Homework 2 Report

This is the report that summarized the accuracy obtained on the test set by using designed multivalued naive Bayes and MCAP logistic regression classifier, the parameters used, and explanation of the result.

For compile and run the test, please refer to the /nblr/README.txt.

1. Results of Naive Bayes

The following chart is the summarized experiment results of Naive Bayes classifier

		Naive Bayes		
text sequence	accuracy on ham	accuracy on spam	total accuracy	train time (s)
1	94.25%	96.15%	94.77%	59.52
2	96.84%	90.00%	94.98%	40.2
3	96.55%	86.92%	93.93%	27.39
4				
4, 0.95	94.25%	96.15%	94.77%	244.39
4, 0.9	94.25%	96.15%	94.77%	232.75
4, 0.85	94.25%	96.92%	94.98%	237.7
4, 0.8	94.25%	96.92%	94.98%	227.84
4, 0.75	93.97%	96.92%	94.77%	235.1
4, 0.7	93.97%	96.92%	94.77%	222.29
4, 0.65	94.25%	96.92%	94.98%	226.86
4, 0.6	93.68%	96.92%	94.56%	219.21
4, 0.55	93.68%	96.15%	94.35%	224.83
4, 0.45	92.82%	95.38%	93.51%	210.62

"**test sequence**" indicates the text sequence used in the training process, 1 for original sequence (including punctuation), 2 for word sequence (excludes punctuation), 3 for word sequence without stop words (the stop words list can be found in /nblr/stopwords), 4 for word sequence after using mutual information to select most relevant words to reduce the vocabulary size.

For "text sequence" 4, the value after comma stands the proportion of the size of vocabulary after attribute selection to the original word sequence (corresponding to "text sequence" 2) vocabulary size.

"**accuracy on ham**", "**accuracy on spam**", and "**total accuracy**" stands for the corresponding accuracies on ham and spam test set, and the total accuracy information.

"**train time**" is the training time, note that I include mutual information attributes selection inside the training, so text sequence 4 use more time than the other three.

I marked the best performance settings with red.

Analysis:

From the above chart, we can see that both word sequence (without punctuation) and applying mutual information for attribute selection can achieve the highest accuracies among the other methods. It is not strange that after applying mutual information for attribute selection, only highest relevant attributes are selected, so it can do well in this prediction task.

While excluding stop words can not achieve better performance than the previous ones, especially on spam test set. It is because that the if an email contains more these words, it is more prone to be spam, if we remove them from the training and prediction process, it will do harm to the prediction accuracy. And using mutual information can help us solve this problem.

2. Results of Logistic Regression

After several tries, I found that setting learning rate ita bellow 0.1, setting prior lambda as small as possible, and setting iteration number around 20~30 can achieve the highest accuracy, I extracted some experiment results in the chart bellow for analysis.

Logistic Regression

ita	lambda	iteration	accuracy on ham	accuracy on spam	total accuracy	train time (s)	text sequence
0.05	0.001	30	94.25%	97.69%	95.19%	166.645024061s	2
0.03	0	30	95.11%	96.15%	95.40%	166.439259052s	2
0.01	0.001	20	93.39%	96.92%	94.35%	120.853749037s	2
0.03	0	20	94.54%	97.69%	95.40%	116.499322891s	2
0.04	0	20	94.83%	96.92%	95.40%	122.534415007s	2
0.03	0.00001	30	95.11%	96.15%	95.40%	171.094090223s	2
0.035	0	20	94.83%	97.69%	95.61%	108.878090858s	2
0.035	0	30	94.54%	95.38%	94.77%	173.222009897s	2
0.03	0	30	96.84%	92.31%	95.61%	176.634380817s	3
0.035	0	30	96.55%	93.08%	95.61%	118.307888031s	3
0.035	0	25	96.55%	92.31%	95.40%	140.344359159s	3
0.01	0.001	20	97.13%	93.08%	96.03%	112.423712969s	3
0.02	0.001	20	97.13%	92.31%	95.82%	121.058682919s	3
0.03	0	30	94.83%	92.31%	94.14%	289.586478949s	4, 0.9
0.035	0.0001	30	94.54%	91.54%	93.72%	269.99255085s	4, 0.8
0.035	0	30	95.11%	91.54%	94.14%	257.785739183s	4, 0.7
0.01	0.001	20	94.25%	96.15%	94.77%	217.550170183s	4, 0.7
0.01	0	20	95.11%	96.15%	95.40%	216.745496035s	4, 0.7

“**ita**” is learning rate;
“**lambda**” is the prior parameter;
“**iteration**” is the iteration number;
“**accuracy on ham**”, “**accuracy on spam**”, and “**total accuracy**” stands for the corresponding accuracies on ham and spam test set, and the total accuracy information;
“**train time(s)**” is the training time, note that I include mutual information attributes selection inside the training, so text sequence 4 use more time than the other three.
“**test sequence**” indicates the text sequence used in the training process, 2 for word sequence (excludes punctuation), 3 for word sequence without stop words (the stop words list can be found in /nblr/stopwords), 4 for word sequence after using mutual information to select most relevant words to reduce the vocabulary size. Note that I do not use 1 (original sequence with punctuation) for avoiding computation overflow.

The rows are marked red whose “total accuracy” > 95%, and the best one among them is marked blue.

Analysis:

From the results, we can see that set learning rate <0.1, iteration number between 20~30, and prior parameter lambda as small as possible can achieve high prediction accuracy in all three text sequence settings.

For logistic regression, word sequence without stop words achieves highest accuracy of 96.03%, while the prediction performance of the other two text sequence settings are comparable, all of the three settings can achieve higher accuracy than multivalued naive Bayes classifier, but the difference is not significant, all around 95%, while the training time of naive Bayes is only a third of the logistic regression.

My implementation of using mutual information to select most relevant attributes to reduce the vocabulary size (text sequence 4) can not outperform the other two text sequence setting lies in that: 1. the prediction accuracy is pretty high, all of them are comparable (all around 95%~96%), so it is difficult to make it higher; 2. logistic regression will assign low weight to unimportant words during iterations, it is likely a attributes selection process.

Note: you can find the original test data in /nblr/doc/data.xls