Homework 4 Report

Rajiv Anisetti, UID: 904801422

November 18, 2018

1 Naive Bayes

My accuracy table is as follows:

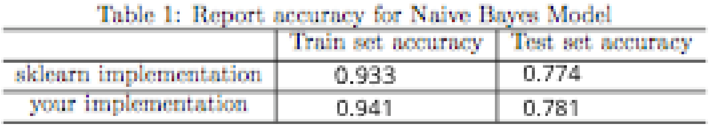
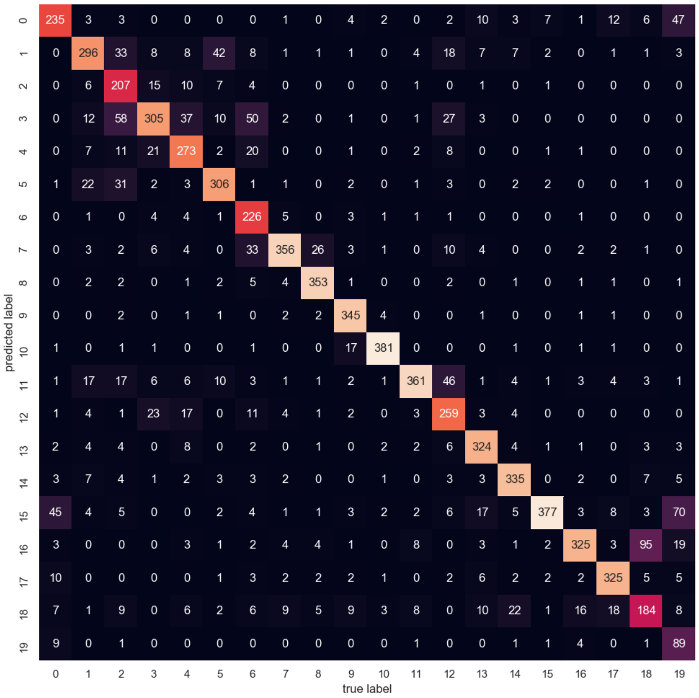


Table 1: Report accuracy for Naive Bayes Model

	Train set accuracy	Test set accuracy
sklearn implementation	0.933	0.774
your implementation	0.941	0.781

My classification matrix is as follows:



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
0	235	3	3	0	0	0	0	1	0	4	2	0	2	10	3	7	1	12	6	47
1	0	296	33	8	8	42	8	1	1	1	0	4	18	7	7	2	0	1	1	3
2	0	6	207	15	10	7	4	0	0	0	0	1	0	1	0	1	0	0	0	0
3	0	12	58	305	37	10	50	2	0	1	0	1	27	3	0	0	0	0	0	0
4	0	7	11	21	273	2	20	0	0	1	0	2	8	0	0	1	1	0	0	0
5	1	22	31	2	3	306	1	1	0	2	0	1	3	0	2	2	0	0	1	0
6	0	1	0	4	4	1	226	5	0	3	1	1	1	0	0	0	1	0	0	0
7	0	3	2	6	4	0	33	356	26	3	1	0	10	4	0	0	2	2	1	0
8	0	2	2	0	1	2	5	4	353	1	0	0	2	0	1	0	1	1	0	1
9	0	0	2	0	1	1	0	2	2	345	4	0	0	1	0	0	1	1	0	0
10	1	0	1	1	0	0	1	0	0	17	381	0	0	0	1	0	1	1	0	0
11	1	17	17	6	6	10	3	1	1	2	1	361	46	1	4	1	3	4	3	1
12	1	4	1	23	17	0	11	4	1	2	0	3	259	3	4	0	0	0	0	0
13	2	4	4	0	8	0	2	0	1	0	2	2	6	324	4	1	1	0	3	3
14	3	7	4	1	2	3	3	2	0	0	1	0	3	3	335	0	2	0	7	5
15	45	4	5	0	0	2	4	1	1	3	2	2	6	17	5	377	3	8	3	70
16	3	0	0	0	3	1	2	4	4	1	0	8	0	3	1	2	325	3	95	19
17	10	0	0	0	0	1	3	2	2	2	1	0	2	6	2	2	2	325	5	5
18	7	1	9	0	6	2	6	9	5	9	3	8	0	10	22	1	16	18	184	8
19	9	0	1	0	0	0	0	0	0	0	0	1	0	0	1	1	4	0	1	89

As the documents were rather large, I included an additional text file, named misclassified.txt that shows an example of a misclassified document.

1.1 Followup Questions

Naive Bayes is a generative model. Generative models learn a *joint* probability distribution rather than a *conditional* probability distribution. In essence, Naive Bayes assumes that each feature (in our example, words) is conditionally independent. Logistic regression, on the other hand, is a discriminative model and does not make the same assumption. As a result, it learns a conditional probability distribution.

Naive Bayes has several pros and cons. In terms of pros, the model is relatively easy to implement and simple in terms of complexity. For cons, Naive Bayes makes the assumption that none of the data points are conditionally dependent, which may not be a realistic assumption in real world data. If this assumption is not valid, Naive Bayes may fail to link such dependencies which can result in a lower classification accuracy, hence why it is called 'naive'.

Naive Bayes could definitely be used to classify spam emails from normal ones. Many spam emails consist of keywords that trigger even the human mind that the email is spam. Words such as "free", "millions", "buy", etc. Naive Bayes could see these words as a high probability of indicating a spam email and could classify emails accurately. This could be done through training the model with a training dataset consisting of both spam and non-spam emails, where the Naive Bayes model could then learn the spam probabilities of each word in the dictionary, realizing which words are more likely to be in a spam email.

2 pLSA

2.1 Appropriate K values

For dataset 1, I found that a K value of around 30 was best for maximizing log likelihood. This gave me a log likelihood of around -3000. For dataset 2, the dataset is much bigger. Because of this, even K values of 100 give a log likelihood of around -75000. I tuned the hyperparameters some more, even trying with K = 20, and number of words in topic = 50, but this led to a lower likelihood of around -98000. Thus, I would say the correct K for the second dataset could be 100 or above. This is a very large text file, so there may be many topics.

2.2 Topic Words

2.2.1 Dataset 1

Topic 1 Words: sea devil fruit grand user fruits blue red water bur

Topic 2 Words: luffy crew pirates straw robin franky hat government joins baroque

Topic 3 Words: pirates luffy island haki piece dressrosa alliance series straw flame

Topic 4 Words: island luffy manga pose grand crew set ace log magnetic

2.2.2 Dataset 2

Topic 1 Words: officials president soviet people noriega government roberts rating children leaders

Topic 2 Words: percent bank bush central billion california economy people administration nation

Topic 3 Words: soviet people union president gorbachev day barry government bush national

Topic 4 Words: rose fire monday school government people oil company police percent

2.3 Followup Questions

pLSA is very similar to general mixture models in the fact that both utilize probability distributions in their analysis. Both also utilize an EM algorithm to get closer to a maximal log likelihood incrementally. Thus, pLSA is almost like an extension of the general mixture model to text data.

The disadvantages of pLSA are multifold. First of all, there are many parameters to tune in order to obtain maximal results. Such extensive hypertuning is a disadvantage in itself. In my case, for a large document such as dataset 2, obtaining semi-optimal parameters took multiple hours, as a individual run could take an hour itself (when testing with a high K). In addition, pLSA doesn't have a great ability to generalize new data, and may fall prey to overfitting. This could result in topics that have high overlap and don't perform well on new, unseen datasets.