Shuffling Features to Improve Performance of Naive Bayes

Marcus Twichel

Editor:

Abstract

This paper describes an experiment into whether or not shuffling 10% of features in a dataset before training a Naive Bayes learner has a negative effect on model accuracy. I predicted that it would have a negative effect on accuracy due to adding more noise into the data. I tested this by implementing a Naive Bayes algorithm on 5 different data sets, and then tested the accuracy (with 5X2 cross validation) of the models with shuffling and without shuffling. In the end, I found no correlation between shuffling in general, although some data sets were positively impacted by shuffling.

1. Problem

I wanted to find if shuffling 10% of the features in a dataset would yield more accurate Naive Bayes learning models. My hypothesis is that it will yield a decease in accuracy as all we are doing is adding noise into the system and eliminating features from our model's consideration. I believe this will lead to less accurate models.

2. Algorithm Implemented

I implemented a variation on the algorithm known as Naive Bayes. It is based on Bayesian decision making, which in its most basic formulation, says the optimal strategy is to look at past take evidence e, find the probability of each choice, and then choose the choice that has the highest probability.

The algorithm does essentially this. You find the probability of each value for each feature for each class. In notation:

Def C: all the classes in a dataset
Def F(c): all the features present in class C in a dataset
Def V(f — c): all the unique values present in feature f given c in a dataset

$$\forall c \in C : \forall f \in F(c) : P(v \in V(f)|C)$$

When you make predictions, you simply multiply the probability of finding the inputted value for each feature, and repeat this for every class. You then multiply the probabilities by the probability of that class in the dataset to weight by class size. You then compare the computed probabilities: whatever probability is highest is the predicted class.

3. Results

The results can be summarized by this table:

Dataset	Without Shuffling	With Shuffling	Difference (without - with)
Iris	0.882	0.889	-0.007
House Votes	0.903	0.900	0.003
Breast Cancer	0.960	0.974	-0.014
Glass	0.148	0.170	-0.022
Soybean	0.862	0.932	-0.07

As you can see, there is most often a slight increase in performance when shuffling 10% of the features. This is most apparent in the Soybean dataset, where performance increases by 7% when shuffling 10% of the features.

4. Conclusion

After doing statistical analysis of a two value t test, (with shuffling as the explanatory variable and accuracy as the response variable) I found a p-value of 0.988, which suggests no statistical correlation between shuffling and the accuracy of a Naive Bayes learner. While I was surprised to see gains at from shuffling, from my experiments, I cannot conclude that there is any link between shuffling 10% of features in a dataset and a change in accuracy of that model.

References

[1] Dr. WIlliam H. Wolberg, Olvi Mangasarian, David W. Aha *Breast Cancer Wisconsin (Original) Data Set.* University of Wisconsin Hospitals, Madison, Wisconsin, 1992.

[2] B. German, Vina Spiehler *Glass Identification Data Set.* Central Research Establishment, Home Office Forensic Science Service Aldermaston, Reading, Berkshire 1987.

[3] R.A. Fisher, