CS573 Data Mining
Homework 2

Fangda Li
li1208@purdue.edu

February 15, 2017

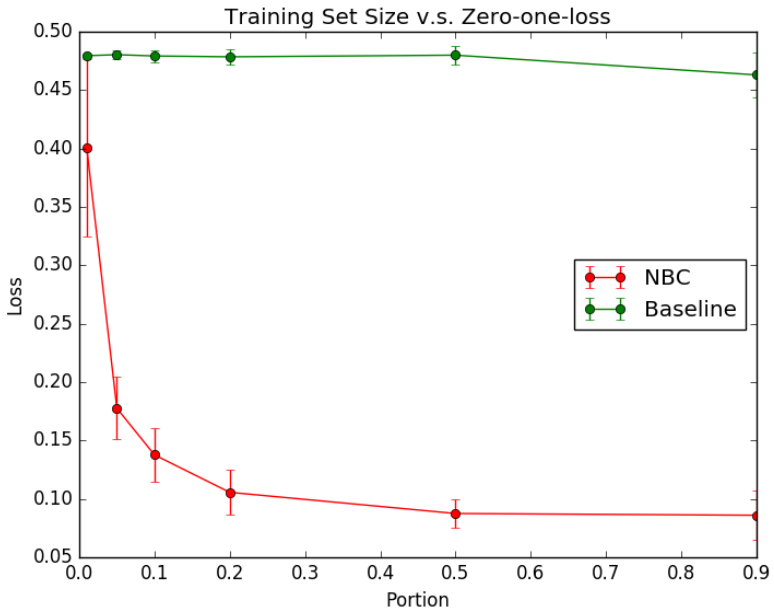

Figure 1: Training set size v.s. zero-one loss. The data points and error bars represent the mean and std across 10 runs, respectively. At a lower portion, the rates of decrement of both the variance and mean decreases significantly as the portion increases. Whereas having a portion greater than 0.5, the NBC performance tends to converge and becomes independent of the randomness.

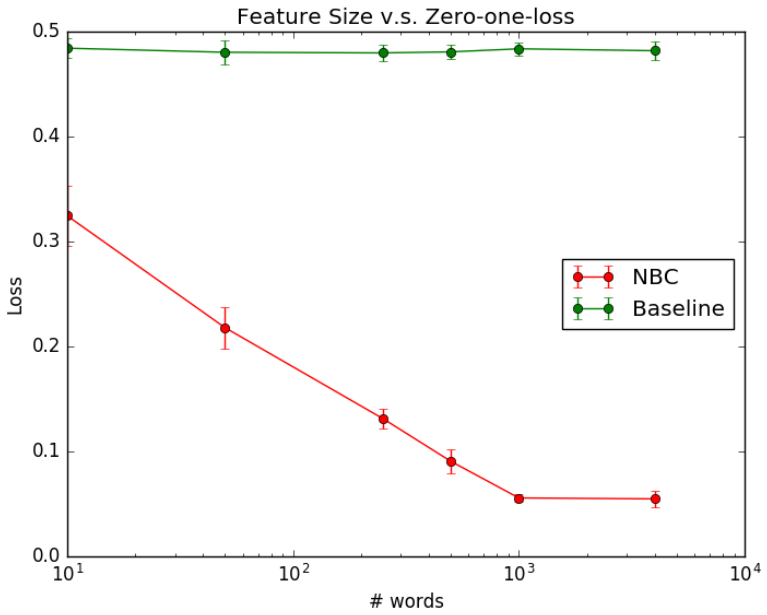

Figure 2: Training set size v.s. zero-one loss. Similar to Figure 1, the performance converges after including 1000 words in the feature vector. Combining with the observations in portions, the NBC model with a portion of 0.5 and a feature vector size of 1000 might provide the best balance between complexity and performance. Note that representing the samples using a mere 10-word feature vector can still provide 15% improvement that the baseline performance. Future work can be in the direction of exploiting the proximity of words.