Ling 572 HW3

Michael Roylance, Olga Whelan

Q1 train accuracy mean = 0.9444444444444444

test accuracy mean = 0.8966666666666666

Q3

cond_prob_delta	Training accuracy	Test accuracy
0.1	0.930740740741	0.88
0.5	0.910740740741	0.863333333333
1.0	0.897407407407	0.84
2.0	0.879259259259	0.823333333333

Q4

cond_prob_delta	Training accuracy	Test accuracy
0.1	0.957777777778	0.91
0.5	0.950740740741	0.91
1.0	0.944814814815	0.893333333333
2.0	0.935555555556	0.89

Q5

cond_prob_delta	Training accuracy	Test accuracy
0.1	0.95962962963	0.903333333333
0.5	0.956666666667	0.903333333333
1.0	0.952962962963	0.896666666667
2.0	0.945185185185	0.896666666667

Q6 (a) First, we can see that Multinomial NB does better than Bernoulli NB, both in training and testing. Possibly this is because Bernoulli performs worse while classifying relatively long documents, mainly because it does not account for multiple occurrences of the words.

Second observation is that binarized Multinomial does even better than non-binarized. Although it only uses binary features as well as Bernoulli, it can be seen that the two algorithms produce significantly different results.

Third, for Bernoulli NB accuracy drops as the smoothing delta grows. This happens in Multinomial NB too, but it is much faster for Bernoulli, maybe because Bernoulli is more sensitive to noisy features.

(b) Not really. Although some features have a noticeably higher probability than the others, they still differ only by order of 10-100, rarely 1000.

(c) Yes. The more frequent a feature is in the class, the more important it becomes. Important features that were encountered in the class a lot of times, have a much higher probability than the rest, usually by order of 10^5-10^6.

(d) Bernoulli NB runs considerably slower than Multinomial NB. This is because for each tested document Bernoulli has to calculate the probability over every feature in the class as well as for every feature not in the class (i.e. it takes into account the non-occurring terms within the document). Multinomial is faster because it is only concerned with the features found in the document.