CS 573: Homework 3

Piyush Dugar (pdugar@purdue.edu)

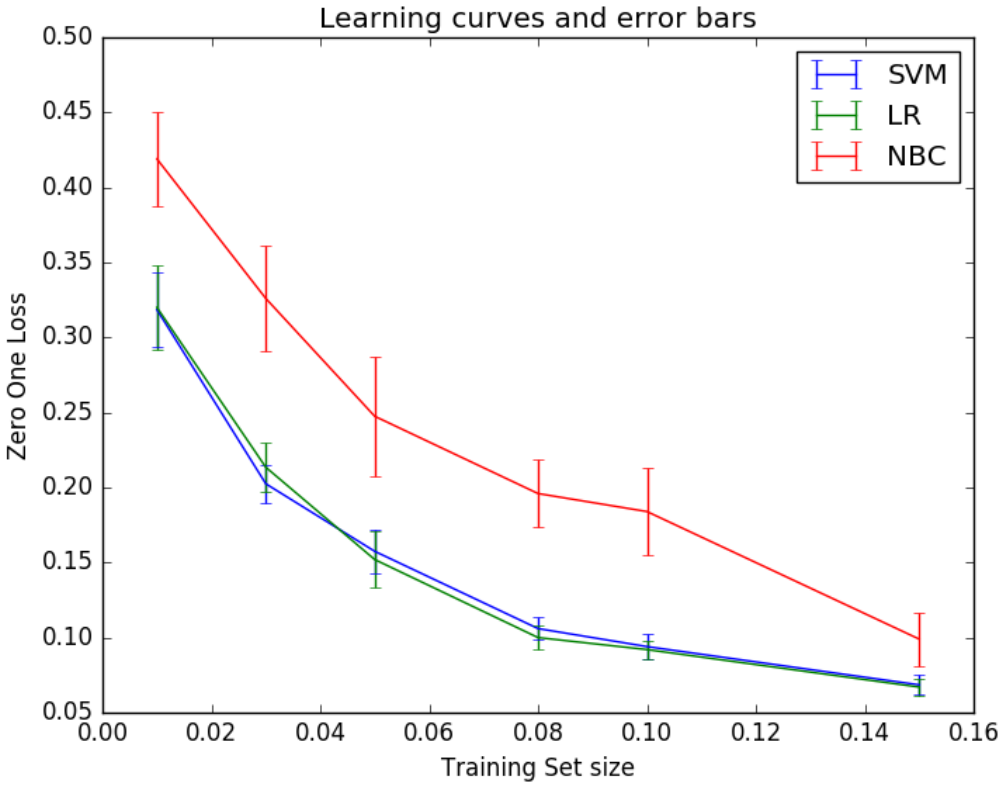
Code.

Given the input training file , testing file and the model index , the code applies the corresponding algorithm and outputs the zero-one loss

Printing the zero one loss corresponding to the Logistics Regression algorithm

Printing the zero one loss corresponding to the Support Vector algorithm

Analysis: Assess whether choice of model improves performance



Learning curves and error bars for the three models from the evaluation based on incremental CV.Here features have value of 0 and 1

Hypothesis

Hypothesis for Logistics regression and Support Vector Machine
H_0:difference of mean of zero-one loss between LR and SVM = 0
H_a: difference of mean of zero-one loss between LR and SVM $\neq$ 0

Hypothesis for Logistics regression and Naive Bayes
H_0: difference of mean of zero-one loss between LR and NB = 0
H_a: difference of mean of zero-one loss between LR and NB $\neq$ 0

Hypothesis for Support Vector Machine and Naive Bayes
H_0:difference of mean of zero-one loss between SVM and NB = 0
H_a:difference of mean of zero-one loss between SVM and NB $\neq$ 0

Before observing the results,We think the following would be true:

1. H_0 is true for LR and SVM
2. H_a is true for LR and NB

3. H_a is true for SVM and NB

NB assumes that the features are uncorrelated i.e are independent and set them by how much they correlate with the label. So if we have two labels are highly correlated with the label but have extra correlation among themselves . NB will give them both strong weights , so their influence is counted double.
Therefore, NB fails to account for correlation among the features and should perform less accurately as compared to the LR and SVM.
On the contrary, LR and SVM and find this correlation and compensate with the weights. Hence they should perform better than the NB classifier.
Generally, LR and SVM perform comparable . Both can find correlations among the inputs and with the class and adjusts the weights accordingly.So, error on the test set should be same for both of them.

Discuss whether the observed data support the hypothesis

To analyze the results, We performed the T-sets on all three models for each train data size.
Following are the results

T statistics and P value for LR and SVM

TSS	T statistics	P value
0.01	0.087	0.931
0.03	0.267	0.793
0.05	0.657	0.52
0.08	0.078	0.938
0.1	1.568	0.134
0.15	0.839	0.412

As we can see that p value for all the TSS are much larger than 0.05, this confirms that our NULL hypothesis for LR and SVM is true

T statistics and P value for LR and NBC

TSS	T statistics	P value
0.01	-2.6379	0.0167
0.03	-2.7783	0.0124
0.05	-2.1797	0.0428
0.08	-2.8343	0.011
0.1	-5.6019	0.0
0.15	-3.2145	0.0048

As we can see that p value for all the TSS are much smaller than 0.05, this confirms that our ALTERNATE hypothsis for LR and NB is true

