National University of Singapore
School of Computing

SWS3018 Predictive Analytics
Lab 5

Learning Objectives

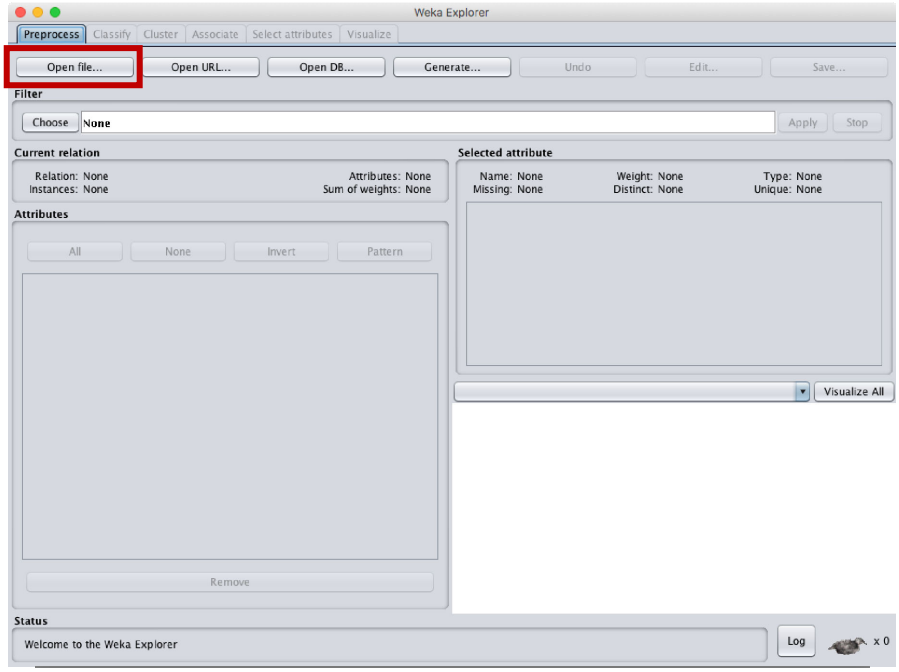
- Read/write with the ARFF format
- Perform Logistic Regression and Naïve Bayes classification using Weka

1. Weka typically works with arff format. R allows you to read/write the ARFF format. To do so, install the **foreign** package in R. After installing, see the help pages of **foreign** for a way to convert the `email_spam.csv` file (from tutorial 4) to `email_spam.arff`. Note: You should first convert spam column to yes, no factor type (`as.factor("character")`) from numeric before the conversion.

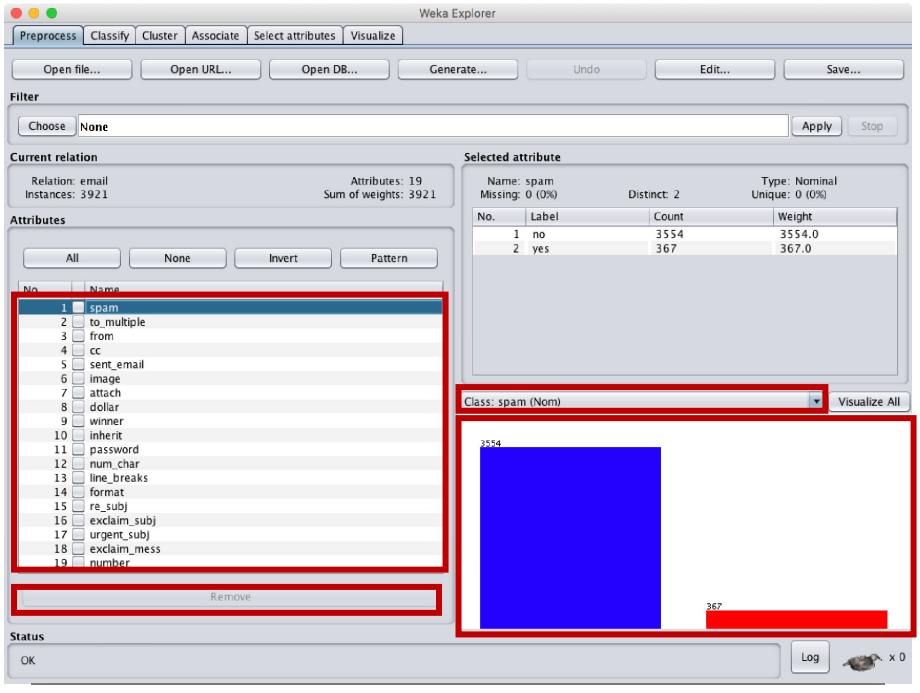
2. Assume that you have a dataset in arff format, how can you read it into a variable in R?

