Naive Bayesian Spam Filter

(COMP3608 - Artificial Intelligence - Assignment 2)
(by Hanley Weng - 309248434)

Aims and Objectives

The aim of this study is to implement a spam filter. Bayesian filtering is a popular method utilized in identifying spam email, however, it's function can be transferred to various other fields such as science, medicine, and engineering. For example, it has been utilized in astrophysics, under a program named AutoClass, where new classes of stars can be identified such as infra-red IRAS Low Resolution Spectral catalogue stars.

The implementation of this spam filter is achieved through the Naive Bayes algorithm, utilizing repeated stratified cross validation to assess accuracy, whoms results can be used in improving the algorithm's performance. Methods of text classification are utilized in this process. The performance of this implementation will be evaluated along existing methods.

Process

Data Preprocessing and Feature Selection

A subsection of Ion Androutsopoulos' LingSpam, a collection of regular and spam emails, were expanded and separated into two separate corpora, body and subject.

These corpora were then preprocessed to a list, with subtracted stop-words, of 200 words with the highest document frequencies (the number of documents a word appears in throughout a corpus).

The subject corpus initially consisted of 1200 unique words, whilst the body corpus consisted of 22'994. This was before a list of 551 stop words were ommitted, along with grouping strings of numbers together in the form #ndigit string, where n is the number of digits. The remaining words were then filtered to only contain more than 3 characters. This left the subject corpus with 899 unique words, and the body corpus with 20'001 .

The top 100 words from each corpus, along with their document frequency (DF) score can be found in the appendix. The selection makes sense, experiencially and logically. The subject corpus has words of much lower DF, as to be expected as their lengths are often much shorter than their bodies. Words prevaling in the subject list are also quite every-day objectives,nouns,

(summary, disc, english, book, etc.), whereas those in the body could be identified moreso with everyday communication (please, many, here, know, used), and more common words may have been ommitable from here. It is interesting to note that strings of four digit numbers hold the highest DF in both corpora, this may be attributed to common cultural norms such as abbreviated dates, years, prices, amongst other metrics.

Feature weight was then applied to each corpus using tf-idf on the 200 terms to construct two csvs representing the terms and their weighted appearance for each dcument in the corpus. The csvs also contained the terms as headers, along with the value ‘isSpam’, a documents associated class.

Naive Bayes

Naive Bayes was then implemented. The version implemented was for numeric attributes using the probability density function over a normal distribution. However, as many values equated to zero, this could result in a zero value when calculating the standard deviation for a probability density functions (pdf) which could result in infinite (or invalid) pdfs. As such, a minimum standard deviation lower bound was applied.

The minimum standard deviation was determined by roughly graphing (see minSd vs accuracy graph in appendix) out the accuracies of the naive bayesian algorithm over various different minimum standard deviations. A rough mutual (for both subject and body corpora) was chosen at 0.002.

Values were also initally normalized during the training stages from 0.0 to 1.0, with a buffer of 0.09 for future test data.

Extension

The naive bayesian implementation was then adapted to optimize accuracy for the body corpus. Utilizing the minimum standard deviation vs accuracy graph (in appendix), the minimum standard deviation was chosen to be 0.001. It was also decided that the normalization value could be addressed, and as such, a graph was generated to determine the accuracies of the algorithm at various degrees of normalization.

The resulting graph (available in appendix as normalizationValue vs accuracy), resulted in a change of the max value to normalize too from the data from 0.91 to 88.

Results

The following are the accuracy results of different existing (Weka) classifiers in comparison with the ones created (MyNBs). All classifiers were tested with 10 fold cross validation.

	Corpus: Subject	Corpus: Body
Classifier	Accuracy (%)	Accuracy (%)
ZeroR	66.67	66.67
OneR	69.83	82.17
1-NN	76.67	81.33
3-NN	76.67	83.83
NB	74.33	80.67
DT	66.67	92.33
MLP	76.67	92.33
SVM	78.67	88.5
MyNB1	80.18	80.47
MyNB2	81.84	90.09

Discussion

In comparing the performance of the Subject and Body corpora, the subject performed significantly worse throughout all the classifiers it was tested upon (with the exception of ZeroR which produced the same result as the corpora were produced from the same original corpus with equal class ratios). This can be attributed to the fact that there is significantly less data to work with with the subject corpus. It originally consisted of only 1200 unique terms whilst the body corpus consisted of 22994, giving the body corpus 19 times more terms to work with.

