Weka Practical 2
Computing Science at University of Stirling

Premium Setting and Claim Prediction for Motor Insurance

1. Go to the assignments page on the course web site and down load the two .csv files for assignment 2. Store them in your own directory

2. Have a quick look at them in Excel. You will see that they both contain data describing motor insurance policies.
   a. The file MotorPremiums.csv contains policy details and the premiums charged on those policies.
   b. The file MotorClaims.csv contains the same policies with a field indicating whether or not a claim was made.

3. Put the premiums and the claims columns together in a new file and use a chart to see how well the premiums predict claims.

4. Your first task is to use the Premiums file to reproduce the insurance companies own set of pricing rules using the examples of prices in this file. Use a multilayer perceptron (MLP) in Weka with 50% test data. How well does the MLP reproduce the prices in the test data?

5. Next you need to use the claims data and produce a better pricing system.
   a. Using Weka, build a model that classifies claims as Yes or No. Try a few models including an MLP and an ID3 tree.
   b. When you have a model you like, verify it using cross validation.
   c. Look at the confusion matrix for your model – what is it telling you?

6. Now we are going to look at lift curves. Right click on your chosen model in the result list and choose Visualize Threshold Curve from the pop-up menu. Below this option will be two more – Yes and No. These are the names of your output class values. Choose ‘Yes’.

   a. The window you get allows you to plot various results from your test data to see how good the model is. You can vary what is plotted on the X and Y axis and also what the colour of the points tells you. The data is sorted by the probability of the output value you chose (‘Yes’ in this case). So, instances that are most likely to be Claim = Yes are plotted at the left hand side of the curve. The curves then assume that you class everything as ‘Yes’ regardless of the output. Those to the left are likely to be correct and those at the right are more likely to be wrong (as their score for Yes is low).
   b. The ROC curve can be seen by plotting True Positive Rate (Y axis) against False Positive Rate (X axis). The faster the curve climbs, the better the result.
   c. Try some other combinations of plotting TP, FP, Threshold, and Instance number and make sure you understand why the chart looks the way it does. Ask for help if you are not sure.