DMML Coursework 2

Naïve Bayes Classifier

Preparing Data and Correlation Script

In the very first step after all the files that are required to do this coursework were loaded, all instances inside an optall.txt file was randomly shuffle by using this simple command in Linux command line terminal and put them into new file below.

```
$ shuf optall.txt > modoptall.txt
```

In part 3 I use R function to find correlation between each field and class field. A Script below showed the way to read files and find correlation. The number 2610 is 50% of all instances in all optN files.

```
> data0 <- read.table("C:\\Users\\Wonchana\\Desktop\\CW2\\opt0.txt",nrows = 2610,header = FALSE)
> temp <-  cor(data0)
> sort(abs(temp[,65]),decreasing =TRUE)
```

Function cor(datao) use Pearson's Correlation to calculate correlation of every columns and put it in temp variable then you can use temp[,65] command to extract only column 65 which is class field and abs() function return absolute value then I sort them with sort function.

Results and Discussion

After running R command in previous section, the list of top 5, top3 and top 2 fields that have strongest relationship to the category field below will be use to produced new version of optall.txt and run “nb.awk”.

Selected top 5 fields + category: `3 4 6 7 11 14 15 19 20 21 22 27 28 29 30 34 35 36 37 38 39 42 43 44 45 46 47 51 54 55 60 61 63 65`

Selected top 3 fields + category: `6 7 11 19 20 21 22 27 29 30 34 36 37 38 42 43 44 46 47 51 54 55 61 63`

Selected top 2 fields + category: `6 11 19 20 21 22 27 29 30 36 37 42 43 46 51 54 55 61 63`

Confusion matrix

Mod	3	6	5	1	4	8	9	0	2	7	Precision
3	227	0	2	0	0	12	9	0	4	1	89.02
6	2	272	1	2	2	1	1	0	0	0	96.8
5	5	0	233	2	2	3	19	0	3	1	86.94
1	1	3	1	224	0	9	12	0	13	1	84.85
4	2	1	2	1	234	5	2	0	0	9	91.41
8	1	1	2	15	3	223	7	0	3	0	87.45
9	11	0	4	1	10	3	219	0	0	7	85.88
0	0	0	1	0	5	1	1	242	0	1	96.41
2	5	0	0	5	0	13	3	0	214	1	88.8
7	1	0	0	0	6	1	2	0	1	273	96.13
Sensitivity	89.02	98.19	94.72	89.6	89.31	82.29	79.64	100	89.92	92.86	

From the table above, Naïve Bayes classifier performed really well. As you can see in precision column, it can 85% or more correctly classify each number. But in order to measure the performance of predictor,

the true positive rate (Sensitivity) of each number should be considered. An overall performance of each number, most of them are higher than 85% except “8” and “9”. Obviously, the data are all handwriting number, the distribution on ink into each pixel of a number like “1”,”2”, and “3” are similar to “8” and “9”.

Top5	**0**	**7**	**4**	**6**	**2**	**5**	**8**	**1**	**9**	**3**	**Precision**
0	257	0	5	1	0	1	2	0	0	0	96.62
7	0	232	9	0	2	5	1	2	7	0	89.92
4	0	7	242	3	1	2	2	3	0	0	93.08
6	1	0	4	247	0	1	2	2	0	0	96.11
2	0	2	2	2	227	0	9	5	10	3	87.31
5	0	0	1	1	0	230	1	1	16	1	91.63
8	0	0	2	1	5	3	222	24	9	0	83.46
1	0	0	3	2	11	1	17	209	18	3	79.17
9	0	7	5	0	0	6	6	5	218	11	84.5
3	2	7	0	0	1	3	10	3	18	226	83.7
Sensitivity	98.85	90.98	88.64	96.11	91.9	91.27	81.62	82.28	73.65	92.62	

The Fact that Naïve Bayes learn a pattern of ink in each pixel, it might not be a good idea to remove some pixel. Because every pixel contained a pattern if some certain pixels were removed like in Top5 table, the overall distribution will change and prediction performances is more likely to be dropped some like “3” is better.

