Programming project-1

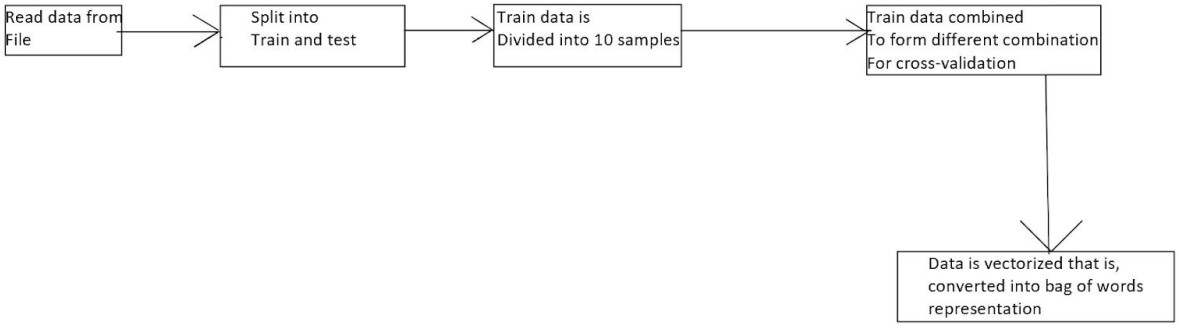
Citation: Discussed on a high level with Aditya Kartikeya

Aim:

1) Plot the learning curve and error for different train-sizes for m=0 and m=1.
2) Run stratified cross-validation and plot accuracy and standard deviation as a function of m for values between 0.1 - 1 and 1 to 10.

Procedure:

Data Preprocessing:



Experiment-1:

In the first experiment, the train data was first divided into 10 sets of samples.
Then the size of the data was increased iteration to iteration and the accuracy for measured for each iteration and later the error was also calculated.In the first iteration, say for example the train data was sample_set_1. Then, for the second iteration the dataset would be sample_set_1+sample_set_2.In this way, the size of the data was recursively increased.This is done for m=0 and m=1.In the end, accuracies and errors for m=0 and m=1 are plotted in a graph.

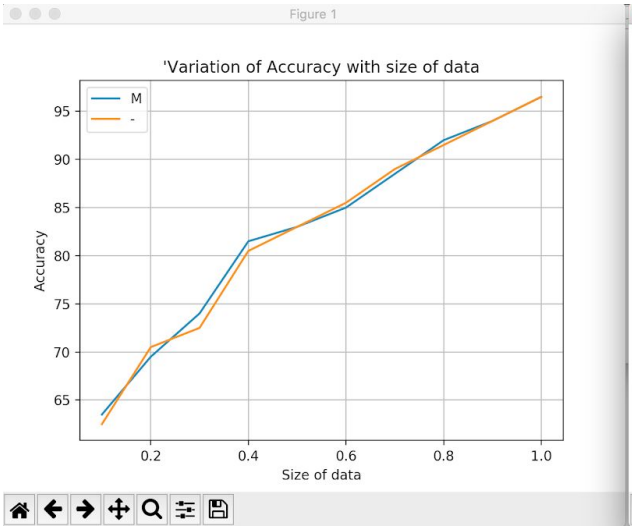
Experiment-2:

In this experiment, we divide the data into 10 parts again. Here, each time we train the model on 9 sets of data and keep the excluded set for testing. In this way, we train 10 different models and measure accuracy and standard deviation for each.Here, we do this

for m=0,0.1,0.2...1 and m=2,3,4…,10.Hence, in the end we plot the accuracies against different values of m.

Results for IMDB dataset:

m=0:orange m=1: blue



m=0:orange m=1: blue

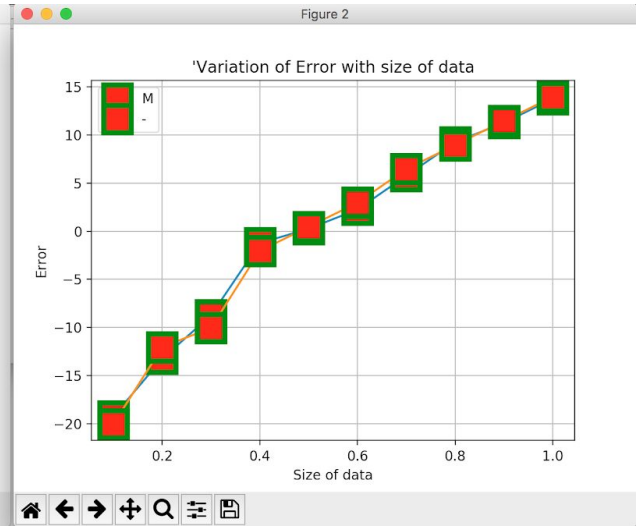
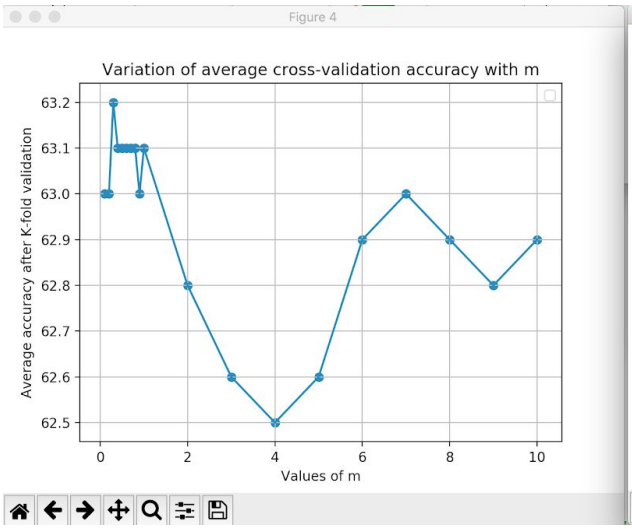


Fig 1: The above figure gives the variation of accuracy and error with the size of data for m=0 and m=1



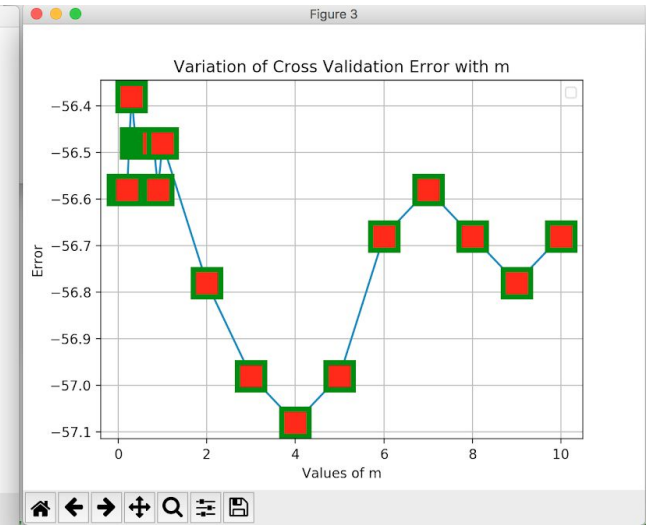


Fig 2: The above gives variation of cross-validation accuracy and error with respect to the smoothening parameter(m)

Conclusions for IMDB dataset:

1) The graph of variation of accuracy with an increase in the size of data clearly shows us that the accuracy increases gradually with increase in size of data.

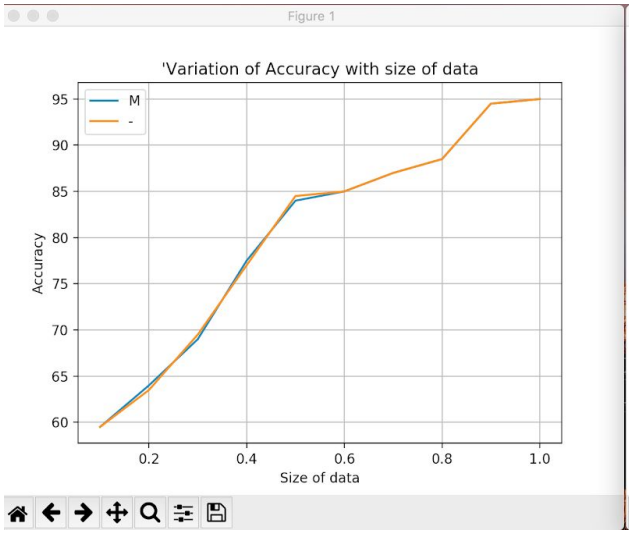
2) The best accuracy obtained is approximately 97% from the graph.This occurs when we use the entire train data for training.
The above points are valid for both m=0 and m=1.

3) From the variation of error bar plot with size, we can see that as we increase the size of data the length of the error bar goes on decreasing which means that error decreases as we increase size of the dataset.

4) In Fig 2, we can observe no particular pattern, but one thing we can say for sure is the fact that for lower values of m, we are obtaining better values for accuracy.