Weka's Naive Bayesian's accuracy on the subject corpus is a bit lower than MyNB1, having a mean difference of -5.85% which is significantly large. However, in comparing the body corpus accuracy of Weka's Naive Bayesian and MyNB1, there is only a mean difference in accuracy of 0.38%. It is notable that MyNB1 deals better on smaller quantities of data than Weka's classifiers situated in the Subject Corpus, being higher than all the other classifiers in that category. However, it's performance whilst classifying in the body corpus is only equal to or less than the performance of Weka's non-rule-based classifiers (of which Multi-Layered-Perceptrons, and Decision Trees (significantly more efficient than MLP), do the best, producing accuracies of 92.33% each).

With MyNB2's altered minimum standard deviation and normalization (mentioned in the process), compared to MyNB1, there was an improvement on the subject corpus to 82% and an improved accuracy on the body corpus by 12.5%, up to 90%.

Conclusions

In conclusion, over all the classifiers, rule-based (ZeroR, OneR) and lazy classifiers (K-NN) performed faster overall, but less accurately than bayes (NB), tree (DT), or function-based classifers (MLP, SVM). Of these, Trees and Function classifiers were the best, however, MLP had a significantly higher complexity than the rest.

MyNBs could out-performs weka's NB, significantly so in it's later iteration where it was improved by optimizing the minimum standard deviation as well as it's maximum normalizing value. It was interesting to note that the different corpora had different optimum minimum standard deviations, this could be attributed to the subject's less abundant data set along with it's different distributions of terms due to it's individual document sizes. NBs in this instance however, fail to be more accurate than weka's default trees or function classifiers.

As future work, the MyNBs could be implemented with automated methods of optimizing the normalization of the data set along with the minimum standard deviation applied to the data. Pre-processing of text could also be better improved through exploring bi-grams as feature terms instead of l-grams. Better filtering could be utilized, and terms could be identified by other factors instead of document frequency such as distribution or frequency of frequency scores. The data could also be post-processed to utilize a lot less, but significant, terms. Outside of the Naive Bayesian implementation, additional attributes could be acquired such as representations of linguistics, the ease of reading, and the general weighting of different 'topics' of an email, for example, within content awareness.

Reflection

This study has taught me much, especially in regards to different classifiers, how they perform, and their efficiencies. The most important concepts I have learnt include the reliability of the measured accuracy (and comparabilities) of a classifier through methods such as repeated 10-fold stratified cross validation. I also found it very intriguing how different initial data-sets (their size and contents) could so significantly effect classifiers, along with how they are normalized and treated in the algorithm's implementation. Initially I came into this assignment slightly dissappointed about exploring supervised classifiers over unsupervised ones, but at it's conclusion, I am glad I did as I have learnt much about the building blocks and outlook of these classifiers and have developed a knowledge of what their suitable contexts.

