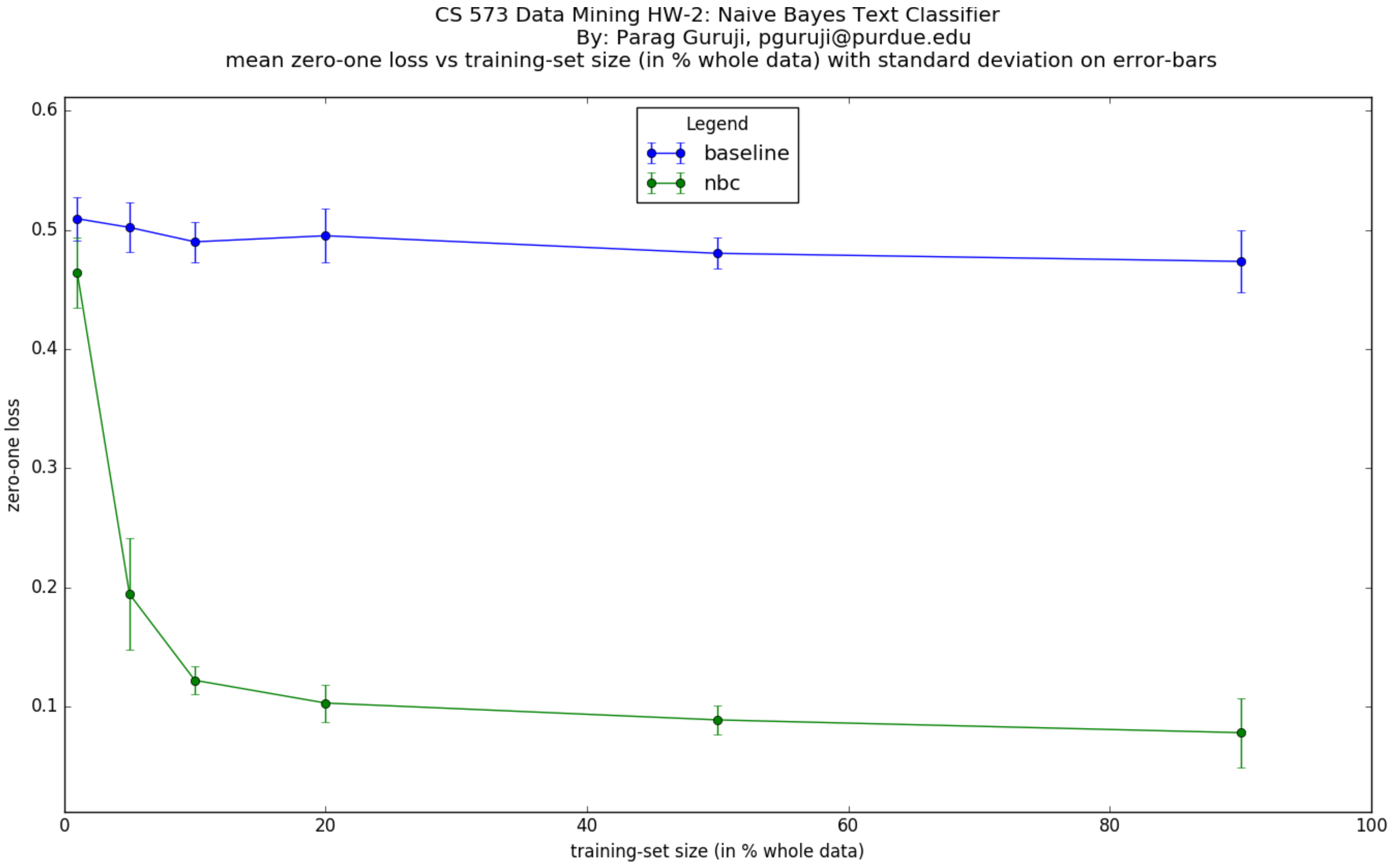
CS 573 Data Mining Homework 2

By: Parag Guruji, pguruji@purdue.edu

Date: 16 Feb 2017

3 Learn and apply the algorithm:



The above graph shows comparison of the NBC classifier's performance with the baseline performance in terms of **mean zero-one loss across 10 trials** against each of the **training set sizes** in [1, 5, 10, 20, 50, 90] (measured **as percent of whole dataset**) with **error-bars showing the standard deviation** of zero one loss. The corresponding test set size is (whole – training). The number of features is 500 in all trials.

- The mean baseline performance remains approximately around 0.5 irrespective of training size as the train-sets are randomly sampled, thus having nearly equal proportion of records for both classes.

