CS534 Implementation Assignment 2: Naive Bayes

Amit Bawaskar, Michael Lam
EECS, Oregon State University

May 1, 2013

Abstract

In this assignment, we implemented the Naive Bayes classifier with the Bernoulli model and Multinomial model, and compared their performance.

1 Introduction

We implemented the Naive Bayes classifier to solve a document classification problem on the 20-newsgroup data set. Two models were implemented and compared for performance: Bernoulli model and Multinomial model.

2 Naive Bayes

Paragraph about Naive Bayes.

2.1 Model

Naive Bayes assumes that the features are independent.

2.2 Inference

Inference is performed by using Bayes rule with the learned likelihood and prior probabilities, and using decision theory to select the class that maximizes the posterior probability.

2.3 Learning

Learning the likelihood probabilities for each feature and class is done by maximum a posteriori estimation. Equivalently this amounts to applying Laplace smoothing to the maximum likelihood estimator. The prior probabilities for each class is the maximum likelihood estimator for it.

2.4 Implementation Details

In this project we operated with the log of probabilities in order to avoid underflow issues. That is, for every multiplication and division operation, we instead used addition and subtraction of the log of the operands. We also stored the log of probabilities. For decision theory, we simply selected the class that maximizes the log of the posterior probability since the log function is a monotonically increasing function. Thus we operated with the log of probabilities for every circumstance.

We also applied Laplace smoothing to the likelihood probability of each feature and class in order to assign a default prior to words that have not been encountered.

3 Bernoulli Model

Paragraph about Bernoulli model.

The overall test accuracy for the Bernoulli model is **0.772152**. Figure 1 shows the confusion matrix.

4 Multinomial Model

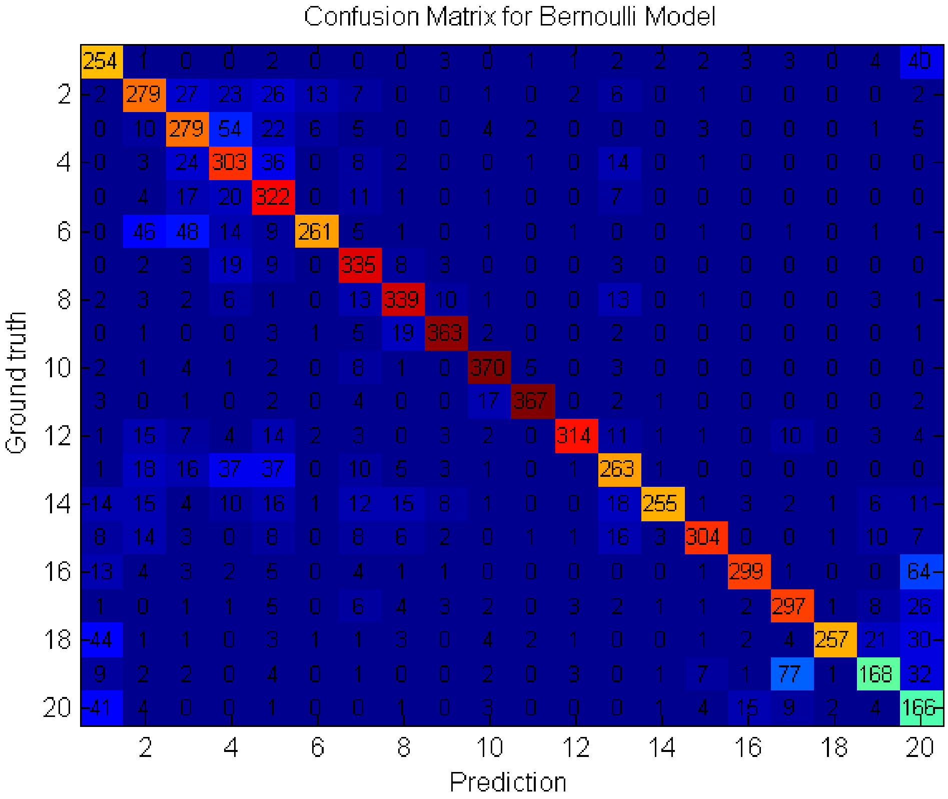
Paragraph about Multinomial model.

Report overall testing accuracy.

Report the confusion matrix.

5 Heuristics for Reducing Vocabulary Size

We explored several heuristic strategies for reducing the vocabulary size in hopes of improving classification accuracy.



