COC131 Data Mining - Basics of Data Understanding

Martin D. Sykora m.d.sykora@lboro.ac.uk
Tutorial 01, Friday 15th February 2008

1. Download the Old Faithful data-set
2. Upload this data in Excel. There are 2 attributes and 2 classes. Sort the data by class (be carefull to sort the entire row), line or bar plot each of the features individually and save the graphs in a Word file. What do you notice on the plots from a visual inspection.
3. For each class feature, compute its minimum, maximum, mean and standard deviation.
4. Generate a pairwise scatter plot for the combinations of: F1 vs F2. Can you visually guess whether these attributes are related or not?
5. Based on the scatter plot generated in point 4, determine the data points that are the outliers (extreme high or low values). Do this manually by visually inspecting the scatter plot, remove at least 5 points.
6. Compute correlation between features for each class separately and create a correlation matrix. What does it show?
7. Normalise all features to the range [0, 1]. There are several ways this can be done, we will use the standard min-max normalisation $v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$, where min_A and max_A are the minimum and maximum values of an attribute A and v is the value we want to put into the range $[new\_min_A, new\_max_A]$. Recompute 6 has it made a difference?