CS573 Assignment3

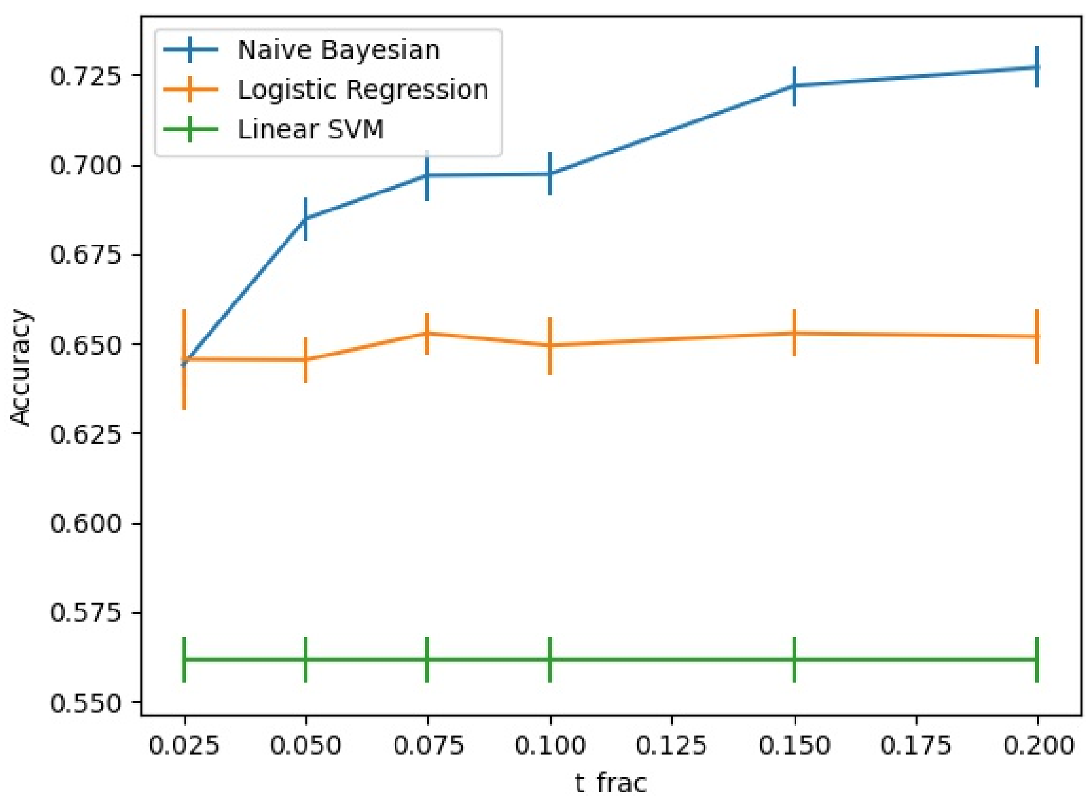
Sandy Hsiao

1. Preprocess
   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, 1, 0]

2. Implement Logistic Regression and Linear SVM
   Training Accuracy LR: 0.66
   Testing Accuracy LR: 0.67
   Training Accuracy SVM: 0.56
   Testing Accuracy SVM: 0.58

3. Learning Curves and Performance Comparison

   (a) Model comparison
   The figure shows that the accuracy increases as training size increase for Naive Bayes Classifier. However, the accuracy does not change much for both Logistic Regression and Linear SVM under the current learning iteration and learning rate specified in part 2. Also, with the current dataset, LR has better accuracy than Linear SVM.



   (b) hypothesis
   Null hypothesis: LR and SVM have similar expected accuracy.
   Alternative hypothesis: There is a difference in the expected accuracy of the two models.

   (c) Evidence
   We can use the paired t-test to test the null hypothesis. The t-test statistic and p-value can be easily computed

using ttest_rel from scipy.stats package. By using the mean accuracy calculated from 3.(a), we get the following t-test result.

```
LR vs SVM
Ttest_relResult(statistic=61.85165540554506, pvalue=2.090892384600833e-08)
```

Since the p-value is way smaller than 0.01, we can reject the null hypothesis. Therefore, we conclude that there is indeed a difference in the expected accuracy of the two models with the current dataset.