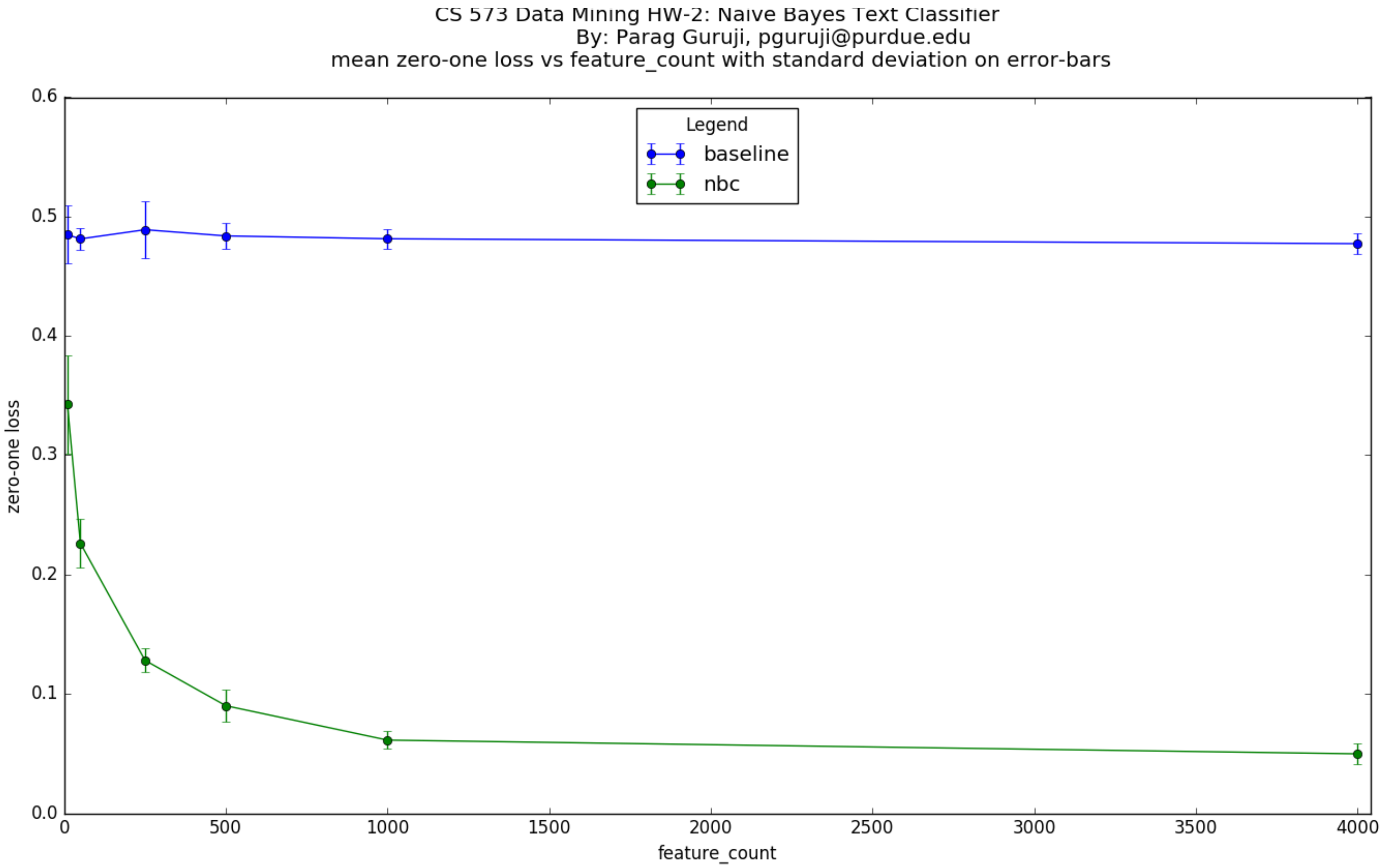
- The NBC performance improves exponentially with the increase in training size (and corresponding decrease in the testing size). Initially, with little number of learning examples and large number of testing examples, the NBC has little to learn from and much larger number of feature combinations to test. With increasing training examples for each class (i.e. feature vectors), the likelihood estimate for possible testing data associated to that class improves, causing the gain in performance.

- However, the zero-one loss converges to a saturation point and doesn't equal to 0 even when same data is used for training and testing because following: (Not all of them are shown in graph but experiments were performed.)

  (a) We are using Laplace Smoothing with alpha = 1 which effectively means ensuring at least one example of each possible combination of each feature-value with each possible class-label occurs in the training set.
  (b) If the same feature combination occurs with different class labels approximately equal no. of times with each class label, then it is effectively inducing ambiguity in learning of the model about that feature combination.
  (c) Even if all feature combinations are unique and alpha is 0 (no smoothing), the products of CPDs of different feature combinations can still be approximately equal. If such combinations are associated different class labels in nearly equal proportion, ambiguity for such feature combinations is induced.

- The standard deviation of zero-one loss shows overall trend of decrease up to 50% training size (and 50% test size). But it is increased at training size=90% significantly – Overall, with larger difference in the proportions of training size and testing size, the variability of the zero-one loss increases. In both models – Baseline & NBC, very low standard deviation is observed at 50% training size.

4 Explore effect of feature space:



The above graph shows comparison of the NBC classifier's performance with the baseline performance in terms of **mean zero-one loss across 10 trials** against each of the different **number of features** in [10, 50, 250, 500, 1000, 4000] with **error-bars showing the standard deviation** of zero one loss. The corresponding train (and test) set size is 50%.

- The mean baseline performance remains approximately around little less than 0.5 irrespective of feature count as it's computation does not consider any feature at all. Due to random sampling, the proportion of positive to negative examples also remains nearly same in all trials with very little variance (std. dev.) (with 2 very small exceptions at feature counts 10 and 250)

- The NBC performance improves exponentially with the increase in number of predictors, i.e. features. The words are added to the feature-set with descending order of their frequency (which we assume to be the indicator of their "importance"). Hence, more number of features add more CPD terms to the multiplicative numerator of model, thus encoding exponentially increasing number of possible data combinations in the training and in turn, increasing the explanatory capability of the model.

- However, the zero-one loss converges to a saturation point and doesn't equal to 0 even when same data is used for training and testing as explained in explanation of previous question.

- Additionally, with increasing number of features, we are in fact including lesser frequent (thus less important/significant) features to the model which produce increasingly sparse distribution of positive values across the training data. Hence, not contributing significantly to the θ value of corresponding CPD.

- The standard deviation of zero-one loss in most of the cases is very low and shows even further decrease as the feature count increases. This is indicative of higher precision of the model at a given feature count and also of the marginal gains in this precision with increasing feature count.