Assignment1

Experiments:

1. For each of the 3 datasets run stratified cross validation to generate learning curves for Naive Bayes with m = 0 and with m = 1. For each dataset, plot averages of the accuracy and standard deviations (as error bars) as a function of train set size. It is insightful to put both m = 0 and m = 1 together in the same plot. What observations can you make about the results.

We have 3 datasets:

- imdb_labelled.txt
- amazon_cells_labelled.txt
- yelp_labelled.txt

In following we can see what are the results.

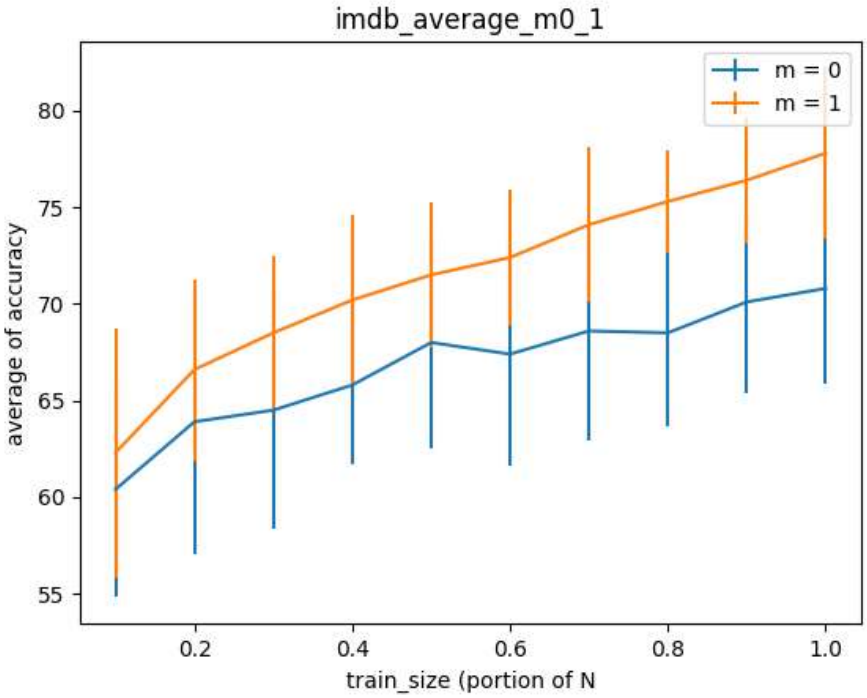

***Figure 1**: averages of the accuracy and standard deviations (as error bars) as a function of train set size for imdb data set*

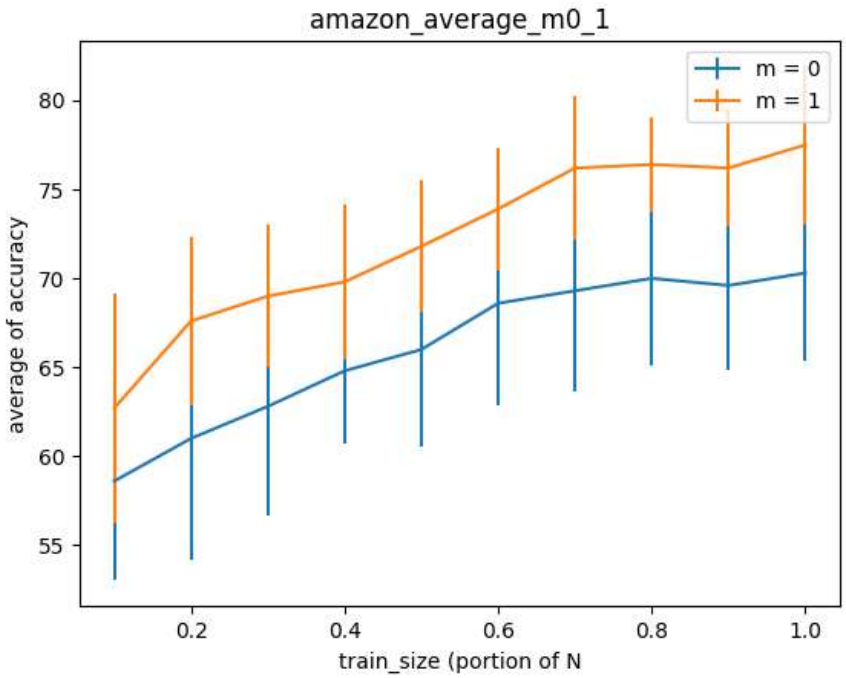

***Figure 2**: averages of the accuracy and standard deviations (as error bars) as a function of train set size for amazon data set*

