Naive Bayes Classification

Alex McCarthy

September 26, 2017

1 Introduction

In this paper I am showing an example of naive bayes classification in the form of a python script written by me versus vowpal wabbit's method of classification. The dataset used is 150 seperate iris type flowers seperated into 3 different species- setosa, versicolor, and virginica. The naive bayes algorithm involves first training the classifier and then testing it on a test set derived from the training set to tune it. The algorithm can then be used on the actual test set. The function of the naive bayes is to give the probability of some class given a set of features and it finds this by multiplying the probability of the set of features given the learned class by the probability of the class itself. In this case the probability of the class is 1/3.

2 Results

My python classifier performed well with an 84.2 percent accuracy. Vowpal Wabbit's classifier was 70 percent accurate. I split my data into three different sets. The first set was training data comprised of 120 entries. I also had a practice set and a test set with 10 entries and 20 entries respectively. When training on the practice set, my naive bayes algorithm was 80 percent accurate. I did not expect this result, I expected vowpal wabbit to perform better than my algorithm because vowpal wabbit seems to be using an algorithm that can be more specifically tuned to the data set than the naive bayes. I handled continuous values in the data set by treating them as discrete values.

3 Conclusion

I think my classifier performed about as well as I expected it to. I figured there would be some overlap between flower species measurements, which there was, and this made the naive bayes less accurate. The reason for this is simple; one flower can have measurements that make it seem like another flower if that is all you have to judge the flower on. I think with more specific features this would happen less often. For example maybe if stripes or colors were included and we could classify based on those features along with the measurements then the bayes classifier would be even more accurate.