Appendix

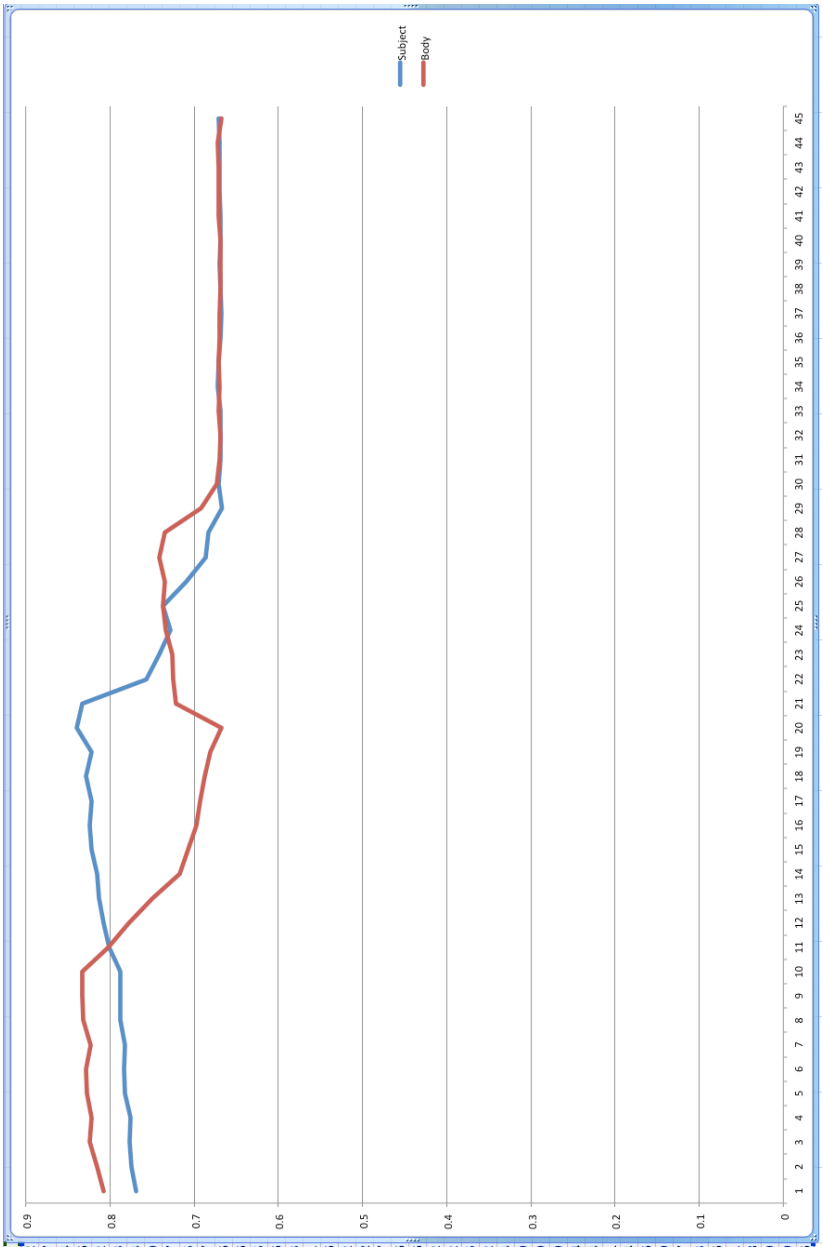
Top 100 words per Corpus by Document Frequency (DF)

Subject Word	DF score	Body Word	DF Score
#4digitString	35	#4digitString	346
summary	26	more	224
english	24	please	217
language	21	information	204
free	19	language	188
disc	19	university	179
query	18	time	171
linguistics	15	list	171
comparative	13	address	165
words	12	english	159
opposites	11	linguistics	156
book	10	http	156
method	9	many	155
email	9	here	150
call	9	know	147
japanese	8	following	147
correction	8	send	146
syntax	7	people	146
program	7	#5digitString	144
million	7	very	142
help	7	free	141
chinese	7	make	140
announcement	7	much	133
workshop	6	such	132
speaker	6	email	132
slip	6	work	128
money	6	number	128
lang	6	first	128
internet	6	mail	127
german	6	over	123
dick	6	well	120
conference	6	name	120
business	6	available	120
armey	6	languages	119
word	5	those	118
want	5	find	118
spanish	5	best	118
software	5	even	114

resources	5	same	111
research	5	want	109
please	5	order	108
part	5	need	108
needed	5	thanks	106
native	5	below	106
list	5	anyone	105
languages	5	call	103
know	5	take	101
jobs	5	research	100
grammar	5	form	100
better	5	being	100
best	5	years	98
unlimited	4	used	98
time	4	subject	98
systems	4	both	98
summer	4	help	97
request	4	each	97
read	4	contact	97
programs	4	world	96
phonetics	4	state	96
need	4	linguistic	96
mail	4	e-mail	96
linguist	4	through	95
intuitions	4	come	95
information	4	between	95
great	4	money	94
books	4	think	92
banning	4	possible	92
american	4	message	91
address	4	thank	90
youthese	3	different	89
world	3	before	89
video	3	word	88
verbal	3	receive	88
uniformitarianism	3	phone	88
tonight	3	using	87
thanks	3	good	87
teaching	3	further	87
teach	3	check	87
synthetic	3	interested	86
sites	3	working	85
site	3	include	85
secrets	3	case	85
school	3	note	83

resolution	3	year	82
requested	3	without	82
released	3	within	82
reference	3	again	82
profit	3	including	81
policy	3	mailing	80
people	3	made	80
opportunity	3	type	79
offer	3	program	79
names	3	place	79
more	3	home	79
misc	3	give	79
millions	3	special	78
make	3	several	78
live	3	right	78
lists	3	example	78
line	3	date	78
life	3	sent	77

