

CS 573 – Homework 3

Priyank Jain - 1 slip day used
jain206@purdue.edu

March 15, 2017

Problem 1

Code

The screenshot below shows the output of LR/SVM models for test cases of Homework 2

```
priyank@priyank:~/Desktop/Dropbox/Spring 2017/DM/My HW Solutions/HW3$ python hw3.py yelp_train0.txt yelp_test0.txt 1
ZERO-ONE-LOSS-LR 0.07
priyank@priyank:~/Desktop/Dropbox/Spring 2017/DM/My HW Solutions/HW3$ python hw3.py yelp_train0.txt yelp_test0.txt 2
ZERO-ONE-LOSS-SVM 0.076666666666666666
priyank@priyank:~/Desktop/Dropbox/Spring 2017/DM/My HW Solutions/HW3$ python hw3.py yelp_train1.txt yelp_test1.txt 1
ZERO-ONE-LOSS-LR 0.07
priyank@priyank:~/Desktop/Dropbox/Spring 2017/DM/My HW Solutions/HW3$ python hw3.py yelp_train1.txt yelp_test1.txt 2
ZERO-ONE-LOSS-SVM 0.11
priyank@priyank:~/Desktop/Dropbox/Spring 2017/DM/My HW Solutions/HW3$
```

Figure 1: Output for testcases of Homework 2

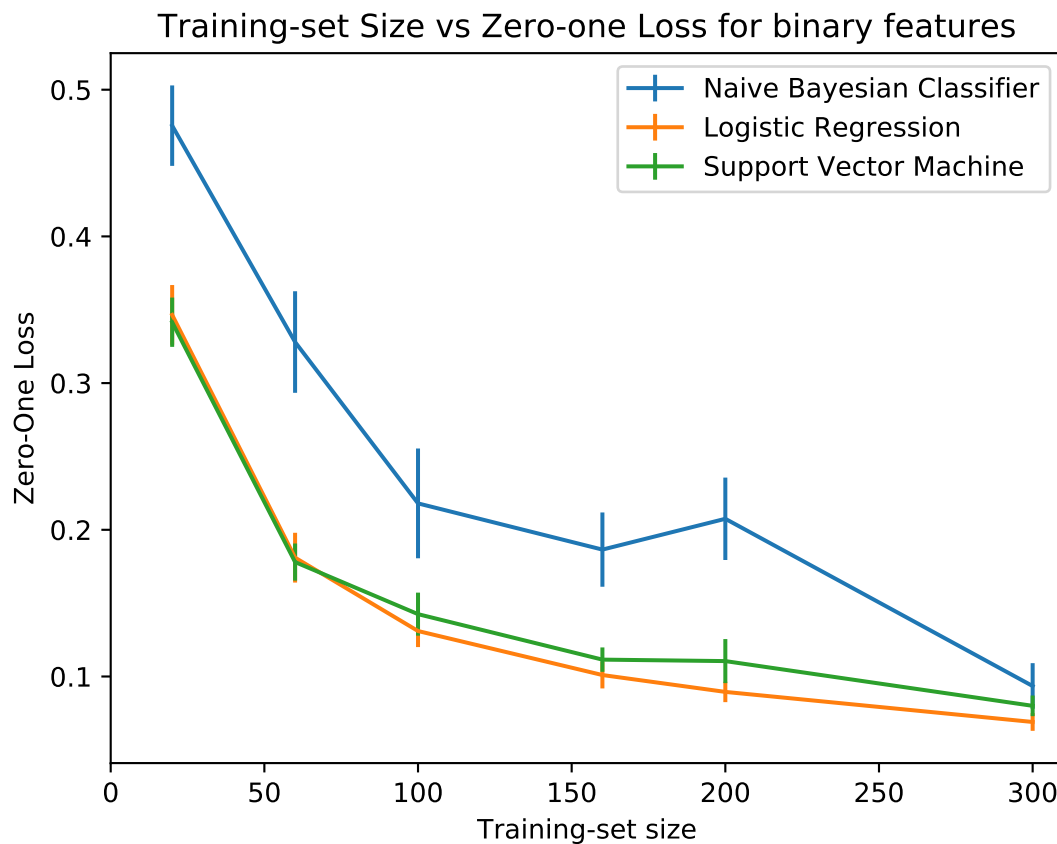


Figure 2: TSS vs Zero-one loss for NBC vs LR vs SVM with binary features

Problem 2

Analysis

1.

