F21DL Data Mining and Machine Learning: Coursework Assignment 2
Handed Out: Thursday 10th October 2013
What must be submitted: A report of maximum 3sides of A4, in PDF format
To be 'Handed in': 23:59pm Sunday November 17th 2013
-- by email to dwcorne@gmail.com with Subject Line: DMML Coursework 2
Worth: 40% of the marks for the module**.**

The point: confusion matrices, correlation and feature selection are all important in data mining and machine learning. So this coursework gives you experience with each of these things.

In this coursework you will work with only the Communities and Crime dataset. You will prepare it in the same way as detailed in the coursework 1 handout, except that you do not need to produce 2-class or normalised versions. Basically, just remove the useless and missing-value fields, and you can use the scripts I supplied in coursework 1 to do this. **However there is one important extra step: the class field in this dataset is a real number between 0 and 1. You need to convert this into ten distinct values (otherwise my Naïve Bayes program will not work properly). So, produce a version of the dataset where the class field is 0 for values between 0 and 0.1, 1 for values between 0.1 and 0.2, 2 for values between 0.2 and 0.3, and so on.**

You will be using my awk program for doing Naïve Bayes machine learning. This program internally discretizes each non-class field into 10 equal width bins, learns a simple Naïve Bayes probability model on the training set (the first 80% of the input field) and provides output giving the overall accuracy on the test set, and the confusion matrix calculated on the test set.
It is at http://www.macs.hw.ac.uk/~dwcorne/Teaching/DMML/nbFixed.awk

What to do

After the preparation indicated above, you will:

1. Produce a version of the dataset that has the instances in a randomised order.
2. Implement a program or script that allows you to work out the correlation between any two fields.
3. Using your program, find out the correlation between each field and the class field *using only the first 80% of instances in the data file*.
4. Using this information, run my Naïve Bayes awk script for each of the following 3 cases:
    4.1. Using only the top 5 non-class fields
    4.2. Using only the top 10 non-class fields
    4.3. Using only the top 20 non-class fields
5. Often it is useful or necessary to find a value for the correlation between a numeric field and a categorical field, or between two categorical fields. This cannot be done with Pearson's *r* value. Do some research (using the www) to find out how it can be done.

What to Submit

What you submit for this assignment is a report of maximum THREE sides of A4, containing the following.

1. up to half a page describing how you did steps 1,2 and 3.
2. up to two pages showing and discussing the results from step 4 (I expect this to include a display of the selected fields, and also a display and discussion of the confusion matrices)
3. up to half a page in a section with the title "Calculating correlation values for categorical data", explaining how this can be done for pairs of fields when either one or two of the pair is non-numeric.

The material for parts 1—3 must be all contained within 3 sides of A4. 20 marks are lost for every extra page, even if there is just one word on the page.

Marking: worth 40% of the module; of that 40%, the above parts break down as follows: 1(2), 2(30), 3(8).