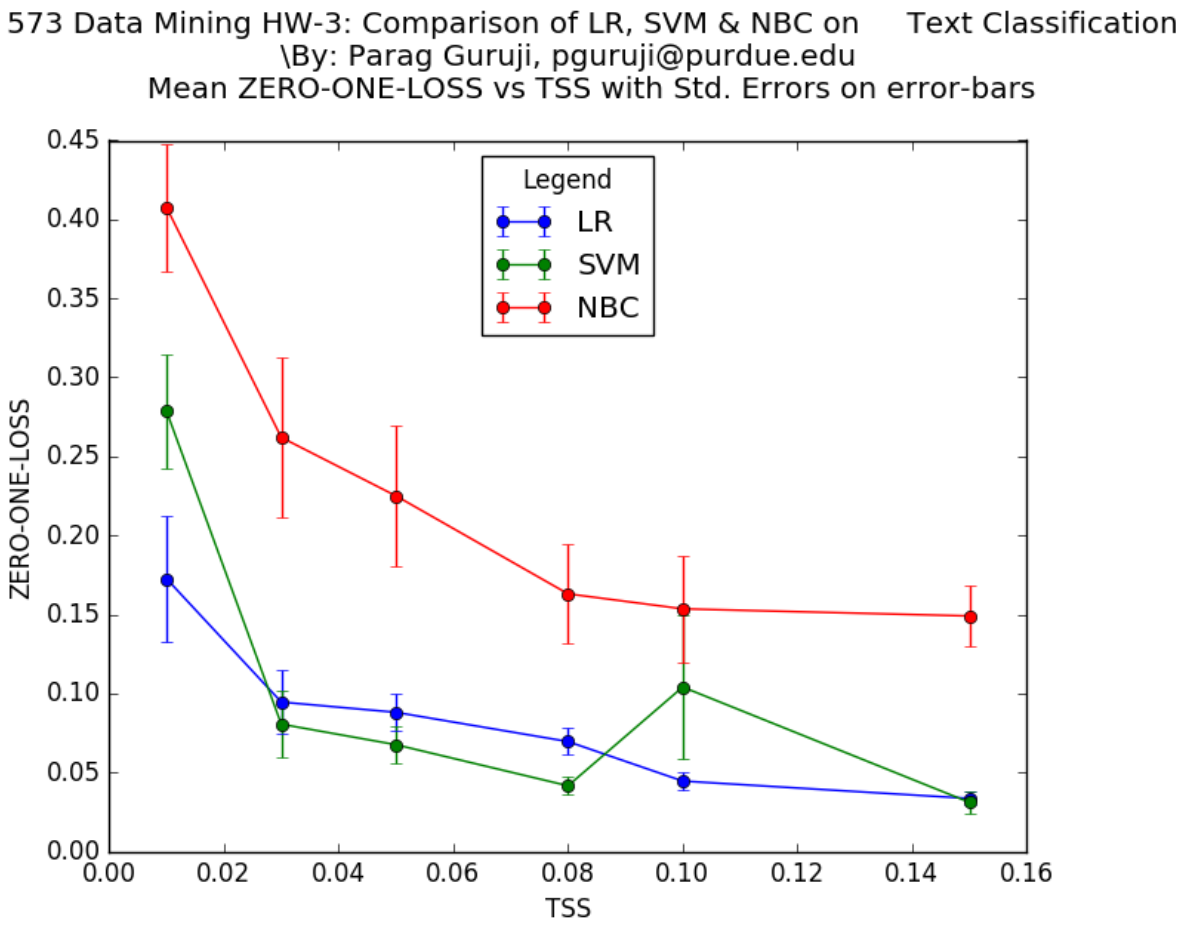
CS 573 Data Mining Homework 2

By: Parag Guruji, pguruji@purdue.edu

Date: 15 Mar 2017 – USING ONE (first) LATE DAY

Analysis:

1. **Choice of model**
   a. **Plot of learning curve**



The above graph shows comparison of the performances of three classifiers in terms of **mean zero-one loss across 10 trials – 1 trial per fold in the K-fold cross-validation setting** (here, K=10) against each of the **training set sizes** (TSS) expressed **as the proportions of the whole yelp dataset** (D) - [0.01, 0.03, 0.05, 0.08, 0.10, 0.15] with **error-bars showing the standard error** of zero one loss. The corresponding test set is disjoint with the training set and is kept at the constant size of |D|/K. The number of true features (excluding bias) is 4000 in all trials.

