CS 573 – Homework 2

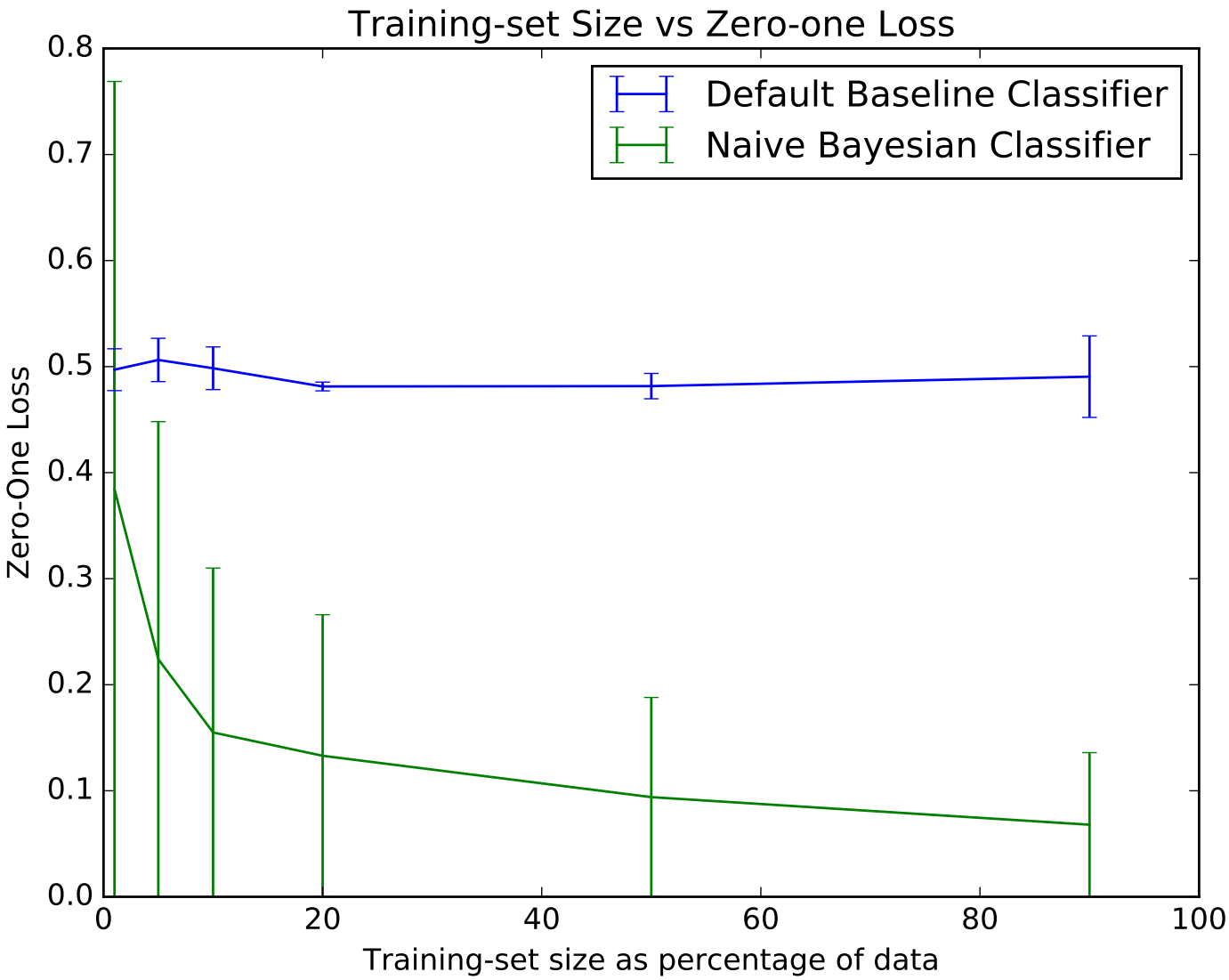
Priyank Jain
jain206@purdue.edu

February 16, 2017

Q1 and Q2
For first 1000 records of the given data used as training data and remaining 1000 records as testing data, below are the top 10 words selected words and the zero-one loss:
WORD1 definitely
WORD2 try
WORD3 people
WORD4 much
WORD5 did
WORD6 come
WORD7 delicious
WORD8 went
WORD9 off
WORD10 amazing
ZERO-ONE-LOSS 0.072

