DMML Coursework 2

Joseph Davidson - 071729468

1 Introduction and methods

This the submission for coursework 2 of DMML written by Joseph Davidson. The methods used in this coursework made use of awk scripts written by myself.

For randomising the dataset, I wrote a script that would produce random numbers between 1 and 1994 (the number of records). When a number comes up that correlates to a record which hasn't yet been chosen, the record is outputted. If the record has already been chosen, the script will just ignore it. After a certain number of 'ignored' records, the script will just output the rest of the records so that the script doesn't run forever.

The script to work out the correlation between two fields calculated Pearson's r in the way shown in the lecture notes. It then worked out the fields which had the strongest correlations and outputted reduced datasets for the top 5, 10 and 20 fields that correlated with the class field.

2 Main results

In using the naïve Bayes script, I noticed that the confusion matrix seemed to be wrong. With that in mind, I dug into the code and found that the creation of the confusion matrix depended on the names of each class. So that I if I were to (say) change all the classes to names of Power Rangers, It would fail.

This is clearly wrong as the identifiers for the class values should have no effect whatsoever on the result. I corrected this by searching for the class in the classes array that the script creates and using the index of the found class instead of its actual value. This appeared to fix the results.

The top 20 fields for correlations is presented below. These are the fields that were used in the running of the naïve Bayes script on the dataset.

Below are the accuracies and confusion matrices for the corrected script for each reduced dataset. The green numbers represent the number of correct guesses that the script has performed.

Top 5 accuracy: 40.6015%
Top 10 accuracy: 41.604%
Top 20 accuracy: 40.8521%

At first glance, we think that ∼40% accuracy for each dataset is quite poor. But we have to consider that naïve Bayes is attempting to distinguish between 10 different classes. So the method is actually producing an excellent result. If we were to change the datasets to 2-class ones, we would probably get close to 100% accuracy.

It appears that class 0 for all the datasets was overwhelmingly the most correctly deduced. This was followed by 1 and then 2. This is because the number of records with a class of 0 is

Table 1: Top 20 correlating fields in the dataset.

Rank	Field	r number	Rank	Field	r number
1	44	-0.733622	11	41	0.548249
2	50	0.733372	12	38	0.520657
3	43	-0.700881	13	67	-0.520048
4	4	-0.681417	14	28	0.519576
5	45	-0.66305	15	32	0.50222
6	46	-0.65588	16	77	0.487983
7	3	0.627746	17	30	0.483021
8	18	0.573232	18	74	0.478652
9	16	-0.572737	19	69	0.468901
10	40	0.551585	20	49	0.463408

Table 2: Confusion Matrix for the naïve Bayes corrected script top 5 fields.

Actual Class and Values		Predicted Class									
Actual class 1	V=2	24	3	2	4	9	12	0	0	0	2
Actual class 2	V=0	12	88	0	0	35	1	0	0	0	0
Actual class 3	V=4	5	0	0	1	1	3	1	0	1	0
Actual class 4	V=5	10	0	0	1	0	2	3	0	2	2
Actual class 5	V=1	26	26	4	2	41	4	0	0	0	1
Actual class 6	V=3	18	1	0	2	6	3	1	0	2	4
Actual class 7	V=7	1	0	2	2	0	0	0	0	0	1
Actual class 8	V=8	1	0	0	1	0	0	0	0	1	4
Actual class 9	V=6	1	0	2	3	0	3	0	1	0	4
Actual class 10	V=9	0	0	1	0	0	0	1	0	0	5

Table 3: Confusion Matrix for the naïve Bayes corrected script on top 10 fields.

Actual Class and Values		Predicted Class									
Actual class 1	V=2	22	3	4	5	9	11	0	0	0	2
Actual class 2	V=0	13	88	0	0	35	0	0	0	0	0
Actual class 3	V=4	4	0	1	0	0	5	0	2	0	0
Actual class 4	V=5	7	0	1	2	0	3	3	0	2	2
Actual class 5	V=1	25	27	2	3	39	5	1	0	1	1
Actual class 6	V=3	12	0	2	2	7	8	0	0	1	5
Actual class 7	V=7	1	0	1	1	0	1	1	1	0	0
Actual class 8	V=8	1	0	0	1	0	0	1	0	1	3
Actual class 9	V=6	1	0	3	3	0	2	1	1	0	3
Actual class 10	V=9	0	0	0	0	0	0	0	1	1	5

713 – almost half the dataset. The accuracy doesn't vary so much between the 3 sets – with the highest accuracy coming from the 10 field dataset – which implies that the number of fields doesn't have much of an effect on the eventual accuracy of the method. Also, it seems that 10 fields is the optimal number for this dataset, any more or less and the method returns results with a sub-optimal accuracy.

Table 4: Confusion Matrix for the naïve Bayes corrected script on top 20 fields.

Actual Class and Values		**Predicted Class**									
Actual class 1	V=2	24	2	4	2	7	13	0	0	2	2
Actual class 2	V=0	14	83	0	0	39	0	0	0	0	0
Actual class 3	V=4	3	0	5	1	0	3	0	0	0	0
Actual class 4	V=5	6	0	3	2	0	4	5	0	0	0
Actual class 5	V=1	22	29	4	1	38	7	0	0	2	1
Actual class 6	V=3	12	0	3	2	8	5	1	1	2	3
Actual class 7	V=7	1	0	1	1	0	0	1	1	1	0
Actual class 8	V=8	1	0	0	2	0	0	1	0	1	2
Actual class 9	V=6	1	0	4	2	0	3	0	1	0	3
Actual class 10	V=9	0	0	0	0	0	0	1	0	1	5

3 Calculating correlation values for categorical data

If we have class fields that are alphanumeric, attempting to use Pearsons r value is a fools errand. In this case, we need to use other methods to find the correlation between a numeric field and a non-numeric one.

The most obvious solution is to substitute the names of the classes with numbers. If you use a consistent scheme, the data will be the same with numbers instead of names in the class field. This is the simplest way to do it but it may misrepresent the classes with the usage of the standard deviations on the class field because this implies that the classes have values attached to them, when in reality the numbers are just substitute names. (As an aside, I have no references for this part as it's a pretty self-evident solution to the problem).

Another method is Chi-Square tests [1]. These look at the interdependence of a variable with another categorical one. It uses the sum of squares of a variable and uses a supplied k as the variables degree of freedom. The strength of the association is given by a P value. A P value that is < 0.05 implies a strong correlation between the independent and dependent variables.

If we can order out categorical data, we can use *concordant and discordant pairs*. We can observe two pairs $\langle S, v_1\rangle$ $\langle S, v_2\rangle$ where S is our subject and v_i are the variables. The two pairs are concordant if a subject that is higher on one variable, is also higher on the other. They are discordant if one is lower than the other. We work out the number of concordant and discordant pairs (Calling them C D respectively) and work out the strength of the associations γ by means of the equation:

$$\gamma = \frac{C - D}{C + D} \tag{1}$$

This will give us a result in the range $-1 \leq \gamma \leq 1$. A positive result indicates a positive association and a negative one indicates a negative association. A γ that is close to zero indicates a very weak association.

There are more methods, but I'm running out of space here to explain all of them I hope this is enough.

[1] http://courses.ncssm.edu/math/Stat_Inst/PDFS/Categorical%20Data%20Analysis.pdf