Meng Deng
Dr. Jennifer Neville
CS 573, hw 3
March 16, 2017

LR and SVM Analysis

1. Assess model performance when applying cross-validation.

   a) Learning Curves of Logistic Regression, SVM, and NB models. The graph shows that in general, as the training sample size increases, all of the three models increase their performance (loss drops). However, both SVM and LR seems to have higher performance than NBC, SVM and LR has similar performance across all size. The standard deviation of NBC seems to be higher than SVM and LR.

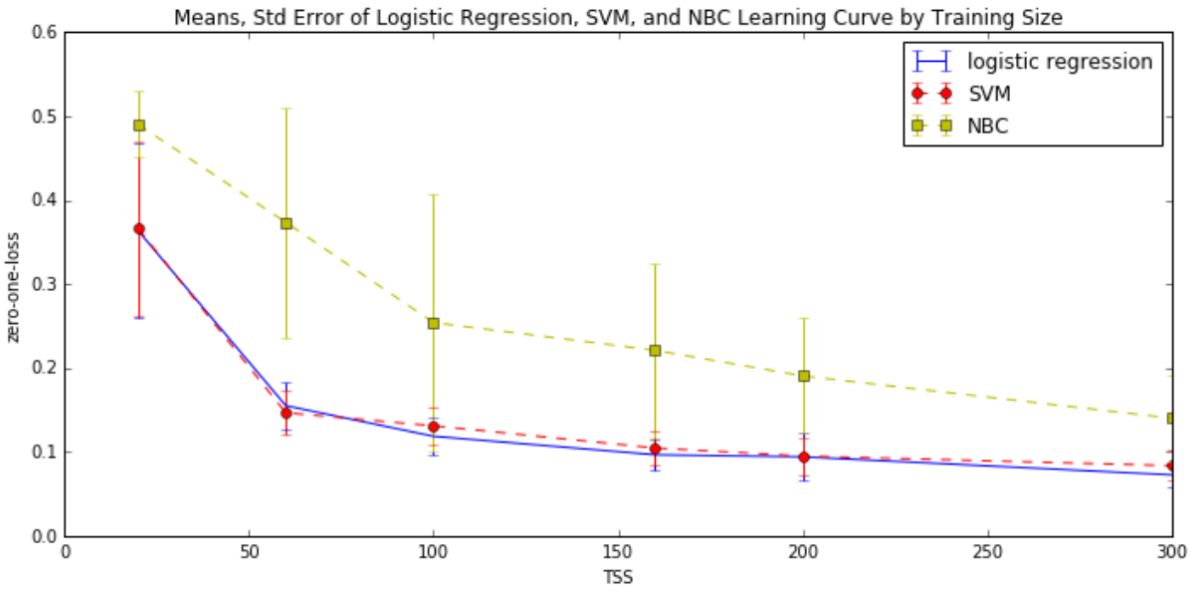


Figure 1: Means, Std Error of NBC, Logistic Regression, and NBC by Training Set Size.

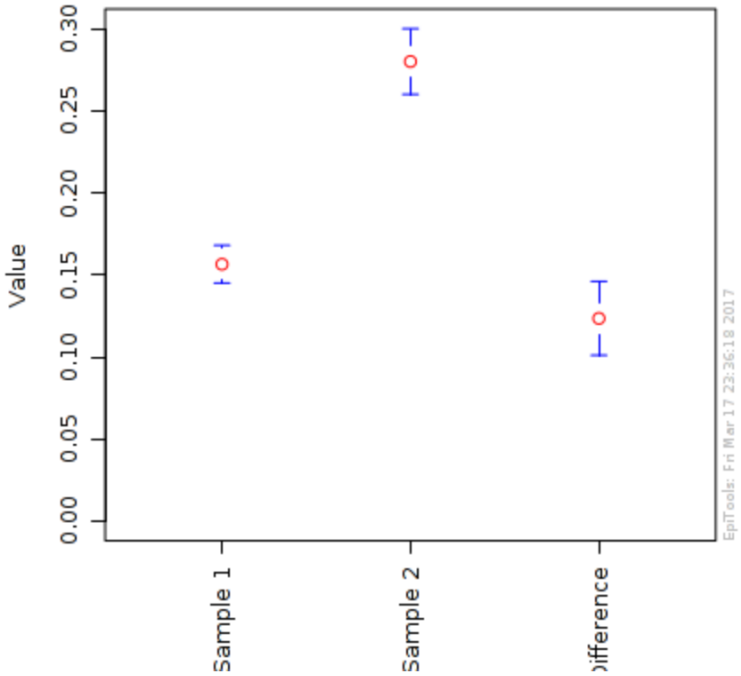
   b) H-null: NBC performs no different than SVM.

