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

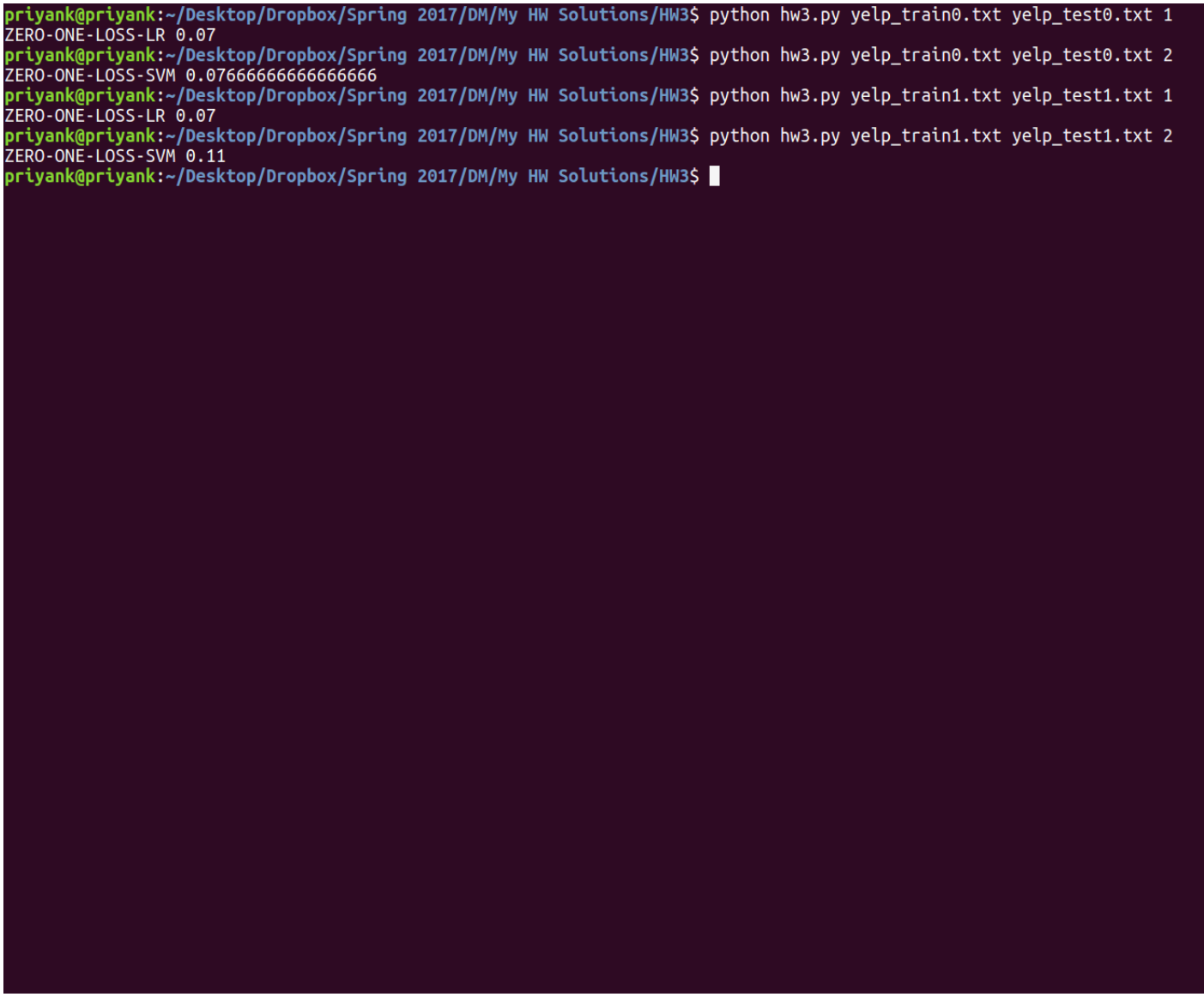

Figure 1: Output for testcases of Homework 2

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

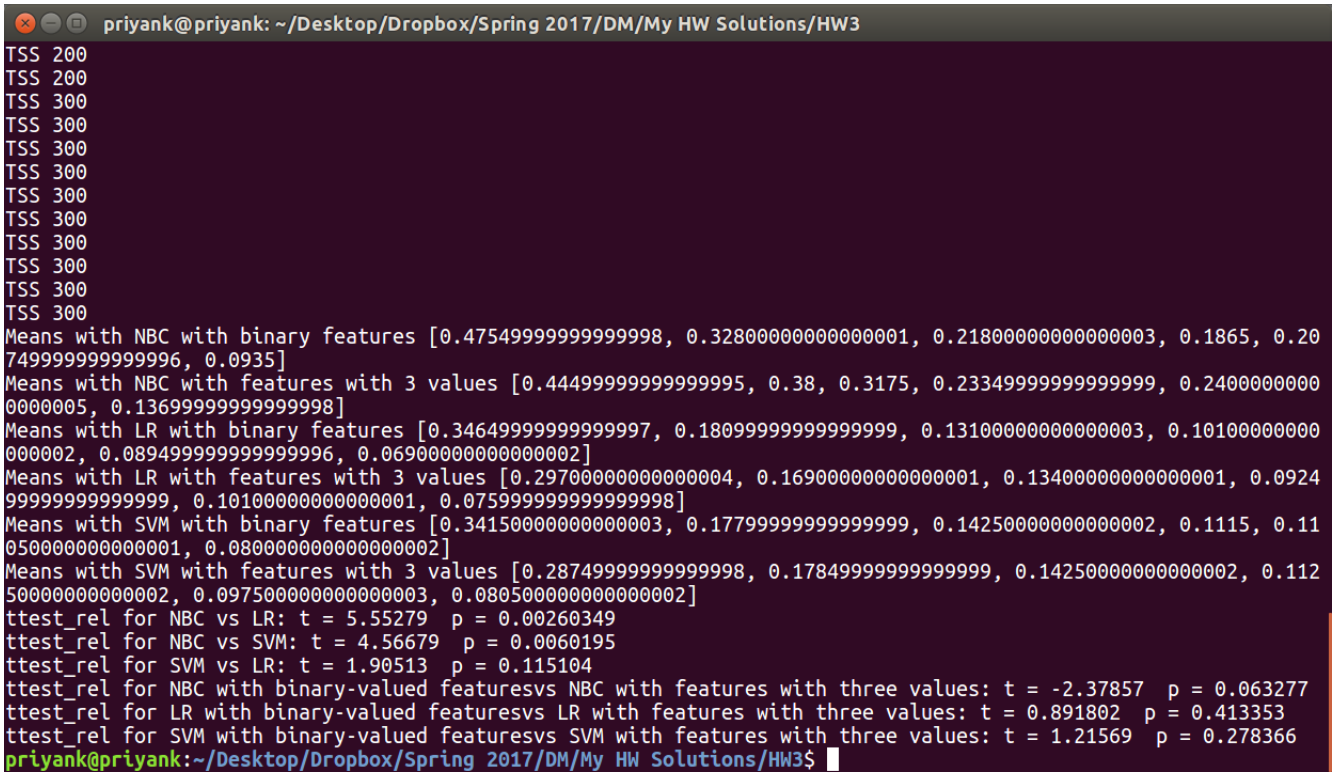

Figure 3: Hypothesis Testing Output