Results for Yelp dataset:

m=0:orange m=1: blue



m=0:orange m=1: blue

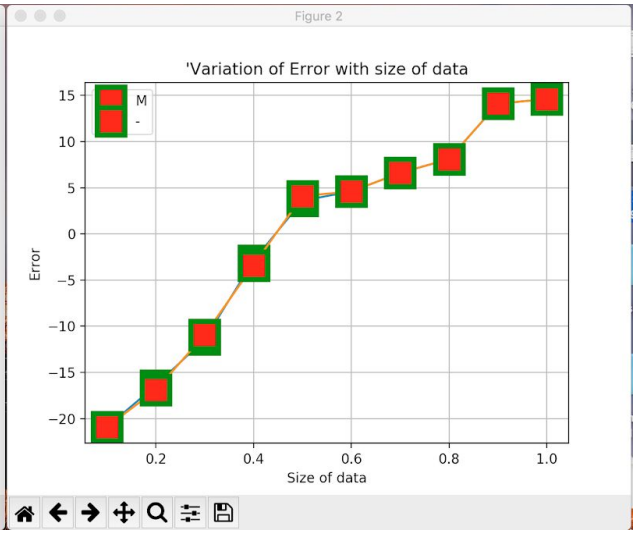
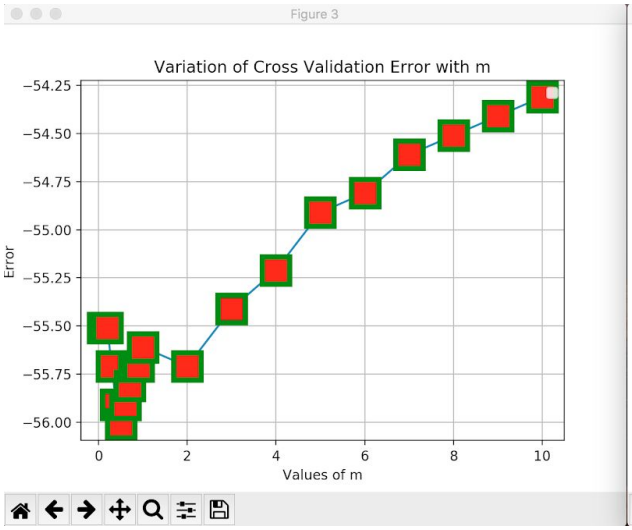


Fig 3: The above figure gives the variation of accuracy and error with the size of data for m=0 and m=1



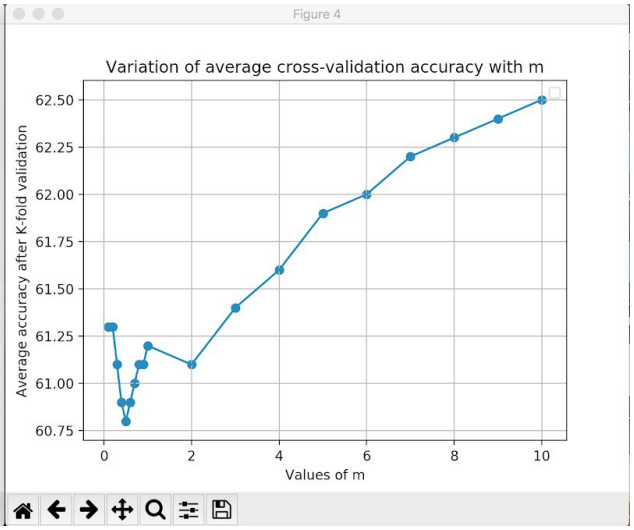


Fig 4: The above gives variation of cross-validation accuracy and error with respect to the smoothening parameter(m)

Conclusions for IMDB dataset:

1) From figure 3, in this case too we can say that the accuracy increases with increase in data size. This applies to both m=0 and m=1.

2) The maximum accuracy obtained for both m=0 and m=1 is 95%.

3) For this dataset, from figure 4, we get a higher cross-validation accuracy for higher values of m. We get maximum accuracy for m=10.

4) The length of error bars are almost contant for different values of m, hence it looks like the error does not vary much with change in m.

