Homework 3

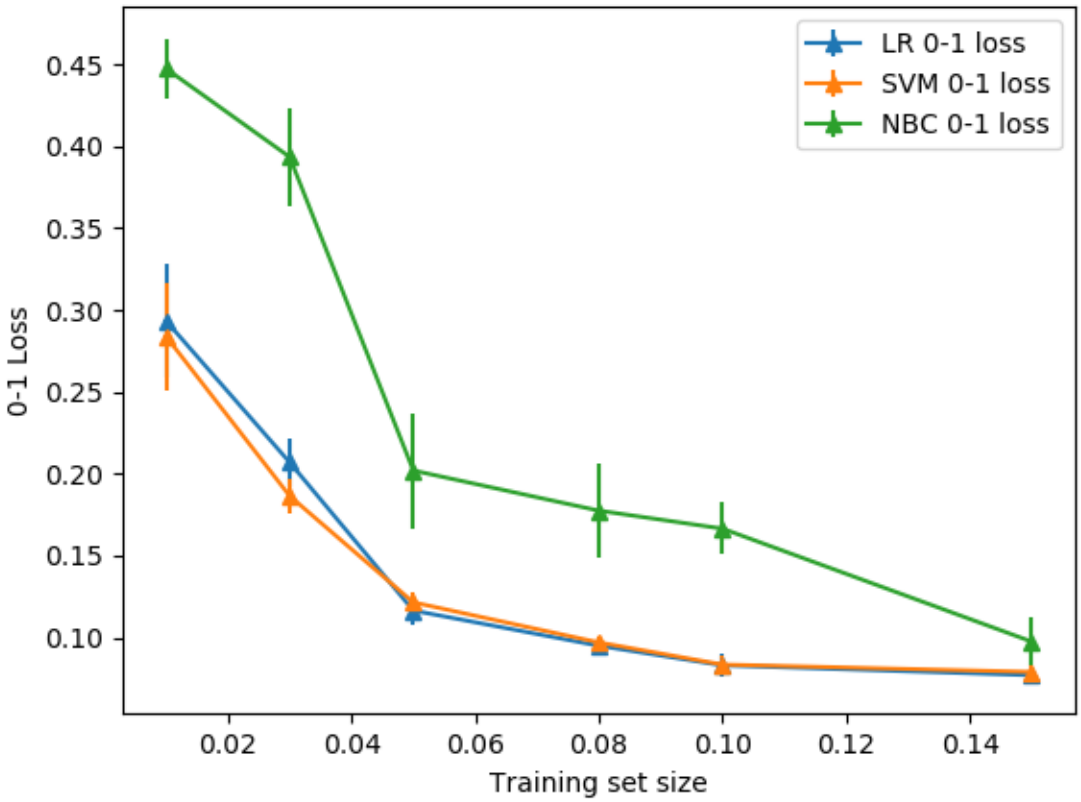
CS 57300

Submitted by:-

Name : Rashmi Soni

PUID: 0029136200

Solution 1a)



1b)

Now, let's formulate a hypothesis about the difference of performance between LR and SVM models.

Stating the hypothesis:

Null Hypothesis : There is no difference in the performance between the LR and SVM models.

Alternative Hypothesis : There is difference between the performance of LR and SVM models.

1c)

Let's first set the significance level (α) for the experiment. If we take 95% confidence interval, then:

$\alpha = 1 - 0.95 = 0.05$

Now, since we have calculated the means of zero-one loss for LR and SVM over six different training size samples, which can be given as:

LRavgZeroOneLossList : [0.29349999999999998, 0.20699999999999999, 0.11650000000000001, 0.095000000000000001, 0.083000000000000004, 0.076999999999999985]

