CS57300: Homework 3

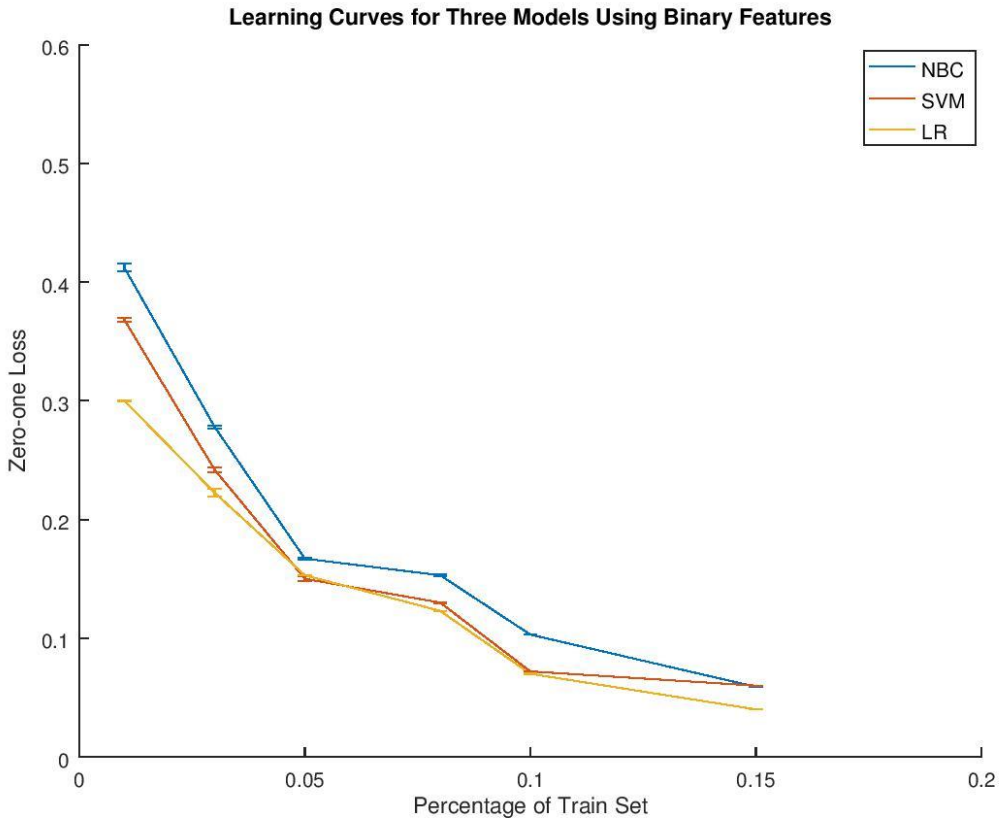
Hengrui Zhang

March 15, 2017

Analysis

1. Assess whether choice of model improves performance.

(a)



We can notice that the LR algorithm performs better than SVM. The reason is that we are using non-linearity features here. The advantage of SVM over LR is the non-linearity obtained by the use of non-linear kernels.

(b)

Hypothesis:

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_a : \mu_1 - \mu_2 \neq 0$$

in which μ_1 is the mean calculated by Logistic Regression, μ_2 is the mean calculated by SVM

(c) Here we can use a two-sample T Test to test our result in question part b.

The test statistic t:

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\frac{{s_1}^2}{{n_1}^2} + \frac{{s_2}^2}{{n_2}^2}}}$$

in which S_1 is the standard deviation of the first distribution, S_2 is the standard deviation of the second distribution. n1 = n2 =10. The degree of freedom is 2n-2,which is 18 in here. Given the t value and the degree of freedom, we can look up the t-table. We have 6 different train set and we did 10 times tests with the condition of these different set. So we can only draw the conclusion given specific condition. We can also use the major vote to decide our final decision. For example, in these 6 situation, 4 of them we concluded Ha, and the left 2 we conclude h0, and then we can conclude that μ_1 does not equal to μ_2.

In the test set of TSS1, which we have 20 train data.

$$t = \frac{0.3 - 0.368}{\sqrt{\frac{0.0036998}{10^2} + \frac{0.01010227}{10^2}}} = -5.788$$

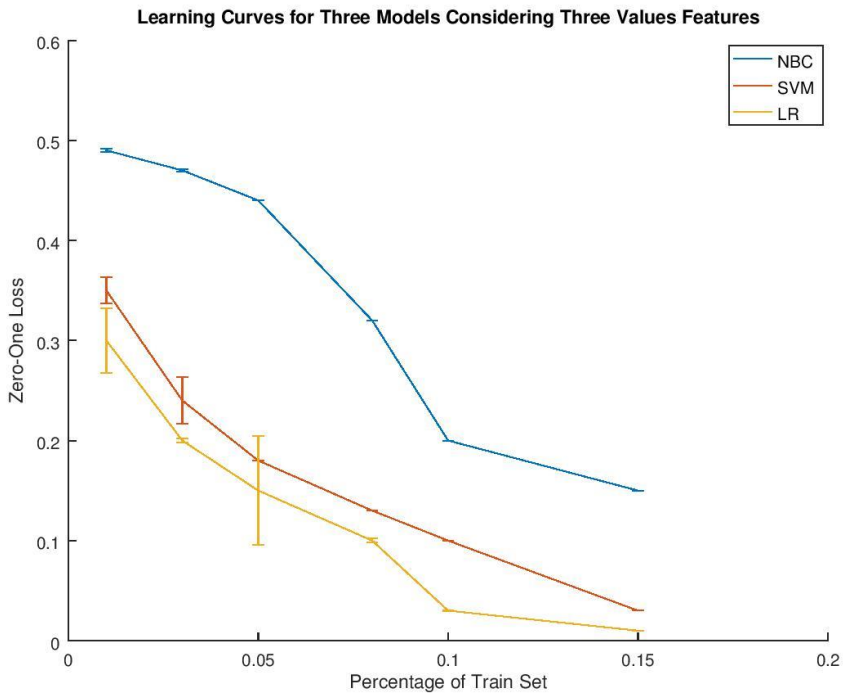
In here, we choose α=0.005.

$t_{.995}(18) = 2.878$. Since $|t| > t_{.995}(18)$, we reject the H_0 hypothesis and conclude H_a hypothesis, which means by using the observed data, there is a significant difference between the mean calculated by Logistic Regression and the mean calculated by SVM.

By using this procedure in different TSS(TSS2,TSS3,TSS4,TSS5,TSS6), the absolute values of t are all larger than 2.878. So we can conclude that no matter which train data we are using, there is a significant difference between SVM and LR

2. Assess whether feature construction affects performance

(a)



(b) Hypothesis:

$$H_0 : \mu_1 - \mu_1' = 0$$

$$H_a : \mu_1 - \mu_1' \neq 0$$

in which μ_1 is the mean calculated by Logistic Regression with binary features, μ_1' is the mean calculated by Logistic Regression by three features.

(c) Here we can also use a two-sample T Test to test our result in question part b.

Using the same method in question 1c, we calculate the t value and compare them to the t table. For each TSS1,TSS2,TSS3,TSS4,TSS5,TSS6, all of the absolute t values are larger than the given value. So we can conclude that there is a significant difference between using binary features and three value features when applying logistic regression.