CS573 Data Mining
Homework 3

Fangda Li
li1208@purdue.edu

March 18, 2017

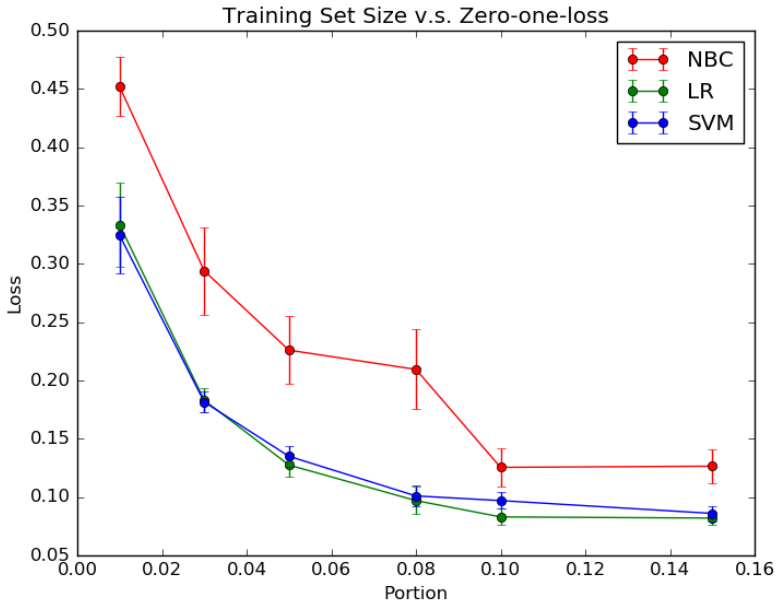

Figure 1: **Analysis 1**. Training set size v.s. zero-one loss. The data points and error bars represent the means and standard errors across 10 folds, respectively. Now, the following hypothesis regarding the performance difference is formed: *Zero-one-losses of LR and SVM across all training set sizes are significantly different*. In order for the hypothesis to be accepted, we use the criteria such that the average of t-test p-values across the training set sizes must be greater than 0.100. Next, *t-test* on the losses of the ten trials for each training set size is performed and the following p-values are obtained: $0.862, 0.920, 0.603, 0.798, 0.194, 0.675$, with an average of 0.676. According to the previously defined passing criteria, we cannot reject the null hypothesis since $0.676 > 0.100$. As a result, the following conclusion is reached: Zero-one-losses of LR and SVM across all training set sizes are *not* significantly different (trinary is worse).

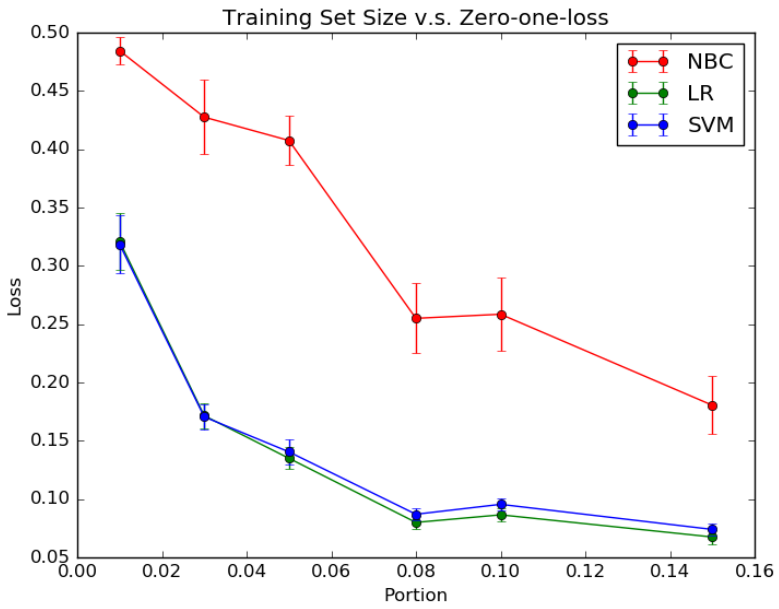

Figure 2: **Analysis 2**. Training set size v.s. zero-one loss with trinary feature values. The data points and error bars represent the means and standard errors across 10 folds, respectively.