Graph of Accuracies of MyNB with different minimum standard deviations
(Accuracies obtained with averaged 5 repetitions of 10-fold stratified cross validation)

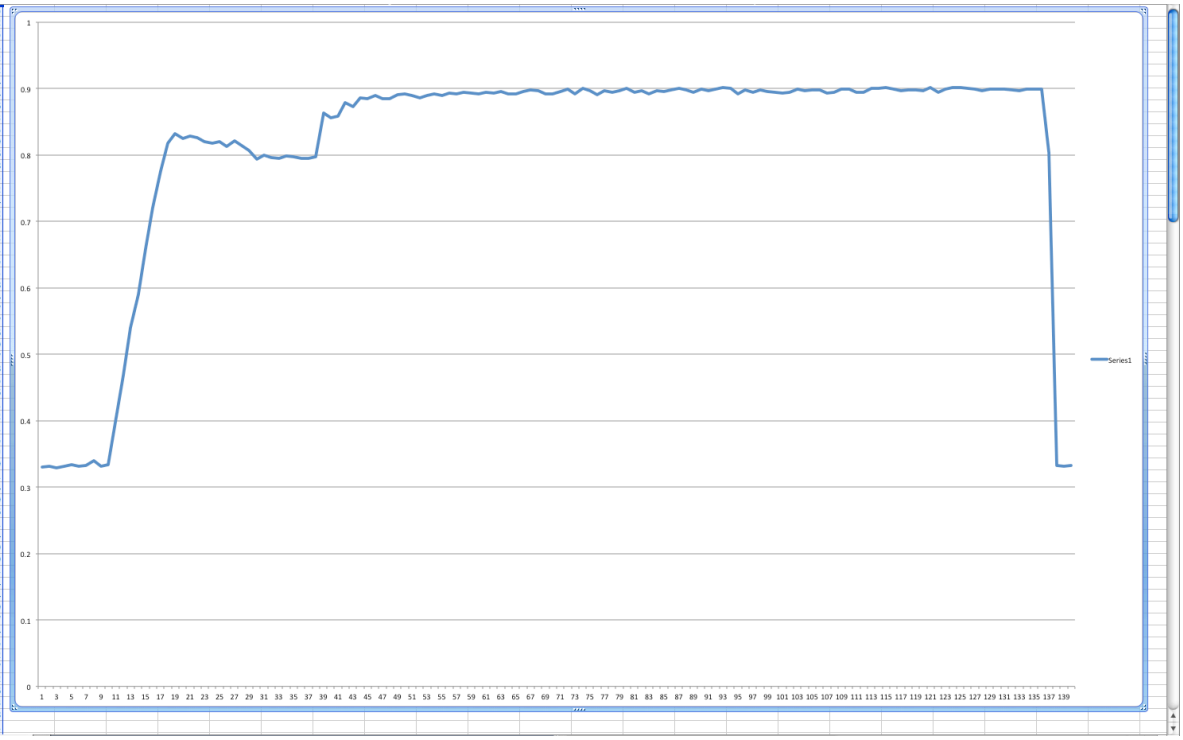


	minSd	Subject	Body
1			
2	0.0001	0.7686144	0.8075742
3	0.0002	0.7751489	0.8148617
4	0.0003	0.777321	0.8244574
5	0.0004	0.7759457	0.8225136
6	0.0005	0.782545	0.8275142
7	0.0006	0.7834874	0.8282823
8	0.0007	0.7818535	0.8234804
9	0.0008	0.7875786	0.8319689
10	0.0009	0.78775	0.8326377
11	0.001	0.7875373	0.8334475
12	0.002	0.8006591	0.8023757
13	0.003	0.8078557	0.7783906
14	0.004	0.8132083	0.7498026
15	0.005	0.8155236	0.7174245
16	0.006	0.8217766	0.7069836
17	0.007	0.8247284	0.6971173
18	0.008	0.8221528	0.6929481
19	0.009	0.828426	0.6877296
20	0.01	0.822068	0.6805802
21	0.02	0.8396738	0.667592
22	0.03	0.833528	0.7218647
23	0.04	0.7573072	0.7253216
24	0.05	0.7411116	0.7256446
25	0.06	0.7285387	0.7334072
26	0.07	0.7368265	0.7370252
27	0.08	0.7097652	0.7353095
28	0.09	0.6868313	0.7413892
29	0.1	0.683103	0.7351754
30	0.2	0.6670957	0.6920389
31	0.3	0.6710165	0.6731079
32	0.4	0.6692636	0.6703799
33	0.5	0.6693275	0.669144
34	0.6	0.6688327	0.6710944
35	0.7	0.6718849	0.669821
36	0.8	0.6707333	0.6707794
37	0.9	0.6693485	0.6701206
38	1	0.6672308	0.6704629
39	2	0.6691356	0.6689107
40	3	0.6694575	0.6693593
41	4	0.6690165	0.6692016
42	5	0.6692079	0.6707851
43	6	0.6698985	0.6709638
44	7	0.6696505	0.6705589
45	8	0.6698485	0.6716459
46	9	0.6710822	0.6679796

Graph of Accuracies of MyNB with different maximum normalized values of training set in classifier.

(Accuracies obtained with averaged 5 repetitions of 10-fold stratified cross validation)

1	0.01	0.32970145
2	0.02	0.33154122
3	0.03	0.32898105
4	0.04	0.33133641
5	0.05	0.33430932
6	0.06	0.33121352
7	0.07	0.33308244
8	0.08	0.33963134
9	0.09	0.33130536
10	0.1	0.333308
11	0.2	0.40461245
12	0.3	0.46861239
13	0.4	0.54006656
