HW5: Assigned 11/20, due 11/27 by 11:59 PM

Total points: 3+1+3+1=8

In this (last!) assignment, you'll use two popular, free, data mining tools - WEKA, KNIME - to analyze ("mine") some data.

Here (data/shells.arff) is the dataset to use. It consists of 4177 rows of data regarding abalone shells (https://www.google.com/search?q=abalone+shells&num=100&rlz=1C1CHBF_enUS723US723&source=lnms&tbm=isch&sa=X&ved=0ahUKEwihrc where each row resulted from measuring 9 parameters/features/values for each shell. The data is in text format (.arff format, for input to WEKA), do take a look at it. The idea is to be able to predict the 9th value, number-of-rings, given the other 8 values, using the existing dataset to learn how to predict.

First, download and install WEKA, and spend a small bit of time, going through the basics (it is all provided in the WEKA site): https://www.cs.waikato.ac.nz/ml/weka/ (https://www.cs.waikato.ac.nz/ml/weka/). Here (_WEKA/slides.html) is another intro' to WEKA.

Bring up WEKA Explorer, read in the shells.arff dataset (look under Preprocess). In the dialog that comes up, you can click on various attributes (age, length...) to see the distribution of data for that attribute. Next, click on Classify, to bring up the UI for choosing mining algorithms. Click Choose, then select classifiers->functions->LinearRegression. Execute with the default algorithm params (eg. cross-validation, 10 folds), look at the result.

0 points (-1 if you don't submit this). What is the MAE (mean absolute error)?

3 points. What is the equation for num_rings? You need to provide an equation, in the form of y=f(x_0,x_1,x_2...), using the output that WEKA provides. As you know, this is the result of the 'mining' - for a new shell, we can predict num_rings, by measuring the other params and inputting them into our equation.

1 point. Re-open the dataset, and keep only these columns: length, diameter, whole_weight, num_rings. Do the linear regression again. What is the equation now? The idea is that now, we would only need do just 3 measurements, to predict num_rings.

Next, download and install KNIME (https://www.knime.com/) ("nime"), and work through the quickstart tutorial. KNIME is also UI-driven, like WEKA; additionally, it's also visual-dataflow-driven, which means we can do data mining with it, by 'connecting the boxes' (where each box reads data or does mining or writes data, etc).

Use KNIME to perform linear regression, just like you did with WEKA. You need these nodes: AARF Reader, Linear Regression Learner. Create and connect the nodes, and execute.

2+1=3 points. What is the linear equation now? Compare this to WEKA's output - what parameters have similar coefficients (where they differ by 0.5 atmost)?

1 point. Set up a 'Decision Tree Learner' predictor, where 'Age' is the predicted variable. Note - think "simple" - no need to partition the data into training and test data, etc! Provide a snapshot (.jpg or .png) of the *entire* decision tree [OK if the nodes are too zoomed out and are therefore unreadable] - hint: look at the *right* side of the split-pane window.

BONUS (1 point). "Watch this space!"

Have fun! The point of this HW is to get you to become familiar with two useful data mining tools that enable analysis, using just point-and-click (no coding!). Also, you can now list 'Data mining tools: WEKA, KNIME' in your resume :)
