F21DL Data Mining and Machine Learning: Coursework 2

Handed Out: Friday 9th October 2015
What must be submitted: A report of maximum 3 sides of A4, in PDF format
To be 'Handed in': 23:59pm Sunday November 29th 2015
-- 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 a 'handwritten digit recognition' dataset. Get it from my site here: http://www.macs.hw.ac.uk/~dwcorne/Teaching/DMML/optall.txt
Also pick up ten other versions of the same dataset, as follows:

http://www.macs.hw.ac.uk/~dwcorne/Teaching/DMML/opt0.txt
http://www.macs.hw.ac.uk/~dwcorne/Teaching/DMML/opt1.txt
http://www.macs.hw.ac.uk/~dwcorne/Teaching/DMML/opt2.txt
[etc…]
http://www.macs.hw.ac.uk/~dwcorne/Teaching/DMML/opt9.txt

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 50% of the input file) 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/nb.awk

In all these datasets, the first 64 fields correspond to an 8x8 array image, so that each field indicates the general amount of ink in a specific area of the image. The last (class) field indicates the specific handwritten digit. In optall.txt, the last field is the handwritten digit itself – i.e. if it is '8', then the image characterised by fields 1—64 was a handwritten '8'. In optN.txt, however, the class field is either 0 or 1, It is '1' if the handwritten digit was actually an 'N', and it is 0 otherwise. For example, in opt7.txt, if the last field is '1', then that instance corresponds to a handwritten '7'; if the last field is '0', then that instance might correspond to a handwritten 0, 1, 2, 3, 4, 5, 6, 8, or 9.

What to do

Everyone:

After collecting the files as above, you will:

1. Produce a version of optall.txt that has the instances in a randomised order.
2. Run my naïve Bayes awk script on the resulting version of optall.txt
3. Implement a program or script that allows you to work out the correlation between any two fields.
4. Using your program, find out the correlation between each field and the class field -- *using only the first 50% of instances in the data file* – for all the optN files. For each optN.txt file, keep a record of the top five fields, in order of absolute correlation value.

5. Using this information, run my Naïve Bayes awk script on optall.txt for each of the following 3 cases:

5.1. Using only the top 2 non-class fields from each optN.txt (i.e. use a reduced version of optall.txt where there are 10 (or perhaps fewer) non-class fields.)

5.2. Using only the top 3 non-class fields from each optN.txt

5.3. Using only the top 5 non-class fields from each optN.txt

Level 11 only (MSc students and MEng final year students):

6. 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.

Everyone:

1. HOW: up to a half page describing how you did steps 1 and 3.
2. RESULTS: up to two and a half pages showing and discussing the results from step 2 and 5 (I expect this to include a display of the selected fields, and display and discussion of the confusion matrices)

Level 11 only:

In addition, about half a page 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.

Your report 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. Level 11 students: My expectation is that your 'RESULTS' part will be two pages, leaving half a page for your extra part.

Marking: worth 40% of the module; of that 40%, the above parts break down as follows:
Level 10 students: 1(5), 2(graphs and tables: 15 ; discussion: 20).
Level 11 students: 1(5), 2(graphs and tables: 10 ; discussion: 15), 3 (10)