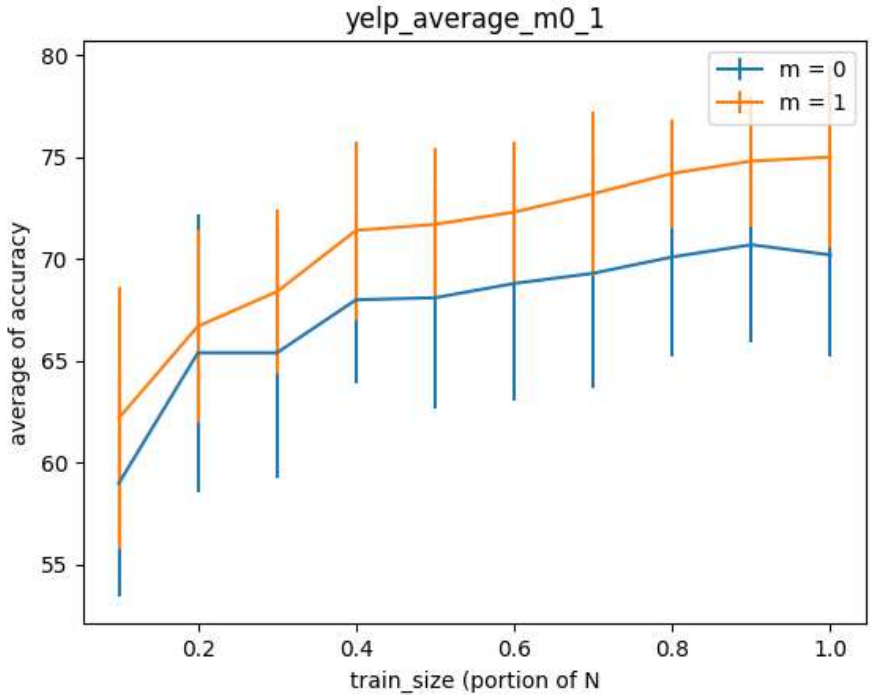

***Figure 3**: averages of the accuracy and standard deviations (as error bars) as a function of train set size for amazon data set*

As we can see, the averages of accuracy for smoothing parameter m = 0 are less than m = 1. It happens because if we have a word in test data that has not appear in training data for a label (positive or negative), so the probability of that word given that class would be 0. Since we multiply the probability of words given class for each sentence to calculate the probability of that sentence given class, the probability become 0 because of that specific word. Therefore, it is not a very accurate result. By adding smoothing parameter, we make our result more accurate.

Also, we can see that in all 3 data sets, by incrementing the number of samples in train data, we can get a better result because we have more data to learn from. (The probability for each word given class (positive or negative) would be more accurate by seeing more data)

2. Run stratified cross validation for Naive Bayes with smoothing parameter m = 0, 0.1, 0.2, . . . , 0.9 and 1, 2, 3, . . . , 10 (i.e., 20 values overall). Plot the cross validation accuracy and standard deviations as a function of the smoothing parameter. What observations can you make about the results?

For this experiment, I use all samples in training data not just the portion as previous experiment.

In following we can see what are the results.