- All the three models show the trend of performance improvement with increasing TSS (the one exception in SVM's learning curve is an aberration as found in several other runs)

- The improvement in performance is rapid initially and decays at higher TSS (exponential decay)

- The LR & SVM generally always beat the NBC in the performance at all TSS.

- The std. error also diminishes wit rise in TSS – more so in LR and SVM as compared to NBC

b. Hypothesis Formulation

Models: LR & SVM at TSS $= 0.15$ (highest TSS, -> most stable available model)

Null Hypothesis:

LR and SVM both perform materially the same. i.e. difference between their mean zero-one loss is statistically insignificant.

$$H_0: \mu_{LR} - \mu_{SVM} = 0$$

Alternative Hypothesis:

LR and SVM perform significantly differently. i.e. difference between their mean zero-one loss is statistically significant.

$$H_a: \mu_{LR} - \mu_{SVM} \neq 0$$

c. Hypothesis Testing

Testing the above hypothesis at $\propto = 0.05$, i.e. at 95% confidence level

Two-Sample T-Test for mean zero-one loss from the mean zero-one loss at TSS $= 0.15$:

```
$ python hw3.py "" "" 1 -t 1 2
```

T-test stats: t $= 1.27205$ p $= 0.219559 > \propto$

Thus, We FAIL TO REJECT the Null Hypothesis H_0 from the evidence given by our data

The difference in mean zero-one-loss of LR and that of SVM is statistically **insignificant at 95.0% Confidence Level**. i.e. Both models **materially perform the same**.

Test Results for similar hypothesis for **LR and NBC**:

```
$ python hw3.py "" "" 1 -t 1 3
```

T-test stats: t $= -3.44669$ p $= 0.00457582 < \propto$

Thus, We REJECT the Null Hypothesis H_0 from the evidence given by our data

The difference in mean zero-one-loss of LR and that of NBC is statistically significant at 95.0% Confidence Level. i.e. Both models **materially DO NOT perform the same**. (LR is better)

Test Results for similar hypothesis for **SVM and NBC**:

```
$ python hw3.py "" "" 1 -t 2 3
```

T-test stats: $t = -4.19446$ $p = 0.00118291 < \propto$

Thus, We REJECT the Null Hypothesis H_0 from the evidence given by our data

The difference in mean zero-one-loss of SVM and that of NBC is statistically significant at 95.0% Confidence Level. i.e. Both models **materially DO NOT perform the same**. (SVM is better)

Thus, we can conclude that choosing model LR or SVM over NBC can improve the performance but the choice in between LR & SVM is immaterial/insignificant w.r.t. performance.

2. **Feature Reconstruction:**

We observe that there is no significant difference in performance for the 2 different ways of feature construction from observation of results obtained at TSS=0.15 for each model.

Feature Values	TSS	LR_mean	SVM_mean	NBC_mean	LR_std_err	SVM_std_err	NBC_std_err
[0, 1]	0.15	0.039	0.029	0.0865	0.005648	0.005468	0.012571
[0, 1, 2]	0.15	0.0335	0.031	0.149	0.004419	0.006738	0.019362

To verify our observation, we perform the 2 sample t-test as above among the mean zero-one loss of each of the model from the results with both the feature-value-sets.

The results are as follows:

Each Null hypothesis is that the mean zero one loss for given model is same across both the feature-value sets.

Each alternative hypothesis is that the mean zero-one loss for given model significantly varies across the feature-value-sets.

Results:

```
$ python hw3.py "" "" 1 -v "D:\spyder
projects\DMHW3\output\comparison_f1.csv" "D:\spyder
projects\DMHW3\output\comparison_f2.csv"
ttest stats: t = 0  p = 1
```

```
The difference in mean zero-one-loss of LR in results with
feature values [0, 1] and [0, 1, 2] is statistically
insignificant at 95.0% Confidence Level. i.e. The model
materially perform the same with both feature-value-sets.

ttest stats: t = 0  p = 1
The difference in mean zero-one-loss of SVM in results with
feature values [0, 1] and [0, 1, 2] is statistically
insignificant at 95.0% Confidence Level. i.e. The model
materially perform the same with both feature-value-sets.

ttest stats: t = 0  p = 1
The difference in mean zero-one-loss of NBC in results with
feature values [0, 1] and [0, 1, 2] is statistically
insignificant at 95.0% Confidence Level. i.e. The model
materially perform the same with both feature-value-sets.
```

Thus, we conclude that both the ways of feature construction are materially perform the same. Choice among them is immaterial as far as the performance is concerned.