CM5239 Tutorial 6 – Data Analysis Using Weka

(1) Install Weka 3.8.6 on your computer by visiting the following link: http://www.cs.waikato.ac.nz/~ml/weka/index.html. Download the glass.arff file from the Canvas Tutorial file folder. Open the glass.arff file using Weka Explorer. How many attributes are present in the Glass data? For the attribute "Type," how many different labels are there? For the "Si" attribute, determine the minimum (Min), maximum (Max), mean, and standard deviation (StdDev) values. Generate a plot of Type vs. Fe by selecting "Visualize" from the top menu and choosing the plot from the "Plot Matrix" option. What conclusions can you draw from the plot?

(2) Use Notepad in our computer to create a file “qsar.arff” using the following data.

```
% isonarcotic activities of ketone, ester and ether
%
% compound        log P log(1/C)
% CH3COCH3              -0.73 0.65
% CH3CO2CH3       -0.38 1.10
% C2H5COCH3       -0.27 1.10
% HCO2C2H5              -0.38 1.20
% C2H5COC2H5            0.59  1.20
% CH3CO2C2H5            0.14  1.50
% C2H5COC3H7            0.31  1.50
% C3H7COCH3       0.31  1.70
% CH3CO2C3H7            0.66  2.00
% C2H5CO2C2H5           0.66  2.00
% (CH3)2CHCO2C2H5 1.05  2.00

@RELATION qsar

@ATTRIBUTE logP REAL
@ATTRIBUTE activity REAL

@DATA

-0.73,0.65
-0.38,1.10
-0.27,1.10
-0.38,1.20
0.59,1.20
0.14,1.50
0.31,1.50
0.31,1.70
0.66,2.00
0.66,2.00
1.05,2.00
%
```

Open the qsar.arff data file using Weka Explorer. Determine the Min, Max, Mean and StdDev values of logP and Activity? Perform a linear regression analysis to identify the best line fit. To do this, navigate to the Classify menu, choose “functions-LinearRegression” and click “Start” to view the result from the Classifier output. What is the linear regression model? What is the correlation coefficient for a 2-fold cross-fold validation.