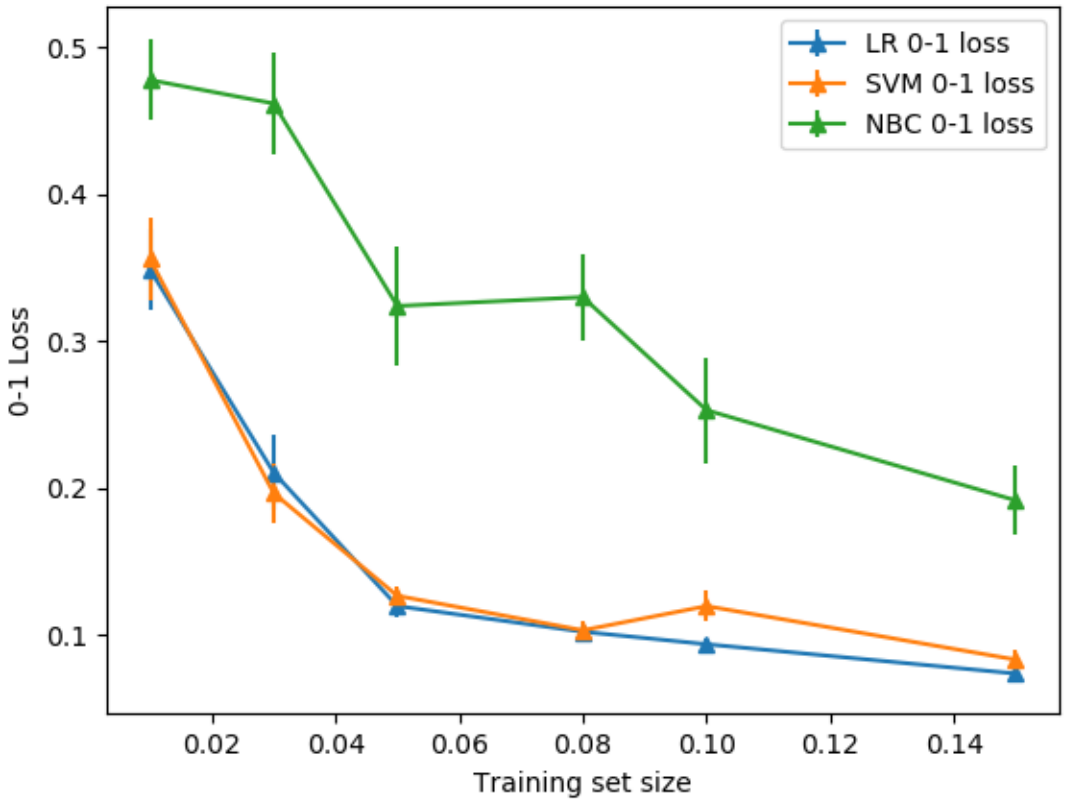
SVMavgZeroOneLossList : [0.28399999999999997, 0.1865, 0.12150000000000001, 0.097000000000000003, 0.083500000000000019, 0.079000000000000001]

Now, to calculate the p-value, we perform the paired t-test by using the mean values of LR and SVM mean zero-one losses as stated above.

So, the two-tailed P value comes as 0.4292. Since p value is greater than the significance level, so we "do not reject the null hypothesis" and the difference between the performance of these two models is not "statistically significant" (as we know that a value of a statistic is significant if it is unlikely to occur under the null hypothesis). Thus, we can say that the difference between the two means of the models occurred in the course of random sampling.

Hence, we can say that the choice of model (LR or SVM) does not significantly improve the performance.

Solution 2a)



2b)

Now, let's formulate a hypothesis about the difference of performance between LR-0-1 model (which constructed binary word features) and LR-0-1-2 model (which constructed the features using three values).

Stating the hypothesis:

Null Hypothesis : There is no difference in the performance between the LR-0-1 and LR-0-1-2 models.

Alternative Hypothesis : There is difference between the performance of LR-0-1 and LR-0-1-2 models.

2c)

Let’s first set the significance level (α) for the experiment. If we take 95% confidence interval, then:

$\alpha = 1 - 0.95 = 0.05$

Now, since we have calculated the means of zero-one loss for LR-0-1 and LR-0-1-2 models over six different training size samples, which can be given as:

LR-0-1-avgZeroOneLossList : [0.29349999999999998, 0.20699999999999999, 0.11650000000000001, 0.095000000000000001, 0.083000000000000004, 0.076999999999999985]

LR-0-1-2-avgZeroOneLossList : [0.34899999999999998, 0.21049999999999999, 0.11950000000000001, 0.10200000000000001, 0.0935, 0.073499999999999996]

Now, to calculate the p-value, we perform the paired t-test by using the mean values of LR and SVM zero-one losses as stated above.

So, the two-tailed P value comes as 0.2085. Since p value is greater than the significance level, so we “do not reject the null hypothesis” and the difference between the performance of these two models is not “statistically significant” (as we know that a value of a statistic is significant if it is unlikely to occur under the null hypothesis). Thus, we can say that the difference between the two means of the models occurred in the course of random sampling.

Hence, we can say that after constructing the features differently also, the respective model (LR) performance is not statistically significant.