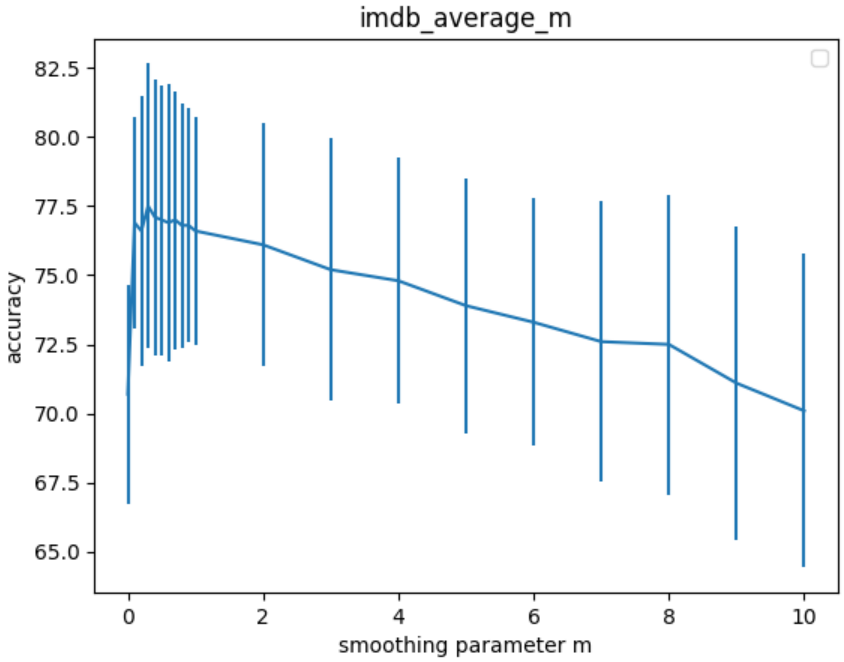

Figure 4: *cross validation accuracy and standard for imbd data set*

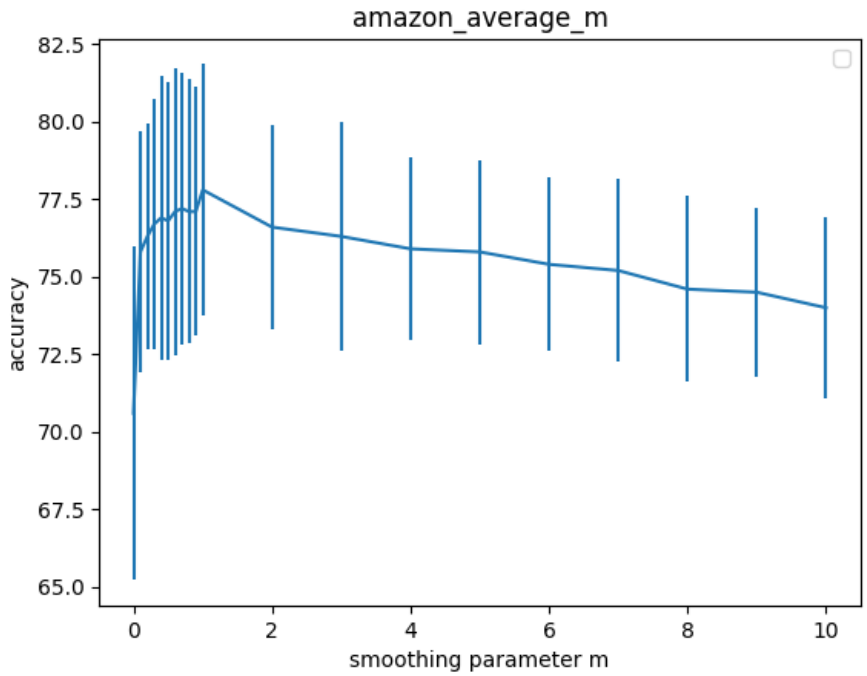

Figure 5*: cross validation accuracy and standard for amazon data set*

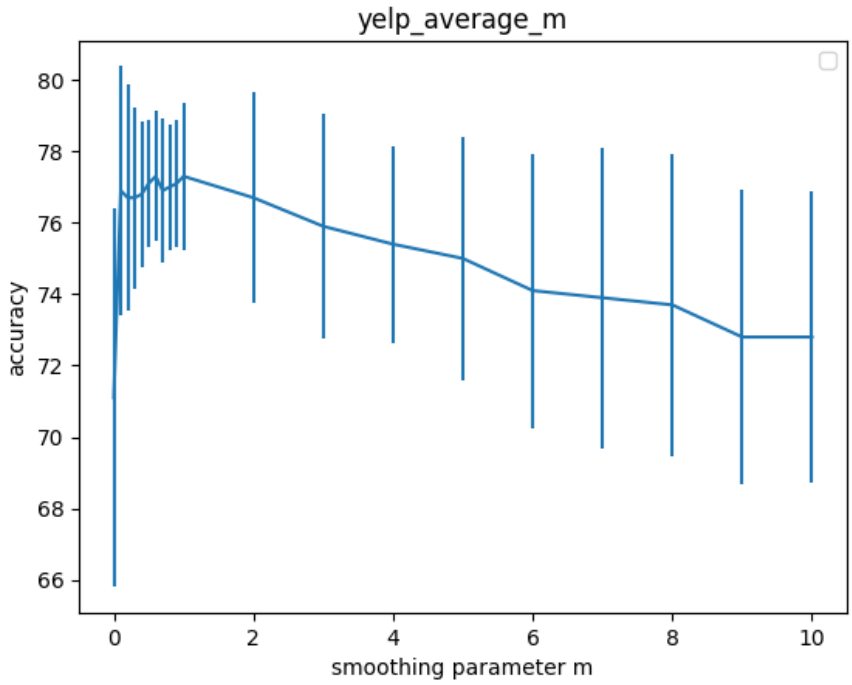

***Figure 6**: cross validation accuracy and standard for yelp data set*

the MAP estimate of p(w|c) is: (#(w∧c)+m) / (#(c)+mV)

where V is the vocabulary size, m is smoothing parameter an, #(w∧c) is the number of word tokens in examples of class c that are the word w and #(c) is the number of word tokens in examples of class c.

