Installing Weka and Building a First Model

As discussed, we will be doing some data processing, model building, and evaluation using the data science/machine learning toolkit Weka. Your first task will be simply to get Weka installed and to run through building a model, to make sure Weka is working fine on your computer.

As your data mining task, you will build a classification tree model (see Chapter 3) from the mailing data set described in the Python tutorial. We will talk about this later in class, but to run Weka through its paces please proceed as follows:

1) Install the proper version of Weka for your machine. Get the installation file download at the following site, and install Weka on your machine: http://www.cs.waikato.ac.nz/ml/weka/downloading.html. Choose the "Stable book 3rd ed. version" (3.6) that corresponds to your computer. Windows users: If you do not know whether you have the Java VM 1.7, then load the version that includes the Java VM. Mac users: you can determine what version of Java you have by typing "java –version" in Terminal. (One of our other tutorials introduces you to Terminal.) If you don't know your Java version, you're probably safe assuming you have Apple's version 1.6 JVM. [FYI: Weka is written in Java. The Java VM is the "virtual machine" that runs Java programs; not important to us, except for choosing the right version of Weka for your computer.]

2) Start Weka. It should be available in the Start/Program menu on Windows. On the Mac it probably will appear in its own Finder folder, and you should move the application (with the blue circle icon and flightless bird) to the Applications folder. After starting Weka, select Explorer from the four choices.

3) Load the dataset "mailing_tree.arff" (available under Resources on NYU Classes). Weka has several panes in which you can work. You should start in the Preprocess pane. Load the data simply with Open file…

4) Build your classification tree. Go to the "Classify" pane. Pick J48 using "Choose"; J48 is under "trees". Before you run the algorithm, please change the minNumObj option to 100. How do you change the option? Strangely, you get to the option menu by clicking in the box next to the 'Choose' button. (In the box it should say 'J48 <etc.>' after you selected J48 above.) To build the classifier, click on "Start".

5) When the bird in the lower right-hand corner settles down, the run is done. Scrolling up and down in the big window will show you the results of the run. Cut the results from the large window and paste them at the end of your homework #1.

6) Look through everything in the window. We will cover almost all that's there in the next few weeks, as we learn about predictive modeling, supervised segmentation with tree induction, and model performance analytics.