Comp 135 hw2

Fanying Ye

Due: October 13 2015

1 Introduction

In this assignment I conducted experiments to implement the Naive Bayes algorithm with smoothing and evaluate it using cross validation on text classification.

2 Plots and Conclusion

2.1 Learning curves for Naive Bayes without smoothing and with Laplace Smoothing

2.1.1 Accuracy means under different size of train set

Conclusion:
According to the plots of ibmmac and sport dataset, we can see that the accuracy averages under different dataset size are all much higher when we choose m=1 than choosing m=0, which proves the benefit of smoothing method. This happens because in Laplace's estimate, we pretend we saw every outcome once more than we actually did. Thus avoiding the case of totally abandon a class when we never see a word in that class.
Besides, we can conclude from the plot that the accuracy average increase with the increase of the train set size. In another word, the prediction improve with the increasing train set size.

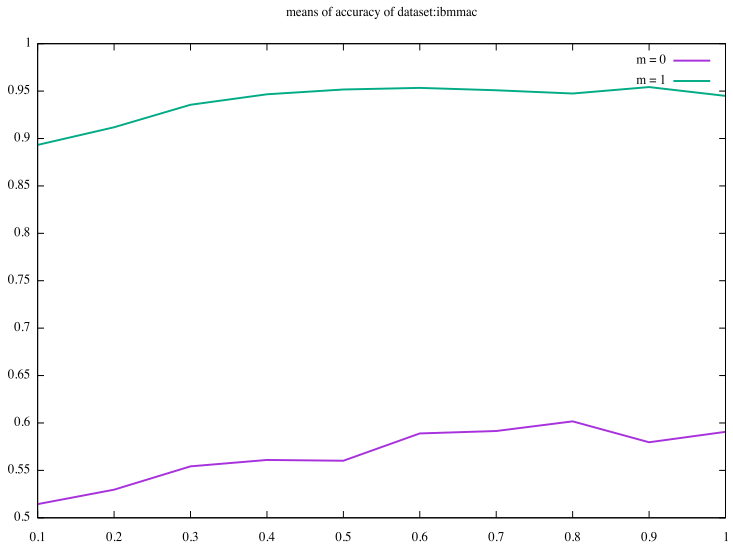

Figure 1: Learning curve with and without smoothing of dataset ibmmac

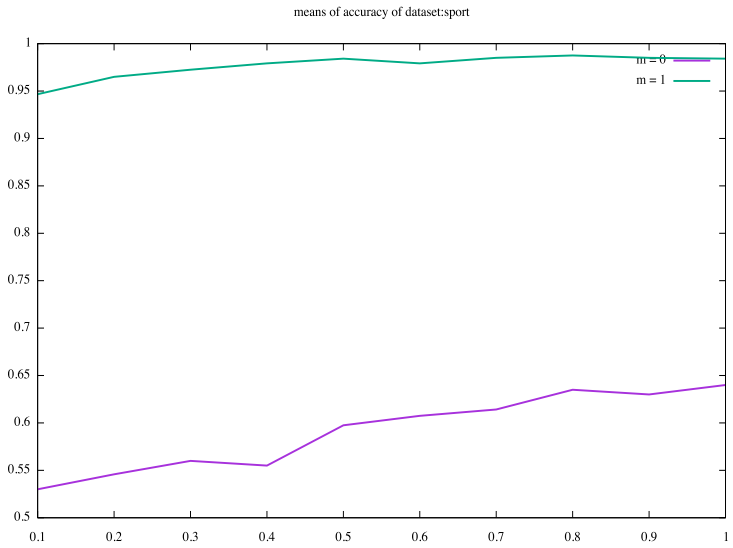

Figure 2: Learning curve with and without smoothing of dataset sport

2.1.2 Standard deviations under different size of train set

Conclusion:
When m=1, standard deviation of 10-fold cross validation tend to drop with increasing of train set size, which is to say, when we increase the train set size, the accuracy of choosing random training and testing set tend to be more closer to the mean. Thus, we can conclude that under higher train set size, we gain higher confidence of the accuracy measurement.
When m=0, standard deviation of 10-fold cross validation tend to be not very stable to show a trend. This makes sense because without smoothing, there are great chances that we choose a class randomly(when there is a word not in class Yes but in class No and another word not in class No but in class Yes). This kind of randomly guess makes the prediction unpredictable.

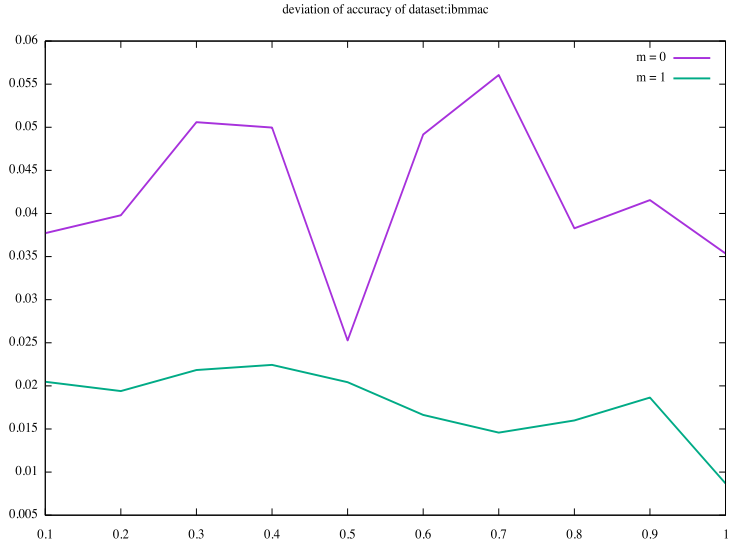

Figure 3: standard deviations under different data size with and without smoothing of dataset ibmmac

