Question. 1

Active Learning Using Support Vector Machines

a) Downloading the dataset

b) Passive and Active Learning

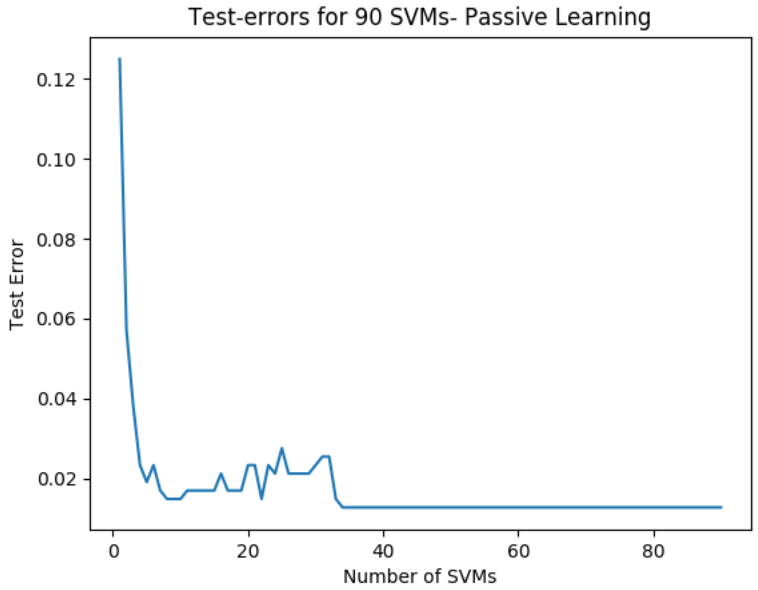
i. Passive Learning:

I shuffled the dataset and then separated out 472 datapoints as Test Data.
This part of the question made use of **LinearSVC** from the classes of SVMs in the sklearn library.

The following was used to tune the hyper-parameter employed for the **L1** penalty on the model.

```
Cs=np.linspace(0.01,5,50)
```

A *10-fold Cross validation* technique was achieved via the use of **GridsearchCV** from the sklearn library.
The following curve was obtained after plotting the Test-errors for all *90 SVMs:*



As observed, the error-rate decreases rapidly after adding more data-points to the Training set, which is obvious. Although, after around 5-6 iterations with about 60-70 data-points in the training set the error-rate does not decrease significantly.

However, there is fluctuation in the error-rate while using 100-300 data-points, it becomes virtually stable after the 37th iteration(370-380+ data-points in training set).

ii. Active Learning:

To perform active learning instead of picking 10 random datapoints from the X_train_remaining(say train_buffer) at each iteration to train the model, I utilized the ***decision_function*** in the LinearSVC class.

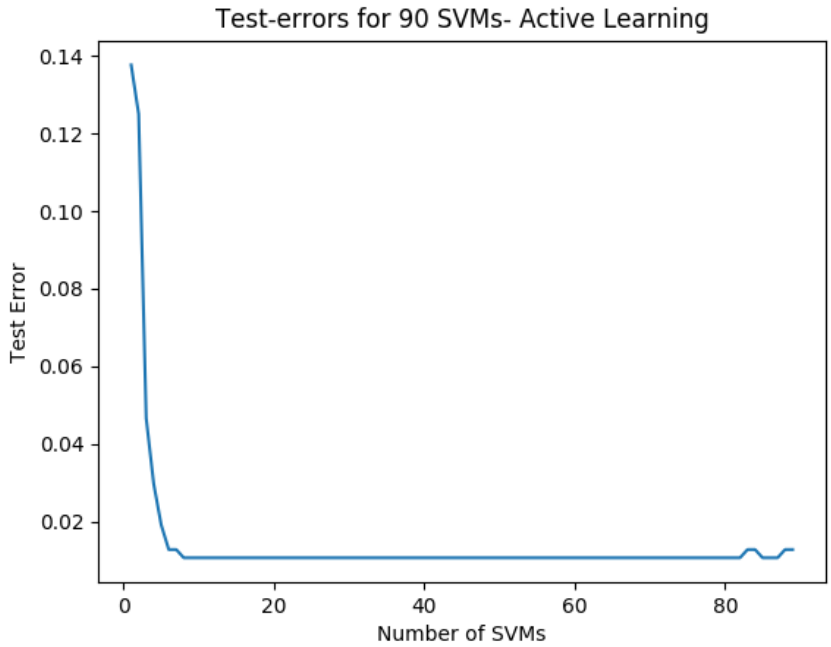
The following code provided me with a vector of the remaining training data and their distance from the hyperplane:

```
X_margin = clf.decision_function(X_rest)
X_margin = np.abs(X_margin)
```

