Nialls Chavez
CS529 - Machine Learning
Project 2

Question 1: In your answer sheet, explain in a sentence or two why it would be difficult to accurately estimate the parameters of this model on a reasonable set of documents (e.g. 1000 documents, each 1000 words long, where each word comes from a 50,000 word vocabulary).

Answer

This would be difficult to accurately estimate with a larger data set because of the amount of time it would take to compute. This calculation would create a huge dataset of MAP values which would need to be traversed for every word in every document which would be very costly in terms of computation time.

Question 2: In your answer sheet, report your overall testing accuracy (Number of correctly classified documents in the test set over the total number of test documents), and print out the confusion matrix (the matrix C, where cij is the number of times a document with ground truth category j was classified as category i). [10 points]

Answer

newsgroup 1 correctly labeled: 249 and incorrectly labeled: 69 Percentage: 0.783018867925
newsgroup 2 correctly labeled: 286 and incorrectly labeled: 103 Percentage: 0.735218508997
newsgroup 3 correctly labeled: 204 and incorrectly labeled: 187 Percentage: 0.521739130435
newsgroup 4 correctly labeled: 277 and incorrectly labeled: 115 Percentage: 0.706632653061
newsgroup 5 correctly labeled: 269 and incorrectly labeled: 114 Percentage: 0.702349869452
newsgroup 6 correctly labeled: 285 and incorrectly labeled: 105 Percentage: 0.730769230769
newsgroup 7 correctly labeled: 270 and incorrectly labeled: 112 Percentage: 0.706806282723
newsgroup 8 correctly labeled: 331 and incorrectly labeled: 64 Percentage: 0.837974683544
newsgroup 9 correctly labeled: 360 and incorrectly labeled: 37 Percentage: 0.906801007557
newsgroup 10 correctly labeled: 352 and incorrectly labeled: 45 Percentage: 0.886649874055
newsgroup 11 correctly labeled: 383 and incorrectly labeled: 16 Percentage: 0.959899749373
newsgroup 12 correctly labeled: 362 and incorrectly labeled: 33 Percentage: 0.916455696203
newsgroup 13 correctly labeled: 264 and incorrectly labeled: 129 Percentage: 0.671755725191
newsgroup 14 correctly labeled: 320 and incorrectly labeled: 73 Percentage: 0.814249363868
newsgroup 15 correctly labeled: 343 and incorrectly labeled: 49 Percentage: 0.875
newsgroup 16 correctly labeled: 362 and incorrectly labeled: 36 Percentage: 0.909547738693
newsgroup 17 correctly labeled: 303 and incorrectly labeled: 61 Percentage: 0.832417582418
newsgroup 18 correctly labeled: 326 and incorrectly labeled: 50 Percentage: 0.867021276596
newsgroup 19 correctly labeled: 196 and incorrectly labeled: 114 Percentage: 0.632258064516
newsgroup 20 correctly labeled: 151 and incorrectly labeled: 100 Percentage: 0.601593625498

The accuracy of our data is: 0.785209860093

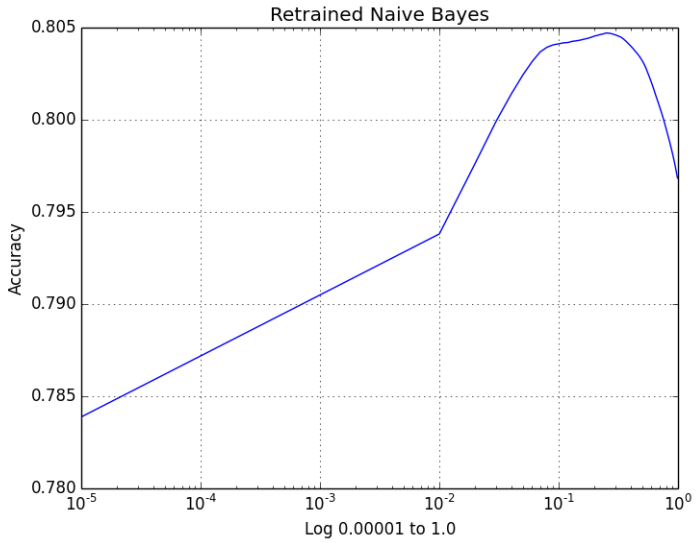
Question 3: Are there any newsgroups that the algorithm confuses more often than others? Why do you think this is? [5 points]

Answer

Yes, talk.politics.misc, talk.religion.misc, comp.os.ms-windows.misc, I think this is because these articles have the highest concentration of extra random data, particularly information like comments header data.

