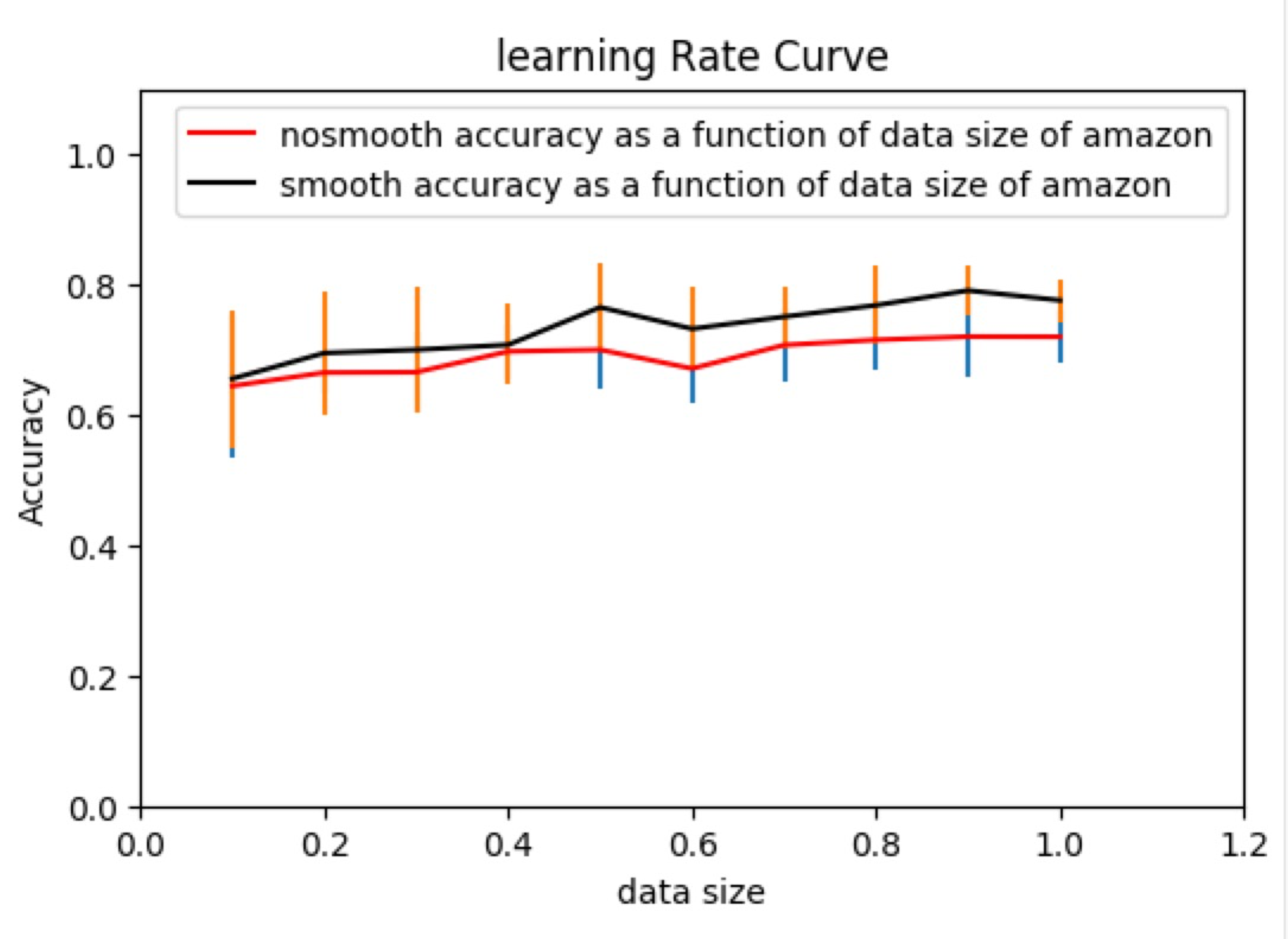
Project 2

Zheyi yi

1. The three graphs show the accuracy as a function of data size of no smooth and smooth(m = 1) for amazon, yelp, and imdb samples.