3. Try performing logistic regression using WEKA using the email spam dataset from tutorial 4 (`email_spam.arff`)

Click on Explorer

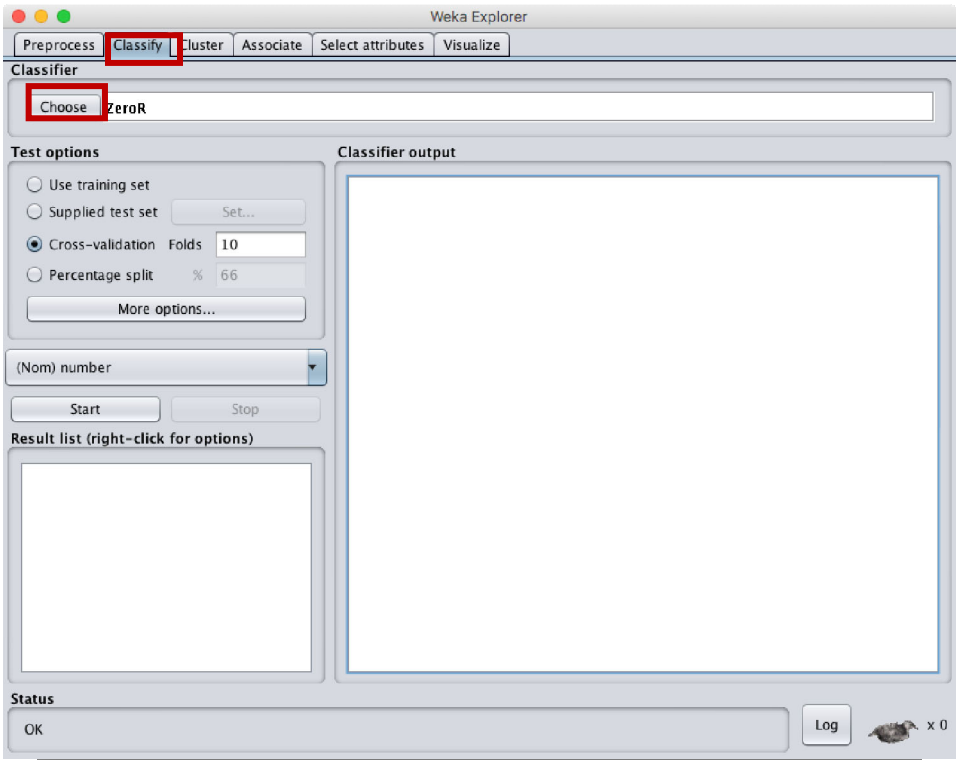


From the Weka Explorer window, click on “Open file” and point to `email_spam.arff`

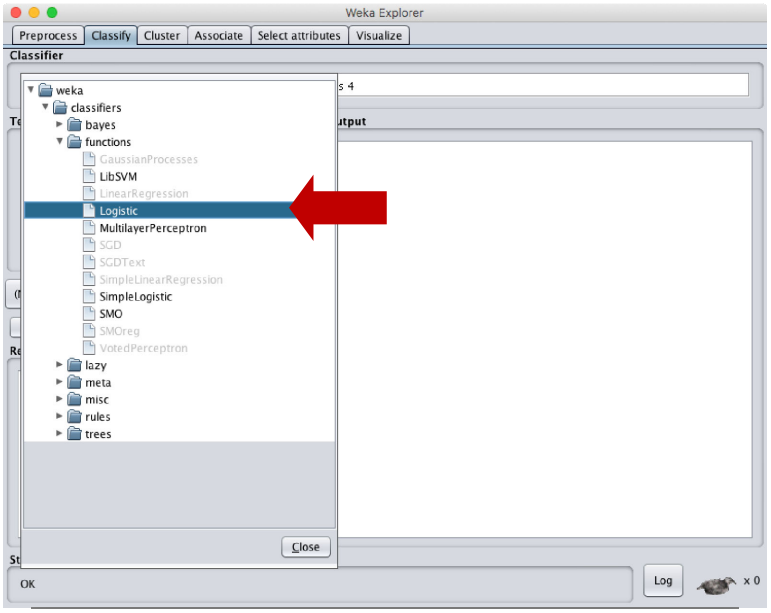


You should see something like this after opening the dataset. From here, you can select attributes/variables to remove (by default, all the attributes/variables are used). It will not modify the original data file. You can also see some charts of the breakdown of the

