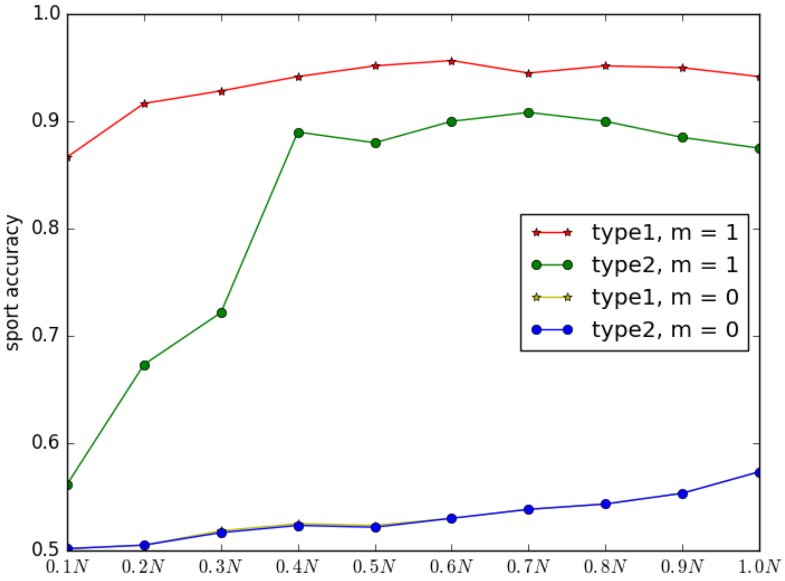
Comp135 project2 Report

By Beibei du
Due data: 10/19/2016

1. **Learning Curves for Naïve Bayes without smoothing and with Laplace smoothing**

   Without smoothing : m = 0   Laplace smoothing: m = 1

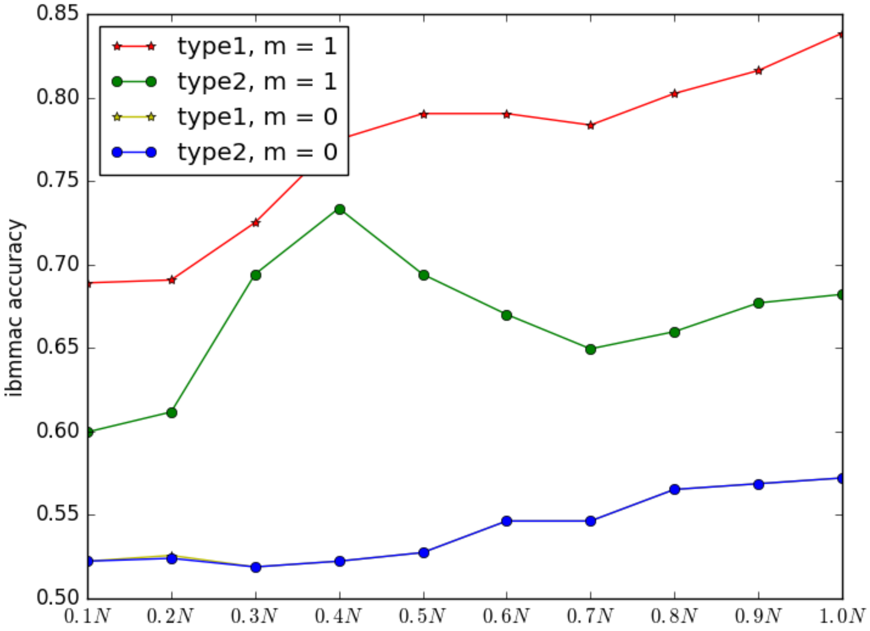
   - Sport data set



from this plot, we can find that when m = 0, type1 and type2 will almost match exactly, with the increase of train data set size, performance will become better for both type1 and type2, the lowest accuracy is about 50%, but highest accuracy is not higher than 60%.

With Laplace smoothing m = 1, type1 and type2 both have better performance than without smoothing. In big direction, type1 is much better than type2, even though type2 involves more features for every test file. Both type1 and type2 performance will become better with the increasing size of train data set.