(a)

Figure 2 shows the learning curves for the three models with binary features.

(b)

Null Hypothesis: Logistic regression performs as good as Naive Bayesian Classifier

Alternate Hypothesis: There is a performance difference between Logistic Regression and Naive Bayesian Classifier models

OR

Null Hypothesis: Support Vector Machine performs as good as the Naive Bayesian Classifier

Alternate Hypothesis: There is a performance difference between Support Vector Machine and Naive Bayesian Classifier models

OR

Null Hypothesis: Support Vector Machine performs as good as Logistic Regression

Alternate Hypothesis: There is a performance difference between Support Vector Machine and Logistic Regression models

I select the significance level as $\alpha = 0.05$

Model/TSS	20	60	100	160	200	300
NBC	0.4755	0.328	0.218	0.1865	0.2075	0.0935
LR	0.3465	0.181	0.131	0.101	0.0895	0.069
SVM	0.3415	0.178	0.1425	0.1115	0.1105	0.08

Table 1: Mean zero-one loss for the three models with binary features

```

priyank@priyank: ~/Desktop/Dropbox/Spring 2017/DM/My HW Solutions/HW3
TSS 200
TSS 200
TSS 300
TSS 300
TSS 300
TSS 300
TSS 300
TSS 300
TSS 300
TSS 300
TSS 300
TSS 300
TSS 300
TSS 300
Means with NBC with binary features [0.4754999999999998, 0.32800000000000001, 0.21800000000000003, 0.1865, 0.20
7499999999999996, 0.0935]
Means with NBC with features with 3 values [0.4449999999999995, 0.38, 0.3175, 0.2334999999999999, 0.2400000000
0000005, 0.1369999999999998]
Means with LR with binary features [0.3464999999999997, 0.1809999999999999, 0.13100000000000003, 0.10100000000
000002, 0.08949999999999996, 0.06900000000000002]
Means with LR with features with 3 values [0.29700000000000004, 0.16900000000000001, 0.13400000000000001, 0.0924
9999999999999, 0.10100000000000001, 0.07599999999999998]
Means with SVM with binary features [0.34150000000000003, 0.1779999999999999, 0.14250000000000002, 0.1115, 0.11
050000000000001, 0.08000000000000002]
Means with SVM with features with 3 values [0.2874999999999998, 0.1784999999999999, 0.14250000000000002, 0.112
500000000000002, 0.09750000000000003, 0.08050000000000002]
ttest_rel for NBC vs LR: t = 5.55279 p = 0.00260349
ttest_rel for NBC vs SVM: t = 4.56679 p = 0.0060195
ttest_rel for SVM vs LR: t = 1.90513 p = 0.115104
ttest_rel for NBC with binary-valued features vs NBC with features with three values: t = -2.37857 p = 0.063277
ttest_rel for LR with binary-valued features vs LR with features with three values: t = 0.891802 p = 0.413353
ttest_rel for SVM with binary-valued features vs SVM with features with three values: t = 1.21569 p = 0.278366
priyank@priyank:~/Desktop/Dropbox/Spring 2017/DM/My HW Solutions/HW3$

```

Figure 3: Hypothesis Testing Output

(c)

Table 1 shows the mean zero-one loss across 10 folds for the three models with binary features. As we can see from figure 3:

For the null hypothesis: "Logistic Regression (LR) performs as good as Naive Bayes Classifier (NBC)", the probability is $p = 0.0026$, which is less than the significance level. So we reject the null-hypothesis. This implies that the alternate hypothesis is true: There is a performance difference between Logistic Regression and Naive Bayesian Classifier

OR

For the null hypothesis: "Support Vector Machine (SVM) performs as good as Naive Bayesian Classifier", the probability is $p = 0.006$, which is less than the significance level. So we reject the null-hypothesis. This implies that the alternate hypothesis is true: There is a performance difference between Support Vector Machine and Naive Bayesian Classifier.