Question 4: Re-train your Naive Bayes classifier for values of β between .00001 and 1 and report the accuracy over the test set for each value of β. Create a plot with values of β on the x-axis and accuracy on the y-axis. Use a logarithmic scale for the x-axis (in Matlab, the semilox command). Explain in a few sentences why accuracy drops for both small and large values of β [5 points]

Answer



The value drops for small and large values of beta because alpha functions as the normalizing factor as we are calculating our MAP estimates. What we are seeing here is an overfitting due to alpha. As alpha increases, to a point, it helps us increase the accuracy of our estimations, until we start overfitting, at which point we see a severe drop in accuracy.

Question 5: Propose a method for ranking the words in the dataset based on how much the classifier relies on them when performing its classification (hint: information theory will help). Your metric should use only the classifier's estimates of P(Y) and P(X|Y). It should give high scores to those words that appear frequently in one or a few of the newsgroups but not in other ones. Words that are used frequently in general English ('the', 'of', etc.) should have lower scores, as well as words that only appear appear extremely rarely throughout the whole dataset. Finally, in your method there should be an overall ranking for the words, not a per-category ranking.[5 points]

Nialls Chavez
CS529 - Machine Learning
Project 2

Answer

My idea is to make two separate lists of words each which influence the estimation in different ways. The first list would contain only words that appear in a single label by the count of those words per label, and the second list would contain words which occur in every label. When we encounter either of these words in a new article, we will add or subtract weight to the word depending on which list it came from. The weights for the singularly existing words will be weighted the same, we will add a value of 75% of the MAP value per word discovered to each probability containing the word found, whereas for each word that is determined to exist in each label, we will subtract 50% of the MAP value from each probability containing these words. This will bias the probability toward infrequent words that only occur in certain labels.

Question 6: Implement your method, set β back to 1/|V|, and print out the 100 words with the highest measure. [5 points]

(29) the: 144433	(16) from: 11109	(83) who: 6494	(1030) does: 4070
(33) to: 72117	(104) by: 10900	(930) out: 6114	(1319) time: 4025
(12) of: 64170	(297) at: 10657	(142) which: 6052	(1015) then: 3957
(23) and: 56701	(139) an: 9969	(67) people: 5914	(823) these: 3666
(30) in: 48337	(749) there: 9638	(863) don: 5862	(942) should: 3605
(60) is: 42727	(477) what: 9560	(44) like: 5773	(137) new: 3566
(233) that: 39128	(995) my: 9465	(458) more: 5707	(902) good: 3521
(42) it: 33357	(235) all: 9219	(340) when: 5659	(792) could: 3468
(81) for: 27074	(766) will: 9206	(828) just: 5574	(245) well: 3452
(474) you: 26534	(813) we: 9067	(49) their: 5467	(978) am: 3355
(251) this: 19929	(100) one: 8976	(663) were: 5385	(1018) because: 3347
(48) on: 19683	(1003) would: 8883	(912) up: 5259	(492) even: 3277
(144) be: 19130	(877) do: 8514	(25) other: 5134	(299) very: 3229
(27) are: 18456	(301) he: 8399	(476) know: 5087	(143) may: 3200
(722) not: 18230	(304) about: 8145	(438) only: 4987	(921) now: 3186
(922) have: 18037	(778) writes: 7844	(1028) how: 4939	(31) us: 3090
(52) with: 17008	(99) so: 7709	(73) get: 4933	(745) why: 3072
(388) as: 15433	(80) com: 7498	(748) them: 4792	(368) into: 3067
(122) or: 14501	(466) has: 7496	(307) than: 4677	(131) see: 3057
(473) if: 13598	(969) your: 7474	(312) had: 4618	(2045) apr: 3053
(51) but: 13380	(886) no: 7314	(630) think: 4524	(1576) two: 2965
(160) they: 13154	(239) any: 6891	(467) been: 4520	(574) way: 2960
(644) was: 12963	(770) article: 6747	(295) his: 4507	(456) first: 2920
(775) edu: 12289	(850) me: 6721	(531) also: 4302	(316) god: 2898
(72) can: 11178	(314) some: 6588	(282) use: 4143	(289) many: 2876

Question 7: If the points in the training dataset were not sampled independently at random from the same distribution of data we plan to classify in the future, we might call that training set biased. Dataset bias is a problem because the performance of a classifier on a biased dataset will not accurately reflect its future performance in the real world. Look again at the words your classifier is ‘relying on’. Do you see any signs of dataset bias? [5 points]

Nialls Chavez
CS529 - Machine Learning
Project 2

Answer

I do see signs of bias, these words were not just taken from the articles themselves but also contained words with typos, comments on articles, and bits of meta data about the articles. What this tells us is that depending on where the “noise” came from different labels will have different biases. These biases will highly affect the prediction that we come up with in the end and will make it more and more difficult to classify articles going forward. We are also ranking highly words that are very commonplace in the english language such as the words shown above in question 6. a good example of how this can be detrimental is the word “com”. This word has 7498 occurrences and has multiple different uses in the english language and may just be a url extension, so if an article has multiple links in each of its articles, it will be harder for our algorithm to accurately classify these articles.