   H-alternative: SVM performs significantly better than NBC.

c) I performed a one-tail t-test. The result fail to accept null hypothesis. I conclude that SVM performs significantly better than NBC.

	SVM	NBC	Difference
Mean	0.1565	0.28	0.1235
Std. Dev.	0.0962	0.1686	0.137
df	199	199	398
t-stat			8.821
p-value			0

Figure 2: 90% Confident Interval (Sample 1= SVM, Sample 2 = NBC)



2. Assessing model performance when modifying the word feature to include three values: 0: word does not occur; 1: word occurs once; 2: word occurs more than once. Rerun the cross-validation procedure and compare the three model performance.

a) Learning Curves of Logistic Regression, SVM, and NB models, modified word feature. I found the similar pattern as in figure 1. The graph shows that in general, as the training sample size increases, all of the three models increase their performance (loss drops). However, both SVM and LR seems to have higher performance than NBC, SVM and LR has similar performance across all size. The standard deviation of NBC seems to be higher than SVM and LR.

Different from figure 1, I found the performance of the three models varied especially when the TSS is small. To be specific, the performance of NBC seems to increase slower when the word feature incorporated the third value. Overall, it performs worse than the previous status. SVM and LR, however, seems to slightly improved across all TSS.

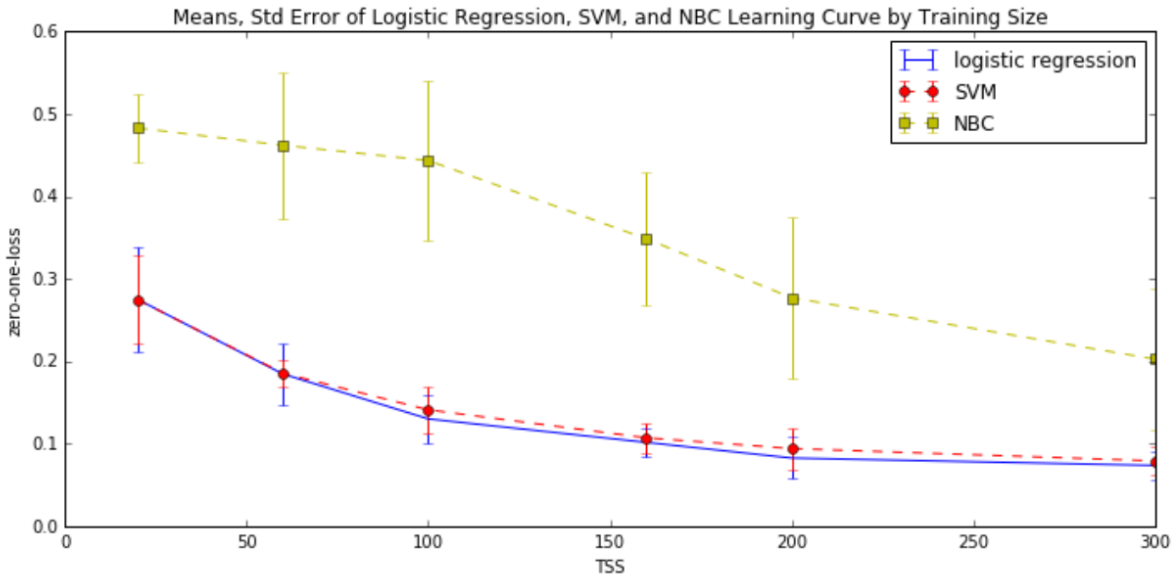


Figure 3: Means, Std Error of NBC, Logistic Regression, and NBC by Training Set Size.

b) H-null: NBC performs no different than SVM.
H-alternative: SVM performs significantly better than NBC.

c) I performed a one-tail t-test. The result fail to accept null hypothesis. I conclude that SVM performs significantly better than NBC.

	SVM	NBC	Difference
Mean	0.1534	0.3195	0.1661
Std. Dev.	0.0886	0.1624	0.131
df	199	199	398
t-stat			12.777
p-value			0

Figure 4: 90% Confident Interval (Sample 1= SVM, Sample 2 = NBC)