T statistics and P value for SVM and NBC

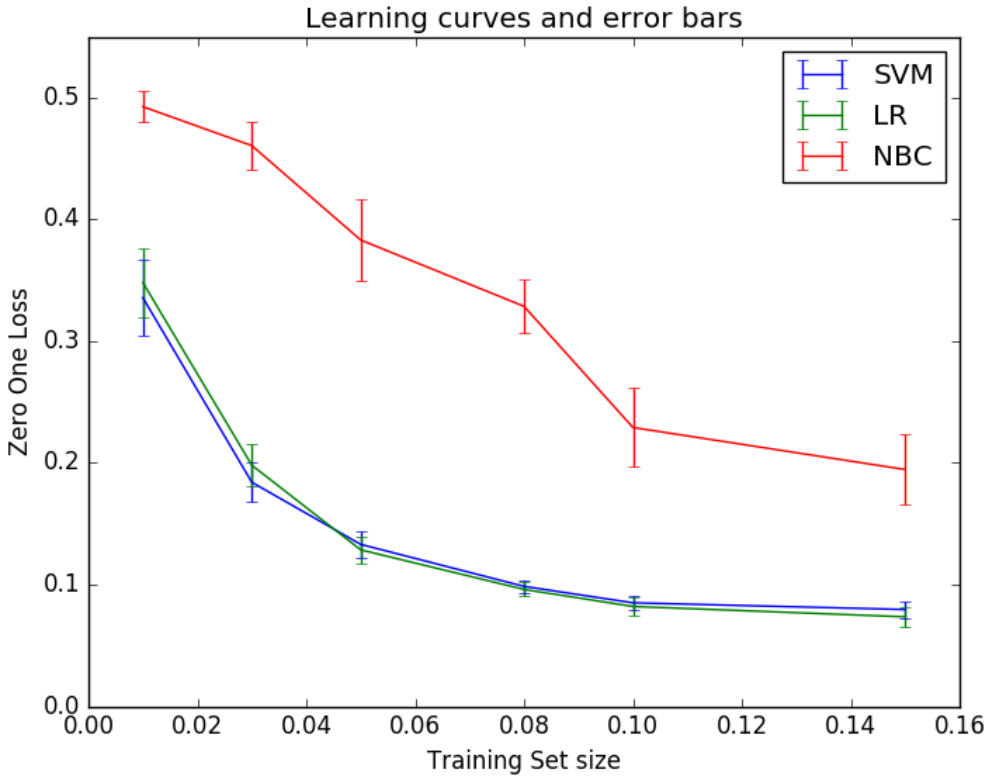
TSS	T statistics	P value
0.01	-2.4129	0.0267
0.03	-2.6834	0.0152
0.05	-2.0327	0.0571
0.08	-2.7579	0.013
0.1	-4.8975	0.0001
0.15	-2.8915	0.0097

As we can see that p value for all the TSS are much smaller than 0.05, this confirms that our ALTERNATE hypothesis for SVM and NB is true

So we have the following conclusions:

1. LR and SVM have have similar performance as their NULL hypothesis is true
2. LR and NBC have different performances . Looking at the graph, the errors of LR is much less than the NBC and so LR performs better.
3. SVM and NBC have different performances. Looking at the graph, the errors of SVM is much less than the NBC and so SVM performs better.

Analysis : Assess whether feature construction affects performance



Effect of feature construction : Learning curves and error bars for the three models from the evaluation based on incremental CV.Here features have value of 0,1 and 2

Hypothesis

Hypothesis for Logistics regression new and Logistics regression old
H_0: Difference of mean of zero-one loss error between old LR and new LR $= 0$
H_a: Difference of mean of zero-one loss error between old LR and new LR $\neq 0$

Hypothesis for support vector new and support vector old
H_0: Difference of mean of zero-one loss error between old SVM and new SVM $= 0$
H_a: Difference of mean of zero-one loss error between old SVM and new SVM $\neq 0$

Hypothesis for Naive Bayes new and Naive Bayes old
H_0: Difference of mean of zero-one loss error between old NBC and new NBC $= 0$
H_a: Difference of mean of zero-one loss error between old NBC and new NBC $\neq 0$

Discuss whether the observed data support the hypothesis

To analyze the results, We performed the T-sets on all three models before and after feature constructions and for each train data size.

Following are the results

T statistics and P values for old and new LR

TSS	T statistics	P value
0.01	0.8576	0.4024
0.03	0.0739	0.9419
0.05	1.9669	0.0648
0.08	-0.2374	0.815
0.1	1.2988	0.2104
0.15	0.1654	0.8705

We observe the following things:

1. The p value is always greater than 0.05(rather much higher) and so we can confidently say that feature construction does not have any effect on the error for LR
2. We accept the NULL hypothesis

T statistics and P values for old and new SVM

TSS	T statistics	P value
0.01	0.5227	0.6076
0.03	0.1404	0.8899
0.05	2.0571	0.0545
0.08	-0.4361	0.6679
0.1	2.1371	0.0466
0.15	0.0	1.0

We observe the following things:

1. The p value is always greater than 0.05(rather much higher) and so we can confidently say that feature construction does not have any effect on the error for SVM
2. We accept the NULL hypothesis

T statistics and P values for old and new NBC

TSS	T statistics	P value
0.01	-0.7747	0.4486
0.03	-1.5192	0.1461
0.05	-1.3505	0.1936
0.08	-2.7494	0.0132
0.1	-0.4001	0.6938
0.15	-2.3093	0.033

We observe the following things:

1. **T statistics is always negative which states that the new nbc error is more than the old nbc error**

2. Also, 2 times the p value is more than 0.05 and 4 times its less than 0.05.So we reject null 2 times and accept it 4 times.

So, Following are the conclusions of increasing the number of values a feature can take:

1. There is no effect on LR and SVM and their error remain approximately same

2. On NBC, although we cannot reject the NULL hypothesis and so cannot confidently say that there is **Significant** difference between the old and new NBC, but we can say from the T statistics value and from the graph that the error of NBC increases with feature construction. The reason being , with an extra value of 2 in NBC, the number of features in NBC will increase drastically and NBC error increases with increasing the number of features