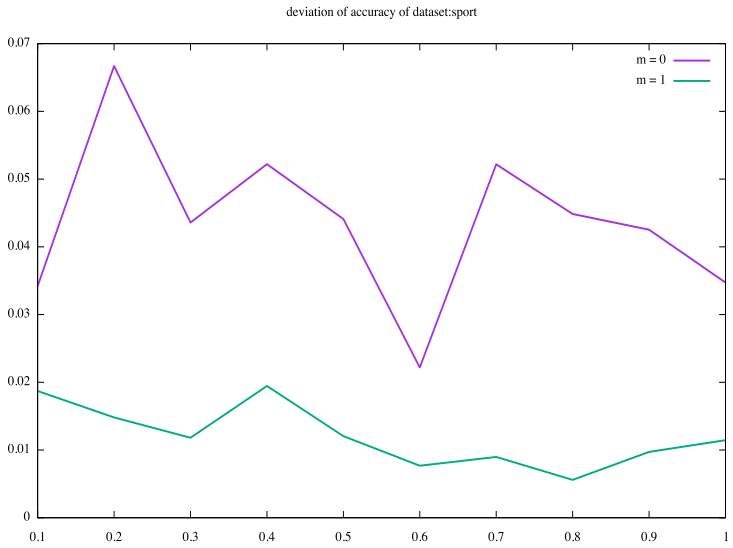

Figure 4: standard deviations under different data size with and without smoothing of dataset sport

2.2 Effect of smoothing parameter

Conclusion:
In both datase, when m is less than 1, the accuracy increase with the increasing of m. However, accuracy tends to decrease with the increasing of m when m is bigger than 1.
This happens because in smoothing method, we pretend we saw every outcome m times more than we actually did. As a result, if we choose a large m, it means that we pretend we saw every outcome many more times more than we actually did. In this case, the probability of prbability p(w/c) depends much more on m rather than number of words in(w and c), which leads to the decrease of total accuracy.

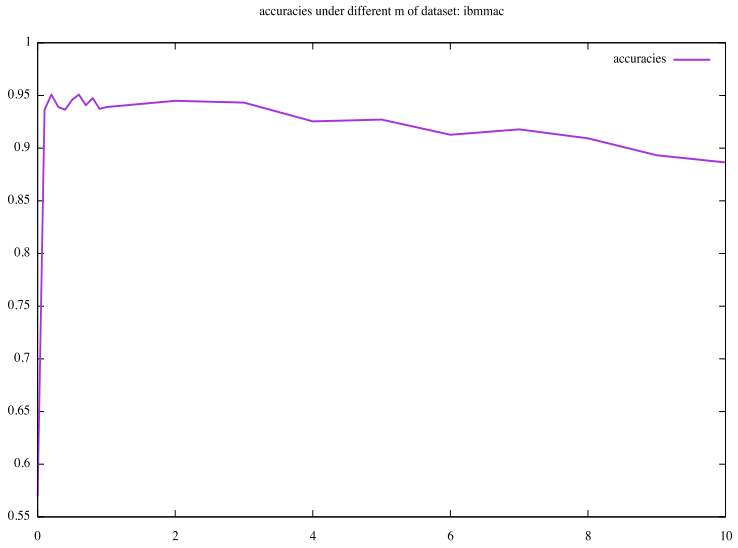

Figure 5: Cross validation performance under different m under dataset ibmmac

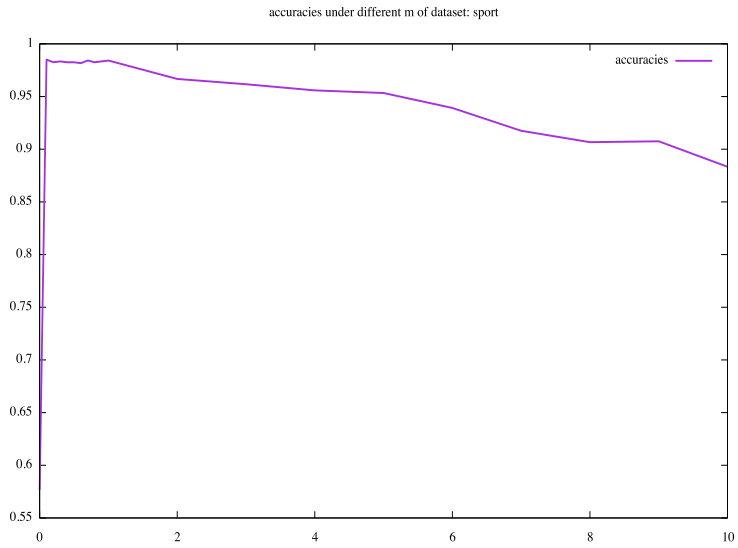

Figure 6: Cross validation performance under different m under dataset sport