As we see in the plots, when the m increases, the MAP probability decreases. It moves toward 1/V. We can see it in the equation. Therefore, the m parameter more than zero and less or equal to 1 would be a good option based on the equation and also the results we got.

Part3: I preprocessed datasets and did all part1 and 2 again

I remove the punctuations from the datasets. So, this time for example “word,” and “word” would be the same and has the same probability and it is more meaningful.

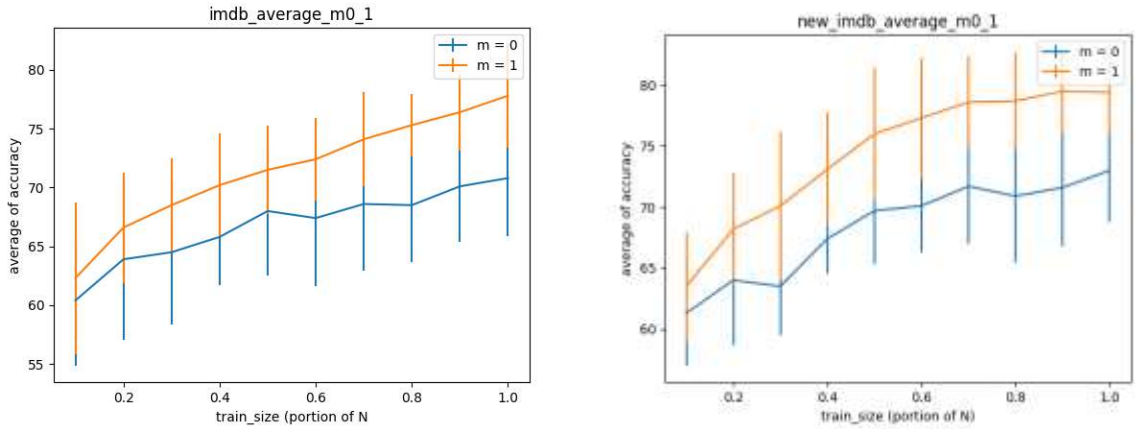

***Figure 7**: averages of the accuracy and standard deviations (as error bars) as a function of train set size for imdb data set (right one with pre-processed data)*