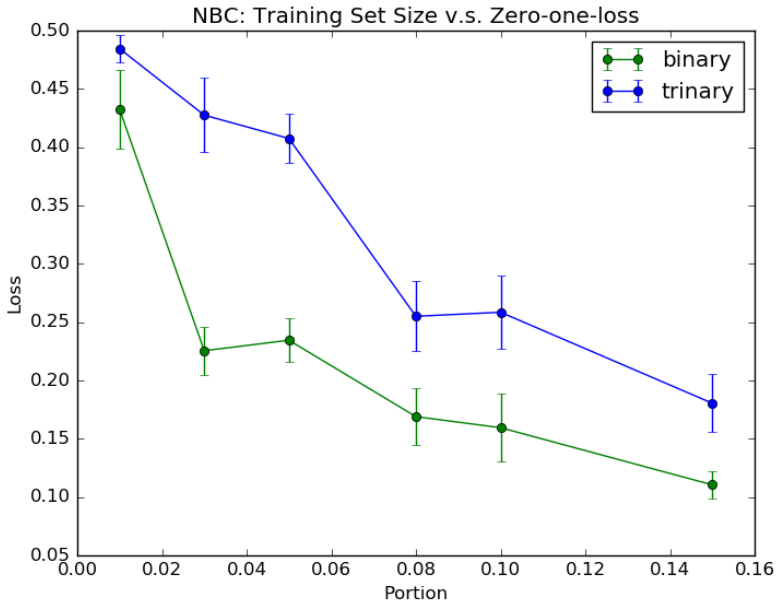

Figure 3: Loss comparison between binary and trinary feature value for NBC. Now, the following hypothesis regarding the performance difference is formed: *Zero-one-losses of NBC with binary and trinary feature value across all training set sizes are significantly different.* In order for the hypothesis to be accepted, we use the criteria such that the average of t-test p-values across the training set sizes must be greater than 0.100. Next, *t-test* on the losses of the ten trials for each training set size is performed and the following p-values are obtained: $1.89e-01, 7.97e-05, 1.67e-05, 5.02e-02, 4.08e-02, 2.74e-02$, with an average of 0.051. According to the previously defined passing criteria, we can reject the null hypothesis since $0.051 < 0.100$. As a result, the following conclusion is reached: Zero-one-losses of NBC with binary and trinary feature value across all training set sizes are significantly different.

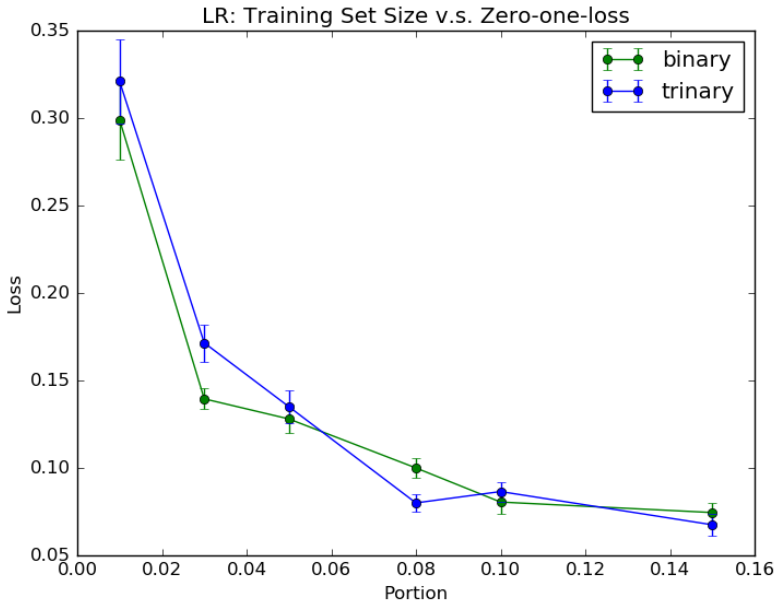

Figure 4: Loss comparison between binary and trinary feature value for LR. Now, the following hypothesis regarding the performance difference is formed: *Zero-one-losses of LR with binary and trinary feature value across all training set sizes are significantly different.* In order for the hypothesis to be accepted, we use the criteria such that the average of t-test p-values across the training set sizes must be greater than 0.100. Next, *t-test* on the losses of the ten trials for each training set size is performed and the following p-values are obtained: $0.535, 0.023, 0.600, 0.024, 0.522, 0.435$, with an average of 0.356. According to the previously defined passing criteria, we cannot reject the null hypothesis since $0.356 > 0.100$. As a result, the following conclusion is reached: Zero-one-losses of LR with binary and trinary feature value across all training set sizes are *not* significantly different.