split? How many for testing?
      ii. How many instances were misclassified when there is a 50% split?
      iii. How many instances were used for training when there is a 25% split? How many for testing?
      iv. How many instances were misclassified when there is a 25% split?
      v. How many instances were used for training when there is a 5% split? How many for testing?
      vi. How many instances were misclassified when there is a 5% split?
      vii. What is the error rate under each of the different splits?
      viii. What do you think is causing the differences in classification error rate under the different splits?

7. **The Classifier: Viewing the Output**

   (a) You can also see how the classifier classifies the individual instances in the test set. Let us examine how to do this when only 5% of the data is in the test set (so that the output is not huge).

   (b) Set the split at 95% in the training set and thus 5% in the test set.

   (c) Click on *More Options*, and then click on the box next to *Output predictions* and then Click on *OK*.

   (d) Run the classifier again, and examine the classifer output. Note that you can now see how each instance in the test set was classified, and the ones that are incorrectly classified are noted by a + sign in the error column.

8. **The Classifier: Rerunning Models**

   (a) Note that on the lower left side of the Explorer window, there is a *Result-list* with an entry for each run of the Classifier. If you click on one of these, you go back to the results for that run. Try it to see that this is the case.

   (b) You can also save a model for future use or reload a previously saved model. Right click on one of the models in the *Result-list*, save it, and then reload it into Weka. Note that only the model is loaded, not the results of testing the model.

   (c) You can also save the results from testing a particular model, and then go back and view the results using a text editor.

9. **The Classifier: Using a New Test Set**

(a) You might want to run a model on a new test set of data instances. This can be a bit tricky, since the test set must be in **EXACTLY** the same format as the original training set. Let us explore the problems involved in doing this.

(b) Now you want to run your model on a new test set. Open the Excel spreadsheet *Mushroom-data-625-test.xls* which can be found in the **Data-sets** directory on the class web site. Save this file on your own machine and name it *Mushroom-data-625-test.xls*. Convert the file to CSV format and save it. Then load the file into Weka and save it in ARFF format.

(c) Reload the original ARFF file *Mushroom-data-625* and build a model using the J48 classifier. Make sure that your class attribute is set to *Status*.

(d) Under *Test options*, click on the radio button to the left of *Supplied test set* and then click on *Set* to the right of *Supplied test set*. Click on *Open file* in the small window that appears, select your ARFF file *Mushroom-data-625-test*, and open it as the file that will be used as test data.

(e) Make sure that your classification attribute is still set to *Status*, and then click on *Start*. You should get an error message saying that the training and test sets are not compatible.

(f) Let's look at the problem. Use an editor to view the ARFF file *Mushroom-data-625* that was used for training and the ARFF file *Mushroom-data-625-test* that was used for testing, and compare them. What differences do you see?

(g) The problem is that the training and testing data files must have the same set of possible values for the attributes and these possible values must be listed in the same order in the attribute definition. Copy the ARFF file *Mushroom-data-625-test* into file *Mushroom-data-625-test2* (*Mushroom-data-625-test2* will also be an ARFF file); then edit the ARFF file *Mushroom-data-625-test2* so that it has the same set of possible values for the attributes as the ARFF training file *Mushroom-data-625* and these possible values are listed in the same order in the attribute definition. Then save *Mushroom-data-625-test2*.

(h) Set the test file as the edited *Mushroom-data-625-test2*. Again click on *Start*. If you have edited the test file correctly, you should get a successful test run on the 10 instances in the test file.

(i) **Questions-7:**

i. How many instances were classified correctly?

ii. Draw the confusion matrix.