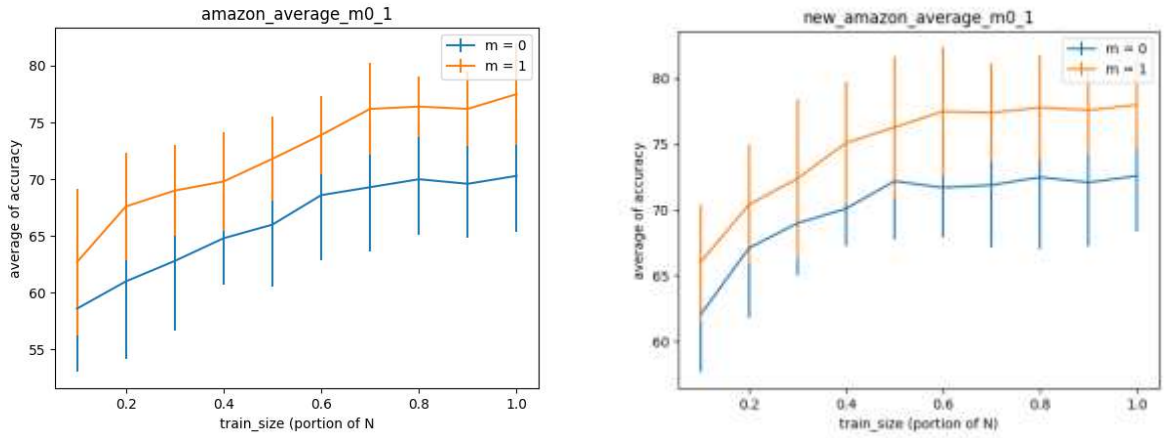

***Figure 8**: averages of the accuracy and standard deviations (as error bars) as a function of train set size for amazon data set (right one with pre-processed data)*

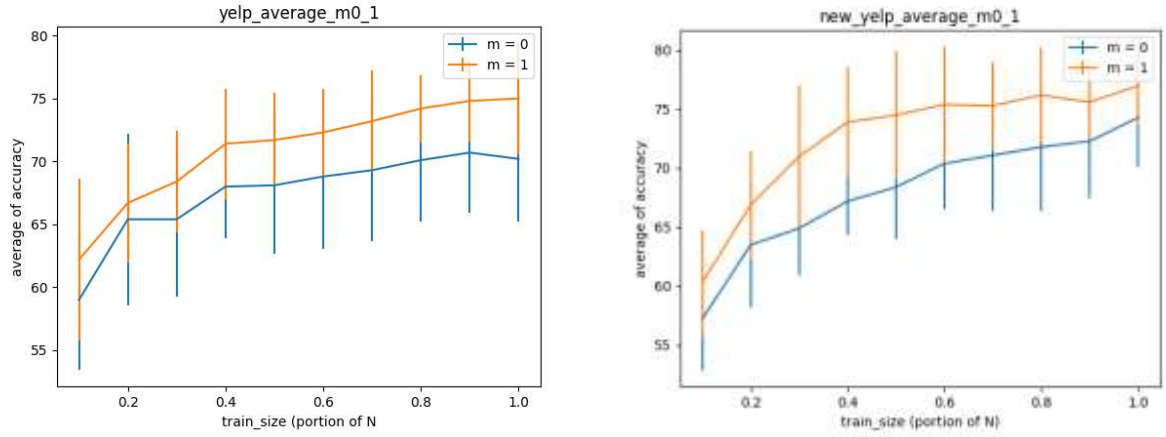

***Figure 9**: averages of the accuracy and standard deviations (as error bars) as a function of train set size for amazon data set (right one with pre-processed data)*

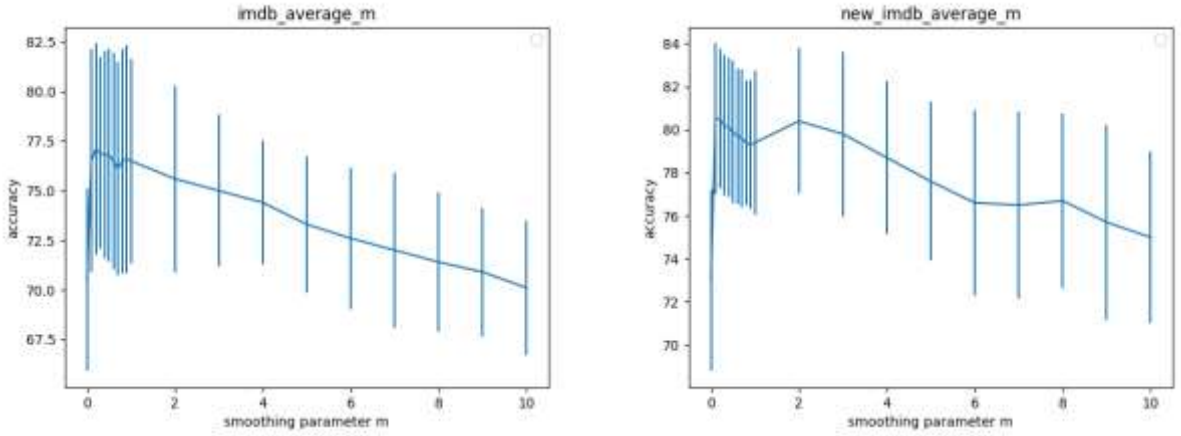

Figure 10: *cross validation accuracy and standard for imbd data set (right one with pre-processed data)*