Fgiure 1

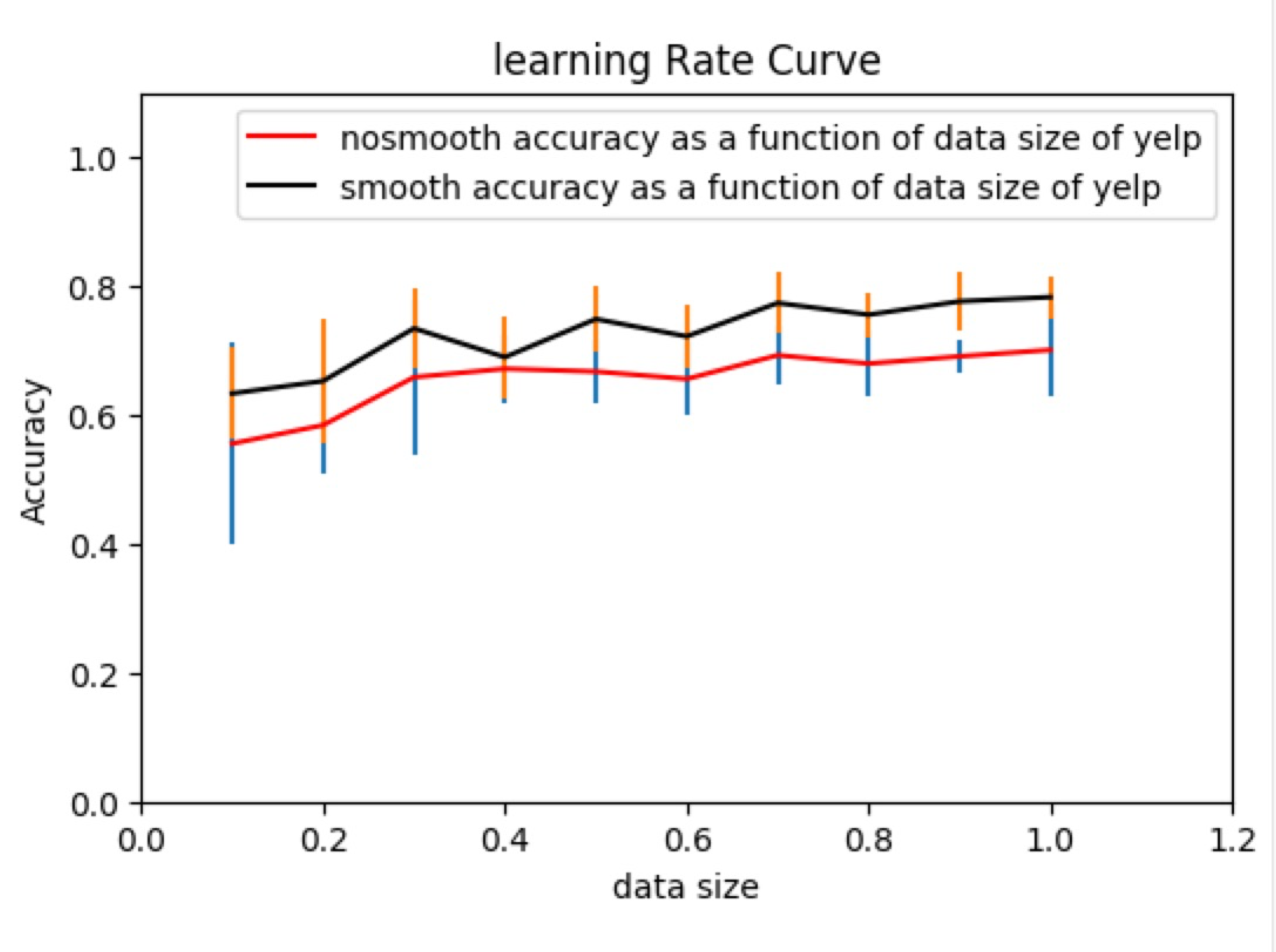


Figure 2

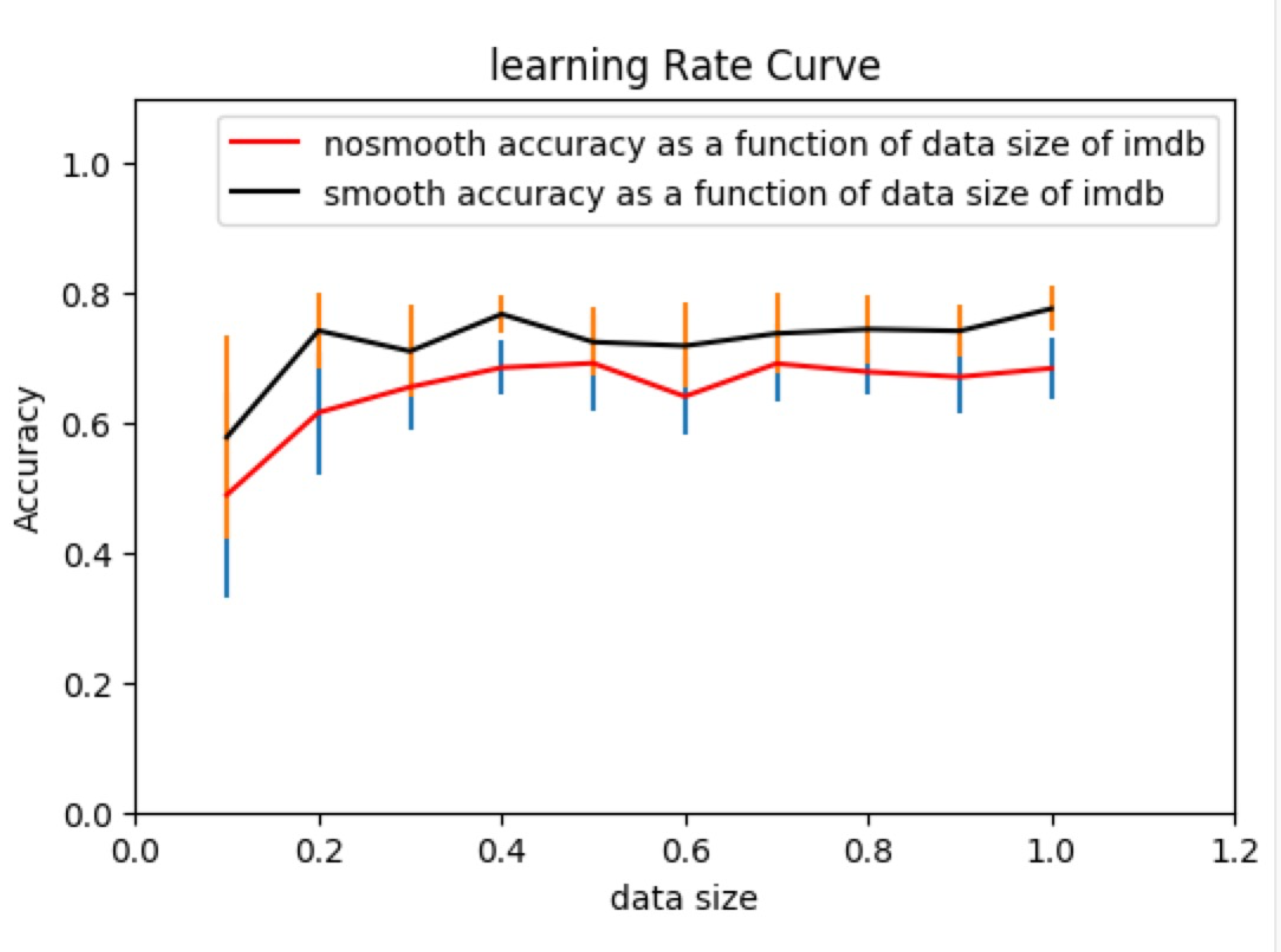


Figure 3

From the three graphs, we see as data size increase, their standard deviation decrease which means the result become more confident.

From the three graphs, we can see the no smoothing accuracy increase from 0.62 to 0.70 and smoothing accuracy from 0.64 to 0.78 for amazon.

no smoothing accuracy increase from 0.58 to 0.65 and smoothing accuracy from 0.64 to 0.77 for amazon.

no smoothing accuracy increase from 0.51 to 0.63 and smoothing accuracy from 0.58 to 0.76 for amazon.

From the three graphs, we can see the accuracy of smooth sample is better than no-smooth one for all three samples. The reason is that when testing data haven’t occur

in current vocabulary and exists in the opposite vocabulary, (for example, we calculate the sentence when we classify it as positive, but one word in testing data not in positive and in the negative vocabulary, so the probability of the sentence will be infinity, the model of bayes lose its function) we treat it negative infinity, so a lot of these sample haven’t been classified by native bayes appropriately, however, after smoothing, these kind of words have a small amount of probability and the sentence also follow the prediction by our trained bayes model.