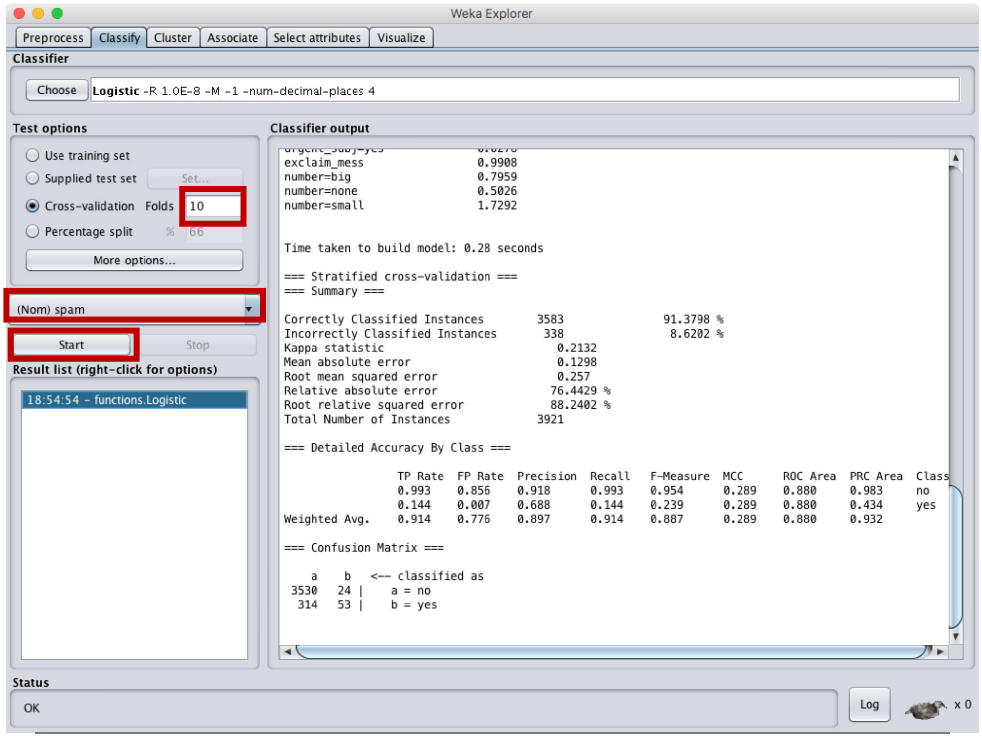
classes (ensure that you select the target/response – spam (Nom) in our case) against a predictor.



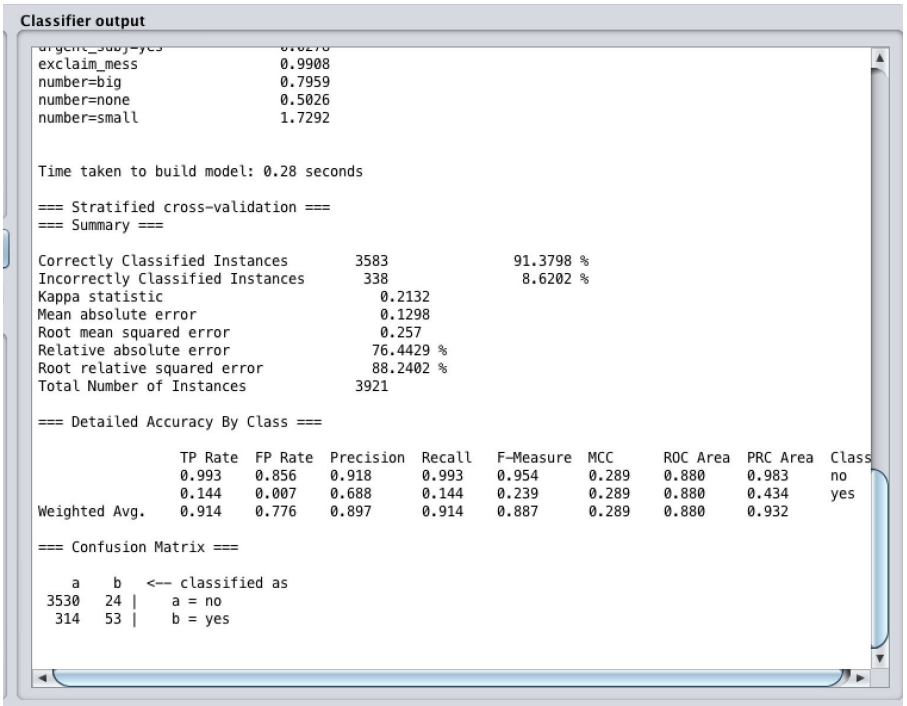
You will also notice that the **Classify** tab is now enabled and a classifier can be chosen.



Logistic regression can be found under classifiers ➔ functions ➔ Logistic



You can then select the attribute to be the response variable (spam) and enter the number of folds of cross-validation to do (the default is 10-fold cross-validation). When you are done, click on Start.



The classification results can be seen on the right box.

4. Try Naïve Bayes classifier (classifiers ➔ bayes ➔ NaïveBayes). Compare the classification accuracy as compared to Logistic Regression. How about the model building time?

5. Try removing all the attributes except spam and `to_multiple` and do classification of spam using the `to_multiple` predictor. What is the accuracy if only `to_multiple` is used for logistic regression and Naïve Bayes? Discuss.