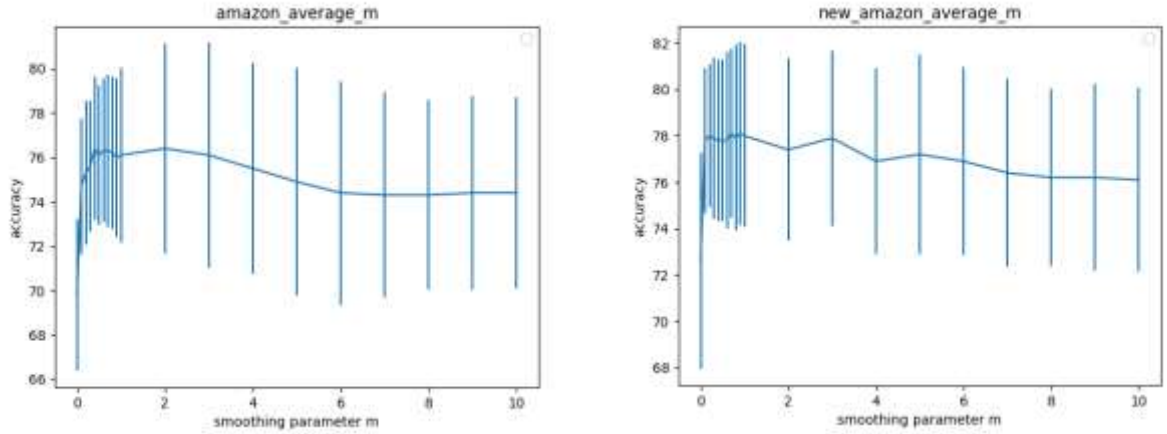

Figure 11*: cross validation accuracy and standard for amazon data set (right one with pre-processed data)*

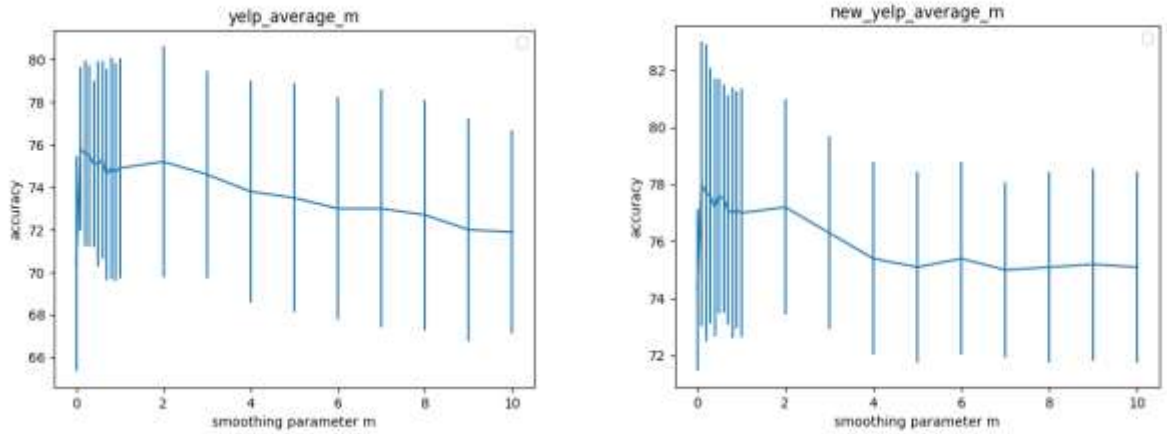

Figure 12*: cross validation accuracy and standard for yelp data set (right one with pre-processed data)*

As we see in the above figures, the average of accuracy become a little better by using preprocessed data.