2. we run stratified cross validation for Naïve Bayes with Smoothing parameter m from 0 to 0.9 and 1 to 10, and get two graph: cross validation performance as a function of the smoothing parameter.

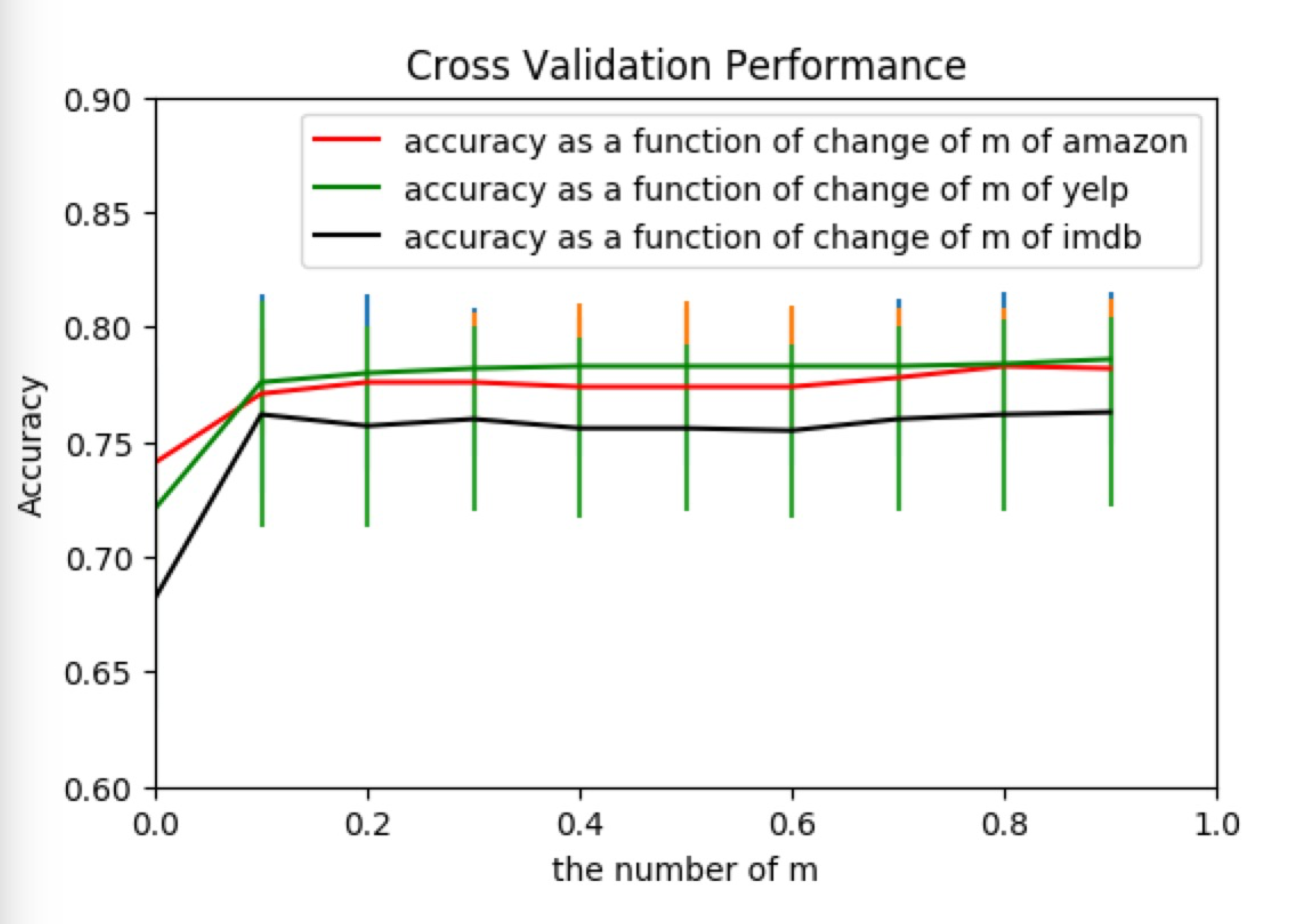


Figure 4

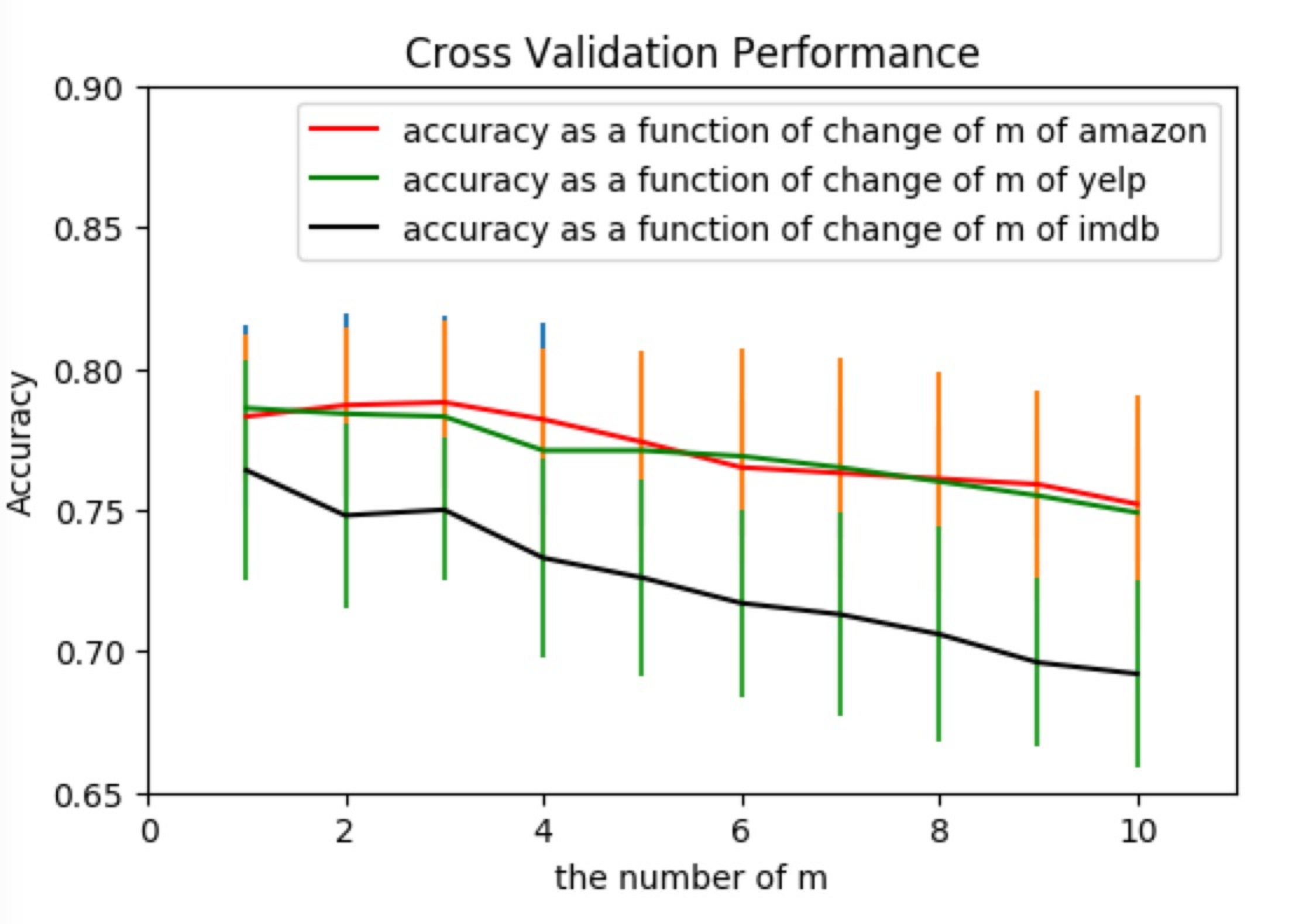


figure 5

From figure 4 and 5, we see the standard deviation does change obviously.

From the figure 4, we can see as m from 0 to 0.1, the accuracy for three samples increase a lot. The reasons is the same with m = 0 to m =1, which be explained above the first part. As m change from 0.1 to 1, the accuracy haven’t change obviously, however, when m change from 1 to 10, the accuracy of three samples decrease. The reason is that as the m increase, The times of the word appear in the vocabulary occupy less portion in the probability and m occupy more part in the probability. The calculation of the word times and set up model based on it which will lose effects because compare bigger m, smaller word number will paly less important role.