CS57300: Assignment 3

Due date: Friday March 8, 11:59pm (submit pdf to Blackboard)

Comparing Methods for Speed Dating Classification

1 Preprocessing (4 pts)

```
$python preprocess-assg3.py
Mapped vector for female in column gender:  [1.]
Mapped vector for Black/African American in column race:  [0.  1.  0.  0.]
Mapped vector for Other in column race_o:  [0.  0.  0.  0.]
Mapped vector for economics in column field:  [0.  0.  0.  0.  0.  0.  0.
0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  1.  0.  0.  0.  0.  0.
0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.]
```

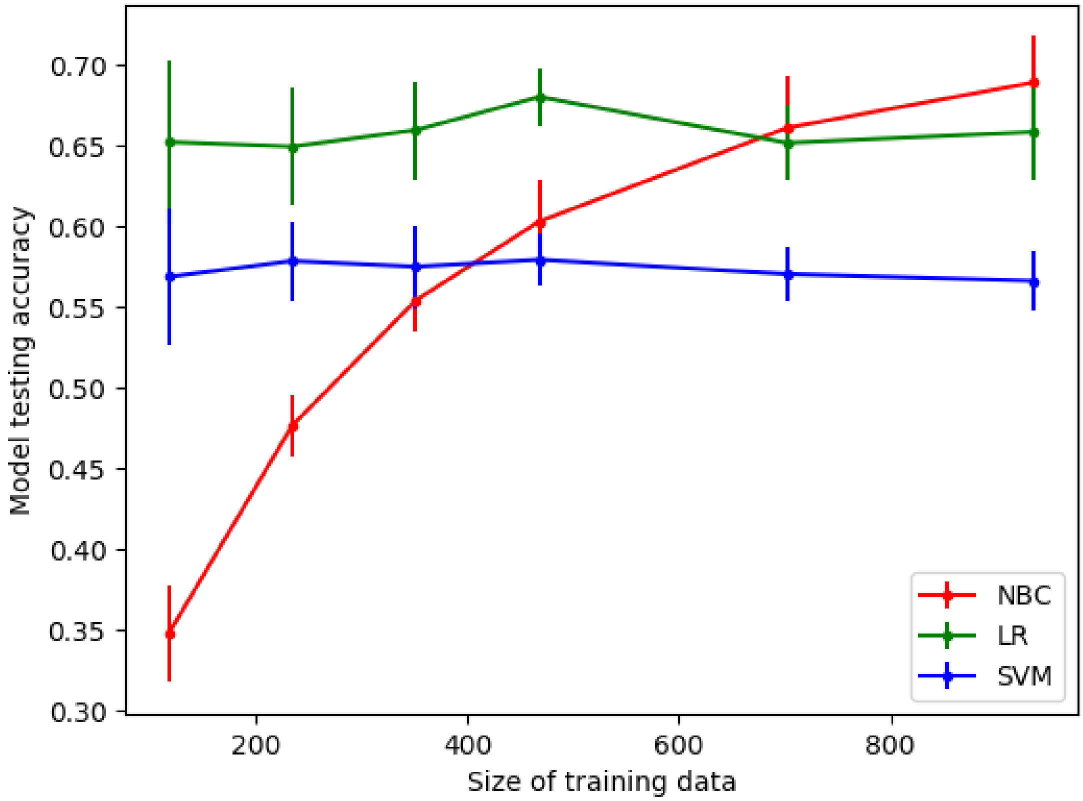
2 Implement Logistic Regression and Linear SVM (16 pts)

1. Expected Output:
   Training Accuracy LR: 0.65
   Testing Accuracy LR: 0.65

2. Expected Output:
   Training Accuracy SVM: 0.57
   Testing Accuracy SVM: 0.57

3 Learning Curves and Performance Comparison (10 pts)

1. Learning Curve (use random_state=47 to preprocess the NBC data while use random_state=25 to preprocess the LR and SVM data)



2, 3. All the possible H0 and corresponding t-statistics, p-values, as well as reject or not are as follows:

```
Fraction:0.025 H0 for NBC and LR: t-statistics=-15.907, p-value=6.767e-08
Reject with significance level of 0.01?  True
Fraction:0.025 H0 for NBC and SVM: t-statistics=-16.45, p-value=5.02e-08
Reject with significance level of 0.01?  True
Fraction:0.025 H0 for LR and SVM: t-statistics=3.14, p-value=0.011 Reject
with significance level of 0.01?  False

Fraction:0.05 H0 for NBC and LR: t-statistics=-18.28, p-value=1.99e-08 Re-
ject with significance level of 0.01?  True
Fraction:0.05 H0 for NBC and SVM: t-statistics=-9.28, p-value=6.58e-06 Re-
ject with significance level of 0.01?  True
Fraction:0.05 H0 for LR and SVM: t-statistics=4.31, p-value=0.0019 Reject
with significance level of 0.01?  True

Fraction:0.075 H0 for NBC and LR: t-statistics=-10.23, p-value=2.95e-06 Re-
ject with significance level of 0.01?  True
Fraction:0.075 H0 for NBC and SVM: t-statistics=-2.32, p-value=0.045 Re-
ject with significance level of 0.01?  False
Fraction:0.075 H0 for LR and SVM: t-statistics=11.29, p-value=1.28e-06 Re-
```

```
ject with significance level of 0.01?  True

Fraction:0.1 H0 for NBC and LR: t-statistics=-8.44, p-value=1.43e-05 Reject with significance level of 0.01?  True
Fraction:0.1 H0 for NBC and SVM: t-statistics=2.68, p-value=0.025 Reject with significance level of 0.01?  False
Fraction:0.1 H0 for LR and SVM: t-statistics=13.34, p-value=3.09e-07 Reject with significance level of 0.01?  True

Fraction:0.15 H0 for NBC and LR: t-statistics=0.66, p-value=0.51 Reject with significance level of 0.01?  False
Fraction:0.15 H0 for NBC and SVM: t-statistics=7.83, p-value=2.60e-05 Reject with significance level of 0.01?  True
Fraction:0.15 H0 for LR and SVM: t-statistics=7.17, p-value=5.24e-05 Reject with significance level of 0.01?  True

Fraction:0.2 H0 for NBC and LR: t-statistics=2.67, p-value=0.025 Reject with significance level of 0.01?  False
Fraction:0.2 H0 for NBC and SVM: t-statistics=10.63, p-value=2.14e-06 Reject with significance level of 0.01?  True
Fraction:0.2 H0 for LR and SVM: t-statistics=6.89, p-value=7.083e-05 Reject with significance level of 0.01?  True
```