I then put this in a dictionary and sorted it preserving their index(as the indices will help me pick the datapoints from the remaining train_buffer).

These 10 are the closest datapoints to the hyperplane, thus they were added to the training set and removed from the buffer for the next iteration.

The following results were observed:



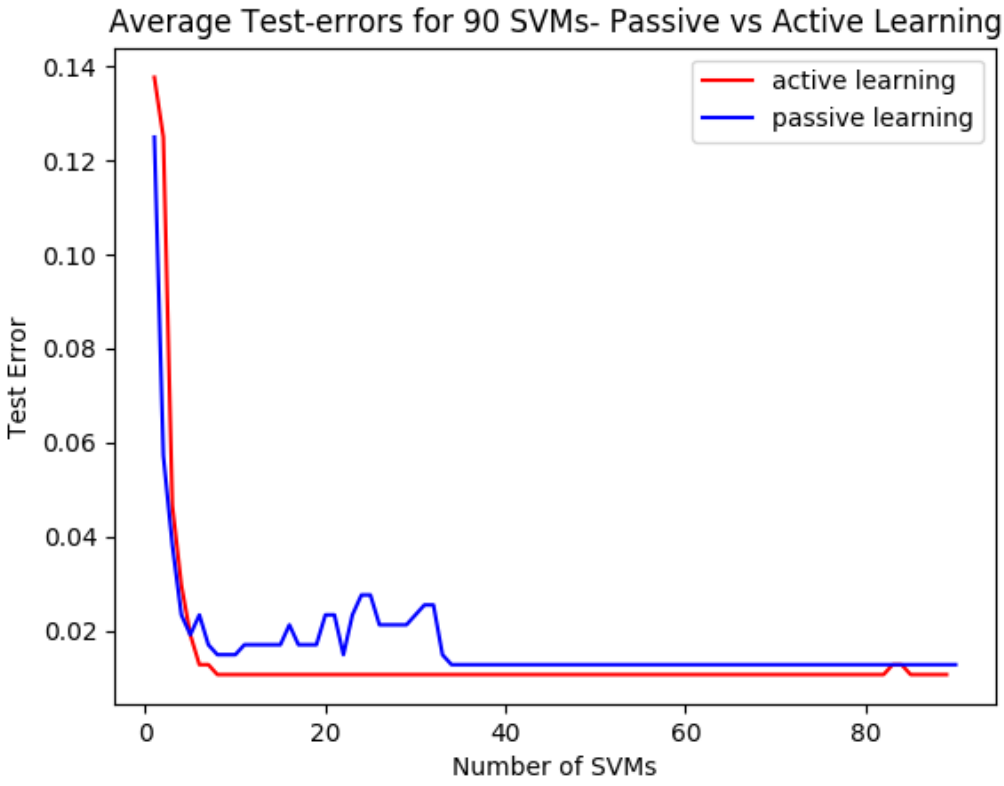
As observed, the error-rate decreases rapidly after adding more data-points to the Training set in a similar fashion to passive learning. Notice, after around 6-7th iterations with about 70-80 data-points in the training set the error-rate does not decrease significantly and becomes fairly stable for the rest of the runtime.

Even though there is fluctuation in the error-rate in the later iterations, it is not of greater significance. This might well be due to existence of certain outliers in the data.

c) Monte Carlo Simulation:

In order to perform the Monte Carlo simulation of the data, both active and passive learning strategies were experimented on the data.

The following results were obtained:



The smoothness in the curve of active learning denotes a more stable model with less error even when the data is not significantly large in amount.