		2		4		6		8		10		12		14		16		18		20
	254	1	0	0	2	0	0	0	3	0	1	1	2	2	2	3	3	0	4	40
2	2	279	27	23	26	13	7	0	0	1	0	2	6	0	1	0	0	0	0	2
	0	10	279	54	22	6	5	0	0	4	2	0	0	0	3	0	0	0	1	5
4	0	3	24	303	36	0	8	2	0	0	1	0	14	0	1	0	0	0	0	0
	0	4	17	20	322	0	11	1	0	1	0	0	7	0	0	0	0	0	0	0
6	0	46	48	14	9	261	5	1	0	1	0	1	0	0	1	0	1	0	1	1
	0	2	3	19	9	0	335	8	3	0	0	0	3	0	0	0	0	0	0	0
8	2	3	2	6	1	0	13	339	10	1	0	0	13	0	1	0	0	0	3	1
	0	1	0	0	3	1	5	19	363	2	0	0	2	0	0	0	0	0	0	1
10	2	1	4	1	2	0	8	1	0	370	5	0	3	0	0	0	0	0	0	0
	3	0	1	0	2	0	4	0	0	17	367	0	2	1	0	0	0	0	0	2
12	1	15	7	4	14	2	3	0	3	2	0	314	11	1	1	0	10	0	3	4
	1	18	16	37	37	0	10	5	3	1	0	1	263	1	0	0	0	0	0	0
14	14	15	4	10	16	1	12	15	8	1	0	0	18	255	1	3	2	1	6	11
	8	14	3	0	8	0	8	6	2	0	1	1	16	3	304	0	0	1	10	7
16	13	4	3	2	5	0	4	1	1	0	0	0	0	0	1	299	1	0	0	64
	1	0	1	1	5	0	6	4	3	2	0	3	2	1	1	2	297	1	8	26
18	44	1	1	0	3	1	1	3	0	4	2	1	0	0	1	2	4	257	21	30
	9	2	2	0	4	0	1	0	0	2	0	3	0	1	7	1	77	1	168	32
20	41	4	0	0	1	0	0	1	0	3	0	0	0	1	4	15	9	2	4	166

Figure 1: Confusion matrix for the Bernoulli model.

We tried eliminating words in the dictionary that have length less than or equal to a certain threshold value. We explored thresholds from 0 to 5, meaning keeping all words to keeping only words with length greater than 5. The results in Figure 2 indicate that the classification accuracy increased negligibly when eliminating words up to 2 letters long, but decreased noticeably as the threshold increases after 2 letters.

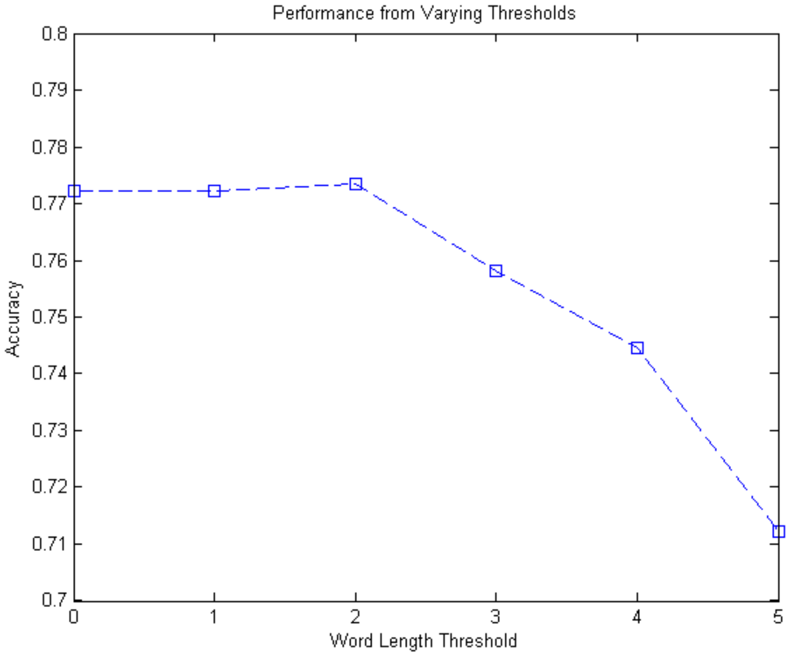


Figure 2: Classification accuracy using the Bernoulli model and varying word length thresholds. Words of length less than or equal to the threshold value are omitted in the vocabulary for learning and inference.