- Ibm/mac data set

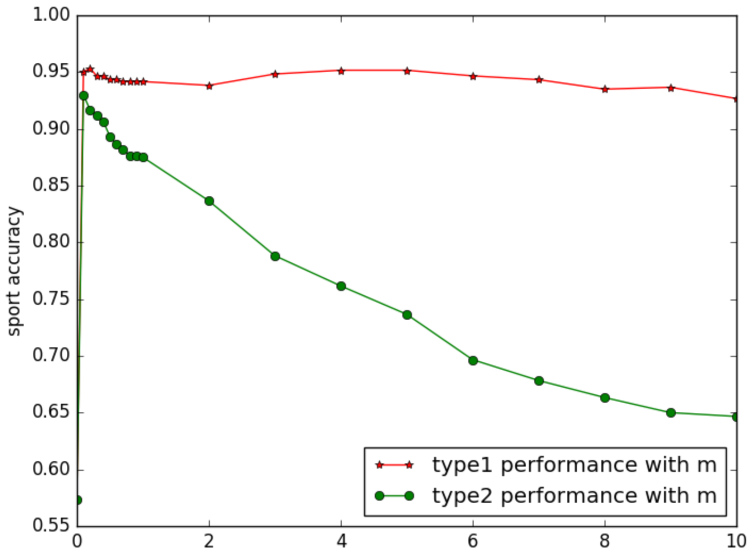


For Ibm/mac data set:
Without smoothing, type1 and type2 test accuracy are almost same, lowest accuracy is about 52%, highest accuracy is no more than 60%. With the increasing of train data size, the accuracy will also become better gradually.
With Laplace smoothing, both type1 and type2 perform better than without smoothing, and also type1 perform better than type2 for every train data set size. With the increase of train data set size, type1 will increase in big direction, but type2 first increase then decrease then increase, and when train data set size = 0.4N, type2 has highest accuracy.

2. Performance with m

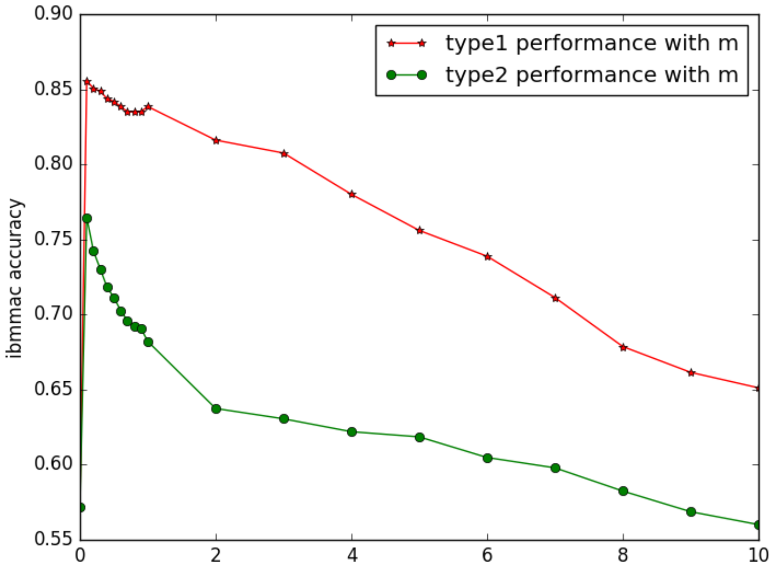
m = (0, 0.1, …, 0.9, 1.0, 2.0, …, 10.0)

- ➢ Sport data set



From this plot, we can find that for typ1 performance better than type2 except when m = 0. When m > 0, type1 accuracy doesn't have big change, just a little vibration around 95%, but type2 has highest accuracy around 93% when m = 0.1, then decrease gradually with the increase of m value, it has dropped to 65% when m = 10.

- ➢ Ibm/mac data set



From this plot, we can find that when m = 0, type1 and type2 have same test accuracy, when m = 0.1, both type1 and type2 have highest test accuracy, type1 with highest test accuracy around 85% and type2 with highest test accuracy around 76%. With the increase of m value, both type1 and type2 test accuracy decrease gradually, and when m = 10, both of them have lowest test accuracy , type1 has lowest accuracy around 65% and type2 has lowest test accuracy around 56%. For any value of m, type1 performs better than type2.