OR

For the null hypothesis: "Support Vector Machine performs as good as Logistic Regression", the probability is $p = 0.115$, which is greater than the significance level. So we accept the null-hypothesis. This implies that the null-hypothesis is true: Support Vector Machine performs as

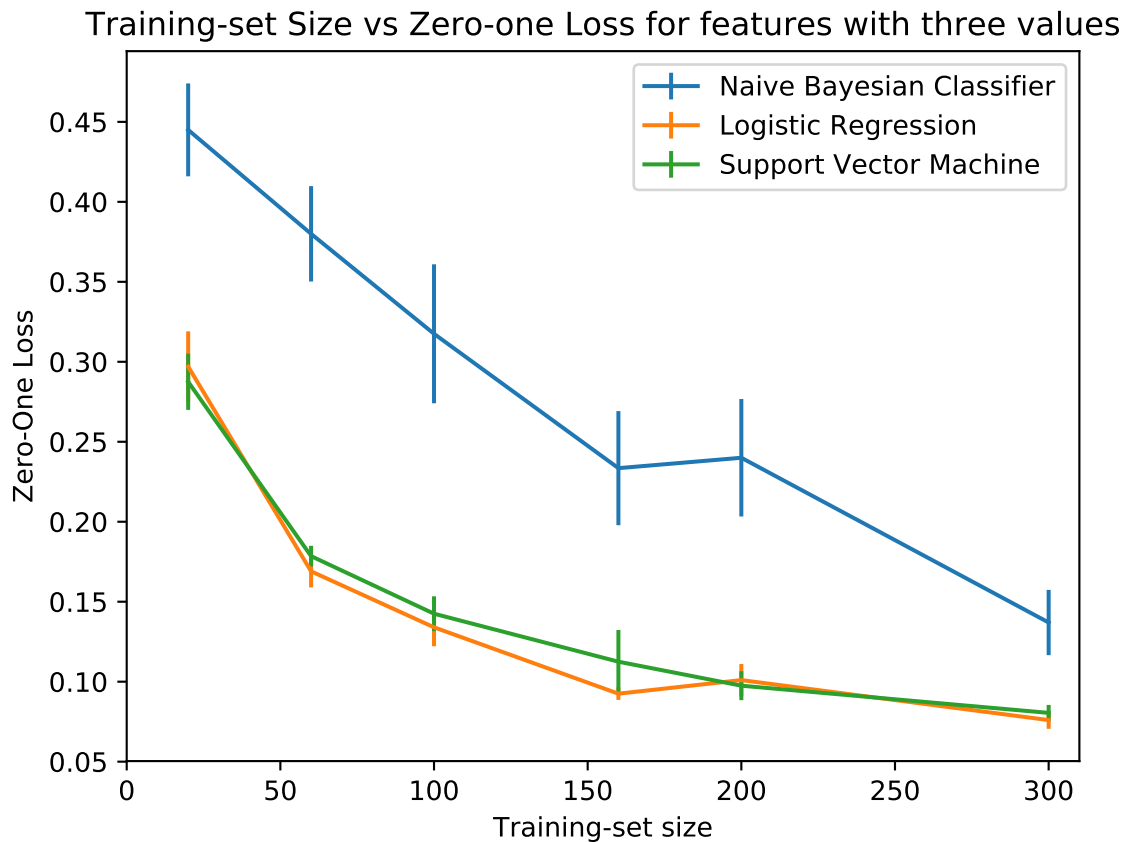


Figure 4: TSS vs Zero-one loss for NBC vs LR vs SVM with features with three values

good as Logistic Regression”

2

(a) Figure 4 shows the learning curves for the three models with features with three values.

(b)

Null Hypothesis: LR with binary features performs as good as LR with features with three values

Alternate Hypothesis: There is a performance difference between LR with binary features and LR with features with three values

OR

Null Hypothesis: SVM with binary features performs as good as SVM with features with three values

Alternate Hypothesis: There is a performance difference between SVM with binary features and SVM with features with three values

I select the significance level as $\alpha = 0.05$

(c)

Table 2 shows the mean zero-one loss across 10 folds for the three models with features with

Model/TSS	20	60	100	160	200	300
NBC	0.445	0.38	0.3175	0.2335	0.24	0.137
LR	0.297	0.169	0.134	0.0925	0.101	0.076
SVM	0.2875	0.1785	0.1425	0.1125	0.0975	0.08

Table 2: Mean zero-one loss for the three models with features with three values

three values. As we can see from figure 3:

For the null hypothesis: "LR with binary features performs as good as LR with features with three values", the probability is $p = 0.413$, which is above the significance level. So we accept the null-hypothesis. This implies that the null hypothesis is true: LR with binary features performs as good as LR with features with three values.

OR

For the null hypothesis: "SVM with binary features performs as good as SVM with features with three values", the probability is $p = 0.278$, which is above the significance level. So we accept the null-hypothesis. This implies that the null hypothesis is true: SVM with binary features performs as good as SVM with features with three values.