Top3	**0**	**7**	**4**	**6**	**2**	**5**	**8**	**1**	**9**	**3**	
0	255	0	8	1	0	2	0	0	0	0	Precision
7	0	288	8	0	1	11	2	4	3	1	95.86
4	0	8	227	4	0	3	3	14	1	0	90.57
6	1	0	5	245	0	2	2	1	0	1	87.31
2	0	2	3	3	216	0	14	3	15	4	95.33
5	0	0	1	0	0	231	1	1	17	0	83.08
8	0	0	4	0	8	4	210	32	8	0	92.03
1	0	0	2	1	16	1	16	197	29	2	78.95
9	0	5	2	0	0	6	5	2	231	7	74.62
3	1	6	0	1	1	3	5	6	16	231	89.53
Sensitivity	99.22	93.2	87.31	96.08	89.26	87.83	81.4	75.77	72.19	93.9	

Top3 table took out even more pixel overall performances are continually dropped but “3”. Originally “3” distribution is similar to “9” but right now “9” distribution changed to become more like “1” so “3” performance is better.

Top2	**0**	**7**	**4**	**6**	**2**	**5**	**8**	**1**	**9**	**3**	**Precision**
0	255	0	5	0	3	3	0	0	0	0	95.86
7	0	224	7	0	1	7	4	5	8	2	86.82
4	0	5	218	5	0	3	5	22	2	0	83.85
6	0	0	8	239	0	0	4	6	0	0	93
2	0	1	4	3	207	0	11	4	15	15	79.62
5	0	0	4	0	5	229	0	0	13	0	91.24
8	0	2	7	0	8	6	192	33	14	4	72.18
1	0	0	1	2	14	3	18	205	20	1	77.65
9	0	5	4	0	12	7	13	4	210	3	81.4
3	1	8	1	0	21	4	3	4	10	218	80.74
Sensitivity	99.61	91.43	84.17	95.98	76.38	87.4	76.8	72.44	71.92	89.71	

After we run Top2, the results showed that all performances are dropped. Because at this point all record ink distribution are even more similar including “3”, now it become similar to “2”.

I conclude that reducing number of column will not increase performances of Naïve Bayes Classifier. Also if the data that we want to classify are more complex the effect of reducing column are unpredictable.

Correlation of non-numeric field.

Non-numerical data can possibly be categorised into 2 group of variable types. First type is ordinal data (Categories that have ranks) such as “High”, “Medium” and “Low”. You can calculate a correlation coefficient by Spearman’s rank correlation.

N=5	Exam marks	Grade	Mark ranking(X)	Grade ranking(Y)	Different of X and Y
1	55	C	1	1	0
2	59	C	2	1	1
3	68	B	4	2	2
4	70	A	5	3	2
5	65	B	3	2	1
6	85	A	6	3	3

Table from above contained two actual columns that you want to find a correlation coefficient between them. In spearman ranking, given the lowest value is 1, the next lowest is 2 and so on. Do the same to both columns. After you apply the ranking you can find correlation by using Pearson correlation coefficient on those 2 ranking columns.

The second group is nominal data such as colour like “red”, “green” and “blue”. As you can see, you can’t give a rank to a data.

N=5	Exam marks	Gender	Grade
1	55	Male	C
2	59	Female	C
3	68	Female	B
4	70	Male	A
5	65	Female	B
6	85	Female	A

Again with the similar data but this time we need to find an association between gender and grade. Both columns are categories data type. Cramer’s V may be a good choice to find a relation between these 2 columns. Unlike correlation, Cremer’s V can only tell the strange of relation between 2 columns but not a patter (negative or positive).

Gender	A	B	C
Male	1	0	1
Female	1	2	1
Total	2	2	2

By generate a table of frequency of each categories as table above. Cremer’s V use a formula below, “min r-1, c-1” is a minimum value of either the rows-1 orcolumns-1. The V value can be interpreted the same way as correlation.

$$V = \sqrt{\frac{\chi^2}{(N)(\min \text{ r -1, c -1})}}$$