Q3

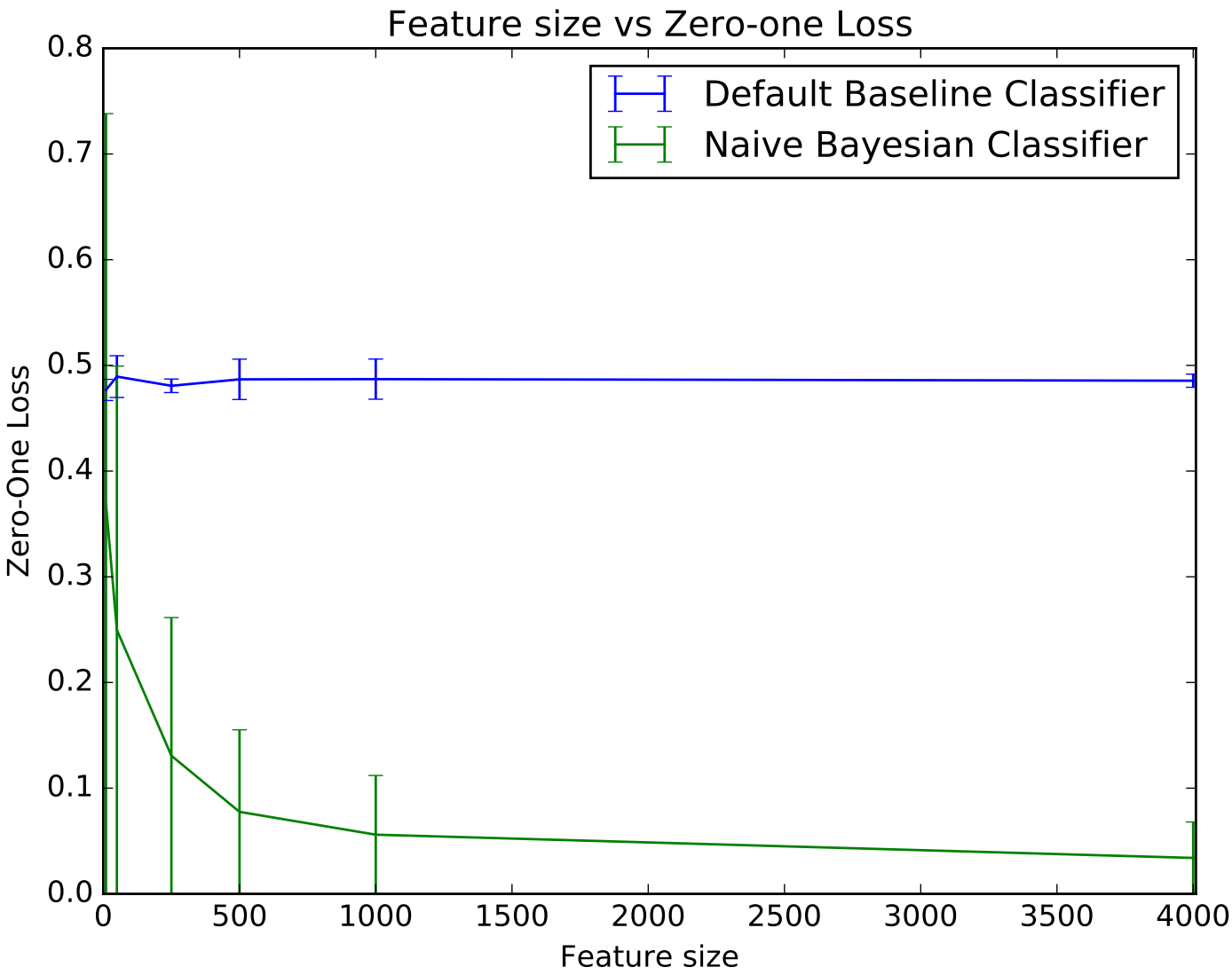
The figure below shows the training set size vs. zero-one loss error for the baseline default and Naive Bayesian classifiers.


Training-set Size vs Zero-one Loss
Default Baseline Classifier
Naive Bayesian Classifier
Zero-One Loss
Training-set size as percentage of data
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0
20
40
60
80
100

As is evident from the figure, Naive Bayesian classifier has a better performance than the baseline default classifier. As the number of training samples increases, the accuracy of Naive Bayesian classifier improves. On the other hand, the baseline default classifier has a high error rate throughout revolving around 0.5. Whereas, the Naive Bayesian classifier has high variance when the number of training samples is less. As the training set size increases, the mean of the zero-one loss for the Naive Bayesian classifier decreases and so does it's variance.
The results could be attributed to the fact that Naive Bayesian classifier learns more and more from the data as we keep increasing the number of training samples. Whereas, the baseline default classifier is relatively highly-biased and does not learn a lot as the number of training samples increases.

Q4

The figure below shows the feature size vs. zero-one loss error for the baseline default and Naive Bayesian classifiers.


Feature size vs Zero-one Loss
Default Baseline Classifier
Naive Bayesian Classifier
Zero-One Loss
Feature size
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0
500
1000
1500
2000
2500
3000
3500
4000

The figure shows that the accuracy of baseline default classifier revolves around 0.5, so its accuracy stays almost the same. This is because the baseline default classifier does not have anything to learn as the number of features increases, since it does not take into account . On the other hand, the Naive Bayesian classifier improves its accuracy as the number of features used for training increases. This is because it can learn from extra features and make a better classification decision. Moreover, for Naive Bayesian classifier, the variance of zero-one loss decreases too as the number of features used for training increases.
Moreover, as the number of features increases, the accuracy of the Naive Bayesian classifier does not drop. This tell us that the Naive Bayesian classifier is pretty robust to overfitting.