(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

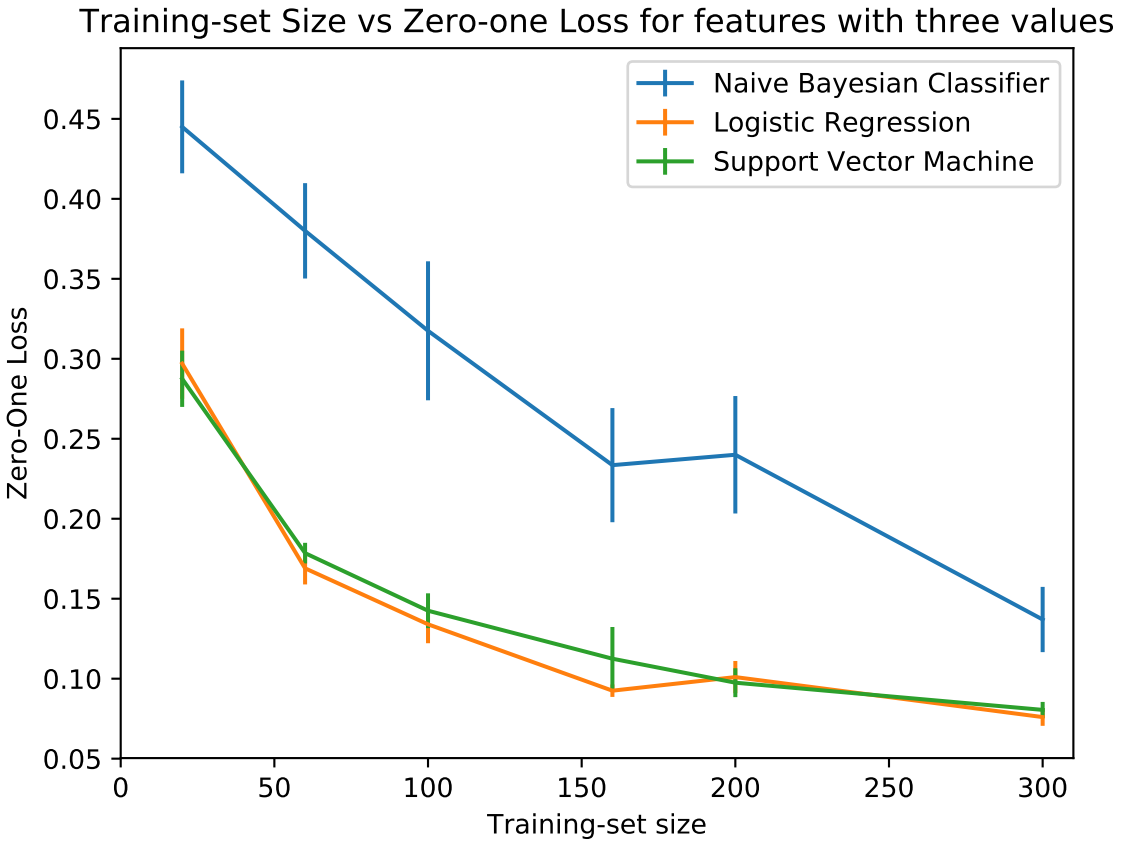

Figure 4: TSS vs Zero-one loss for NBC vs LR vs SVM with features with three values

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 sigificance 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.