Solution (3.2)

Accuracy of Naive Bayes classifier over testing data: 0.785209860093

Confusion Matrix is as follows:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
A	249	0	0	0	0	1	0	0	1	0	0	2	0	3	3	24	2	3	4	26
B	0	286	13	14	9	22	4	1	1	0	1	11	8	6	10	1	2	0	0	0
C	1	33	204	57	19	21	4	2	3	0	0	12	5	10	8	3	1	0	5	3
D	0	11	30	277	20	1	10	2	1	0	1	4	32	1	2	0	0	0	0	0
E	0	17	13	30	269	0	12	2	2	0	0	3	21	8	4	0	1	0	1	0
F	0	54	16	6	3	285	1	1	3	0	0	5	3	6	4	0	1	1	1	0
G	0	7	5	32	16	1	270	17	8	1	2	0	7	4	6	0	2	1	2	1
H	0	3	1	2	0	0	14	331	17	0	0	1	13	0	4	2	0	0	6	1
I	0	1	0	1	0	0	2	27	360	0	0	0	3	1	0	0	1	1	0	0
J	0	0	0	1	1	0	2	1	2	352	17	0	1	3	3	5	2	1	5	1
K	2	0	1	0	0	0	2	1	2	4	383	0	0	0	0	1	2	0	1	0
L	0	3	0	3	4	1	0	0	0	1	1	362	2	2	2	0	9	0	5	0
M	3	20	4	25	7	4	8	11	6	0	0	21	264	9	7	1	3	0	0	0
N	5	7	0	3	0	0	3	5	4	1	0	1	8	320	8	7	6	5	8	2
O	0	8	0	1	0	3	1	0	1	0	1	4	6	5	343	3	2	1	12	1
P	11	2	0	0	0	2	1	0	0	0	0	0	0	2	0	362	0	1	2	15
Q	1	1	0	0	0	1	1	2	1	1	0	4	0	5	2	1	303	5	23	13
R	12	1	0	1	0	0	1	2	0	2	0	2	1	0	0	6	3	326	18	1
S	6	1	0	0	1	1	0	0	0	0	0	5	0	10	6	2	63	6	196	13
T	39	3	0	0	0	0	0	0	1	1	0	1	0	2	6	27	10	3	7	151

Where the labels have following meaning:

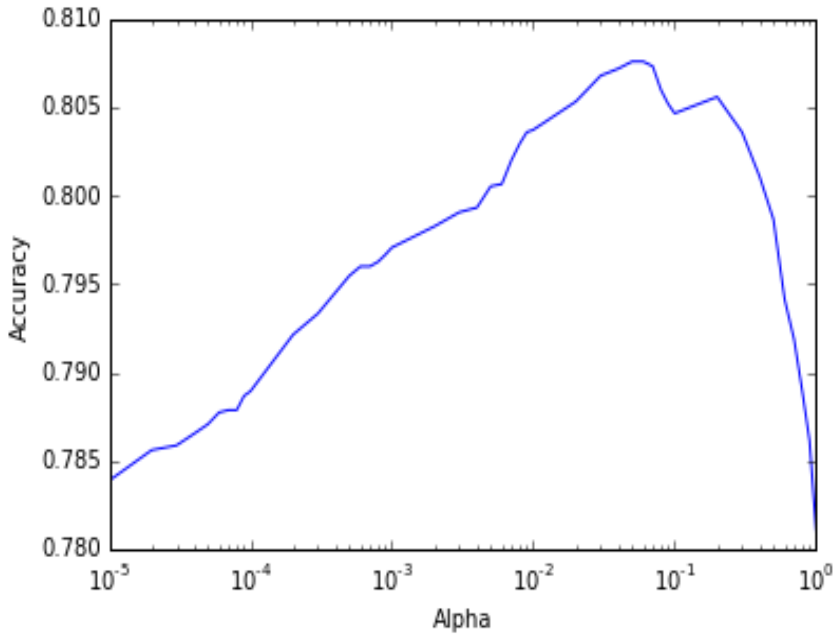
{**A**: 'alt.atheism', **B**: 'comp.graphics', **C**: 'comp.os.ms-windows.misc',
D: 'comp.sys.ibm.pc.hardware', **E**: 'comp.sys.mac.hardware', **F**: 'comp.windows.x',
G: 'misc.forsale', **H**: 'rec.autos', **I**: 'rec.motorcycles', **J**: 'rec.sport.baseball',
K: 'rec.sport.hockey', **L**: 'sci.crypt', **M**: 'sci.electronics', **N**: 'sci.med',
O: 'sci.space', **P**: 'soc.religion.christian', **Q**: 'talk.politics.guns',
R: 'talk.politics.mideast', **S**: 'talk.politics.misc', **T**: 'talk.religion.misc' }

Solution (3.3)

From the confusion matrix, it is clear that newsgroups with a similar topics are confused frequently. Notably, those related to computers (eg comp.os.ms-windows.misc and comp.sys.ibm.pc.hardware), those related to politics (e.g. talk.politics.guns and talk.politics.misc), and those related to religion (alt.atheism and talk.religon.misc). Newsgroups with similar topics have similar words that identify them. For example, we would expect the computer-related groups to all use computer terms frequently.

Solution (3.4)

The plot representing the accuracy of our trained Naive Bayes classifier for different values of alpha is as follows:



For very small values of alpha, the probability of rare words not seen during training for a given class tends to zero. There are many testing documents that contain words seen only in one or two training documents, and often these training documents are of a different class than the test documents. As alpha tends to zero, the probability of these rare words tends to dominate as the priors are being ignored.

For large values of alpha, the model underfits the training data: the final parameter estimates are close tend toward the prior as alpha increases. In particular, the classifier tends to underestimate the importance of rare words.