Results for Amazon dataset:

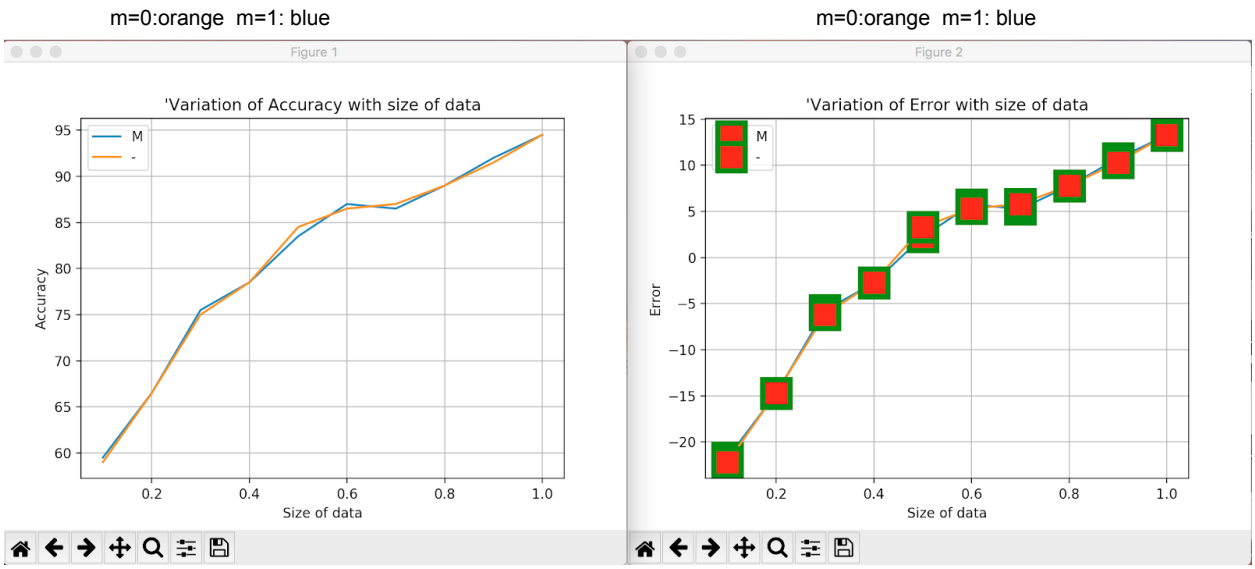


Fig 5: The above figure gives the variation of accuracy and error with the size of data for m=0 and m=1

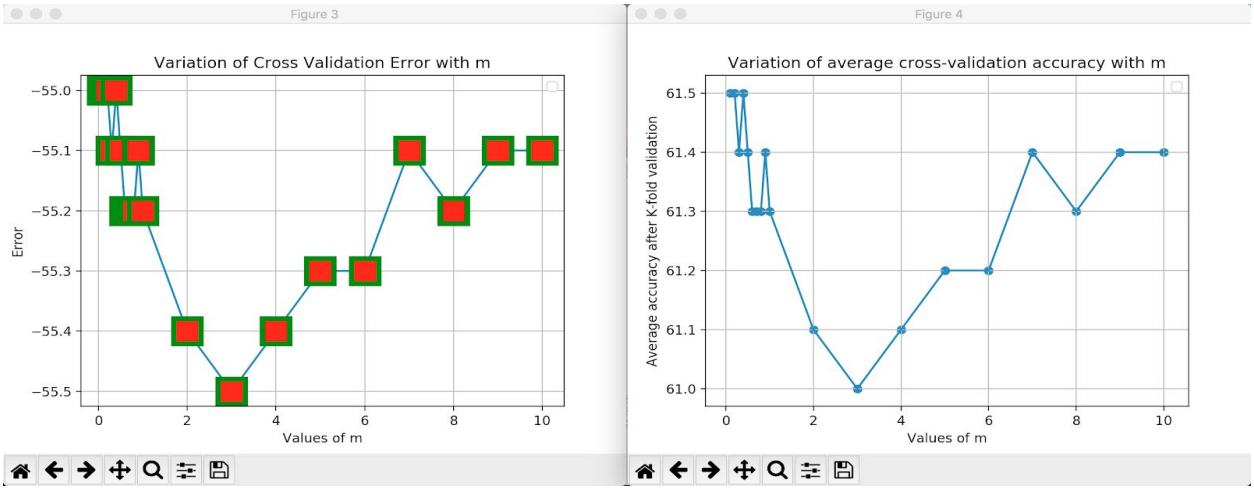


Fig 6: The above gives variation of cross-validation accuracy and error with respect to the smoothening parameter(m)

Conclusions for Amazon dataset:

1) Figure 5 shows us that accuracy increases for both m=0 and m=1 with increase in size of data.

2) The maximum accuracy is approximately 94%.

3) Also, the length of error bars are little shorter for large-sized datasets.

4) The cross-validation accuracy can be observed to be maximum for very low values of m. Hence, ideal value of m should lie somewhere between 0 and 1.