14	0.5	0.59056798
15	0.6	0.6583043
16	0.7	0.72082792
17	0.8	0.77562724
18	0.9	0.81742192
19	0.91	0.8317297
20	0.92	0.82478742
21	0.93	0.82841001
22	0.94	0.82672665
23	0.95	0.8199605
24	0.96	0.81807148
25	0.97	0.81996594
26	0.98	0.81326987
27	0.99	0.82118577
28	1	0.81429946
29	1.01	0.80767049
30	1.02	0.79382907
31	1.03	0.80033738
32	1.04	0.79638415
33	1.05	0.79554306
34	1.06	0.79839241
35	1.07	0.79806332
36	1.08	0.79503931
37	1.09	0.79557245
38	2	0.7971059
39	3	0.86300429
40	4	0.85665131
41	5	0.85818884
42	6	0.87949309
43	7	0.87277636
44	8	0.88614951
45	9	0.88524834
46	10	0.88984099
47	11	0.88520089
48	12	0.8844244
49	13	0.89156154
50	14	0.89242594
51	15	0.88991849
52	16	0.88621247
53	17	0.89025094
54	18	0.89216725
55	19	0.88985402
56	20	0.89362677
57	21	0.89276882
58	22	0.89461916
59	23	0.89298143
60	24	0.89188458
61	25	0.89514691



(also available as xlxs: http://goo.gl/7YUTk)

Bibliography

- (Program) Weka, http://www.cs.waikato.ac.nz/ml/weka/, *University of Waikato, New Zealand,* 2010
- Ames Research Cente, AutoClass C, http://ti.arc.nasa.gov/tech/rse/synthesis-projects-applications/autoclass/autoclass-c/, *National Aeronautics And Space Administration (NASA),* 2009
- Fabrizio Sebastiani, Machine learning in automated text categorization, *ACM Computing Surveys*, 34(1):1-47, 2002
- I. Androutsopoulos, J. Koutsias, K.V. Chandrinos, George Paliouras, C.D. Spyropoulos, An Evaluation of Naive Bayesian Anti-Spam Filtering, *Proc. of Workshop on Machine Learning in the New Information Age, 11th European Conference on Machine Learning, Barcelona, Spain,* 2000
- Jon Kågströ, Improving Naive Bayesian Spam Filtering, *Mid Sweden University, Department of Information Technology and Media,* 2005
- Paul Graham, Better Bayesian Filtering, http://www.paulgraham.com/better.html, 2003
- William A. Gale, Good-Turing Smoothing Without Tears, *AT&T Bell Labratories*, 2001