CS57300: Homework 2

1. **Write code to transform a given input file to a bag-of-words representation. (5 pts)**

(B) Output the top ten selected words with highest frequency (i.e., words 101-110). (Make sure you count every word once in each review (even though it appears multiple times) since we are using bag of words representation.)

Ans: The top ten words:

WORD1 burger
WORD2 definitely
WORD3 try
WORD4 much people
WORD5 did come
WORD6 delicious
WORD7 went
WORD8 off
WORD9 has amazing them
WORD10 made

2. **Implement a Naive Bayes Algorithm. (20 pts)**

(a) Write code to read in training data and learn the NBC model.

(b) Write code to read in test data to apply the learned NBC and evaluate the resulting predictions with zero-one loss.

Ans: ZERO-ONE-LOSS 0.089

3. **Learn and apply the algorithm (15 pts)**

(a) For each % in [1, 5, 10, 20, 50, 90]:

Ans:

Percentage Size of Train Set	Average Zero-One Loss	Standard Deviation of Zero-One Loss
1%	0.4643	0.0514
5%	0.3122	0.1741
10%	0.2845	0.1967
20%	0.2369	0.1910
50%	0.1051	0.0194
90%	0.0985	0.0271

(b) Plot a learning curve for the results (training set size vs. zero-one loss). Compare to the baseline *default* error that would be achieved if you just predicted the most frequent class label. Discuss the results.

Ans:

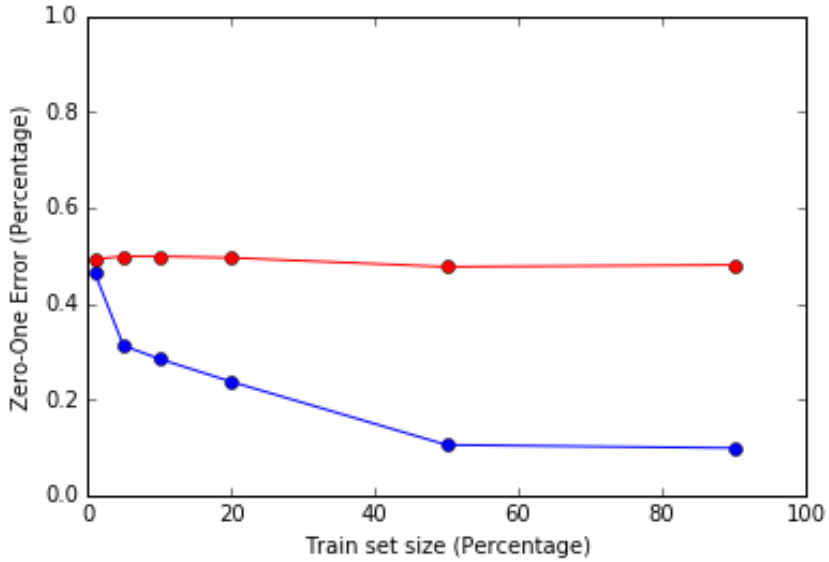


Fig. 1 Train set size against Zero-One Error

In the above figure, we can see that the value of the Baseline Error (Red) over all fraction of sample sizes is higher than that of the Naïve Bayes Classifier (Blue). From this we can conclude that NBC delivers predictions with higher accuracy than its Baseline counterpart.

Further, we see that as the training-test ratio increases, the performance of the Naïve Bayes Classifier performs immensely better than the Baseline Classifier.

4. **Explore effect of feature space (10 pts)**

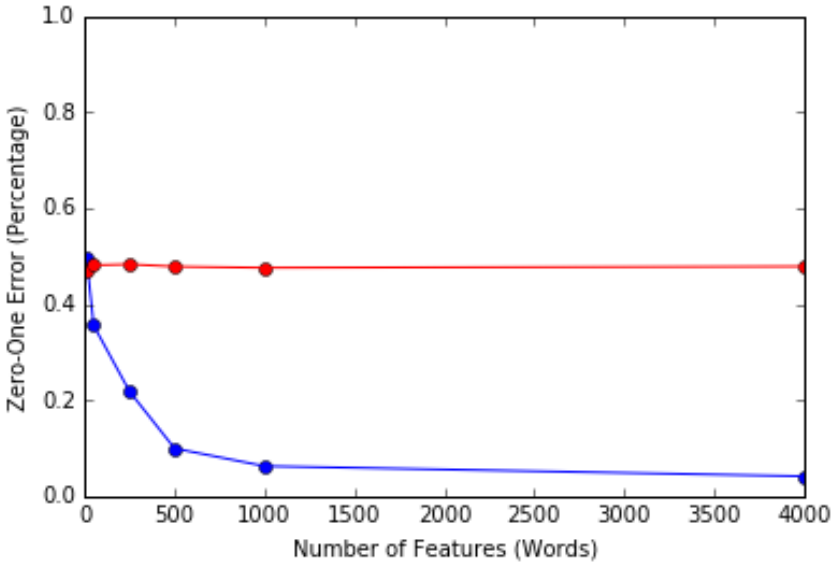
(a) For each *W* in [10, 50, 250, 500, 1000, 4000]:

Ans:

Training/Test Ratio	Number of Features	Average Zero-One Loss	Standard Deviation of Zero-One Loss
50%	10	0.4980	0.0374
50%	50	0.3554	0.0302
50%	250	0.2176	0.1118
50%	500	0.0993	0.0210
50%	1000	0.0625	0.0076
50%	4000	0.0414	0.0069

(b) Plot a learning curves for the results (feature size vs. zero-one loss). Again compare to the baseline *default* error that would be achieved if you just predicted the most frequent class label. Discuss the results.

Ans:



From the above graph, we can see that on majority, as more number of feature words are selected, the accuracy of the Naïve Bayes Classifier increases drastically till a size of 1000 after which it does not show any such drastic improvement in its accuracy.

When small number (approx. around 10) of features are selected, the accuracy of the Baseline Assumption Model is better. Apart from this, any number of features above 50 results in a very large difference between the two models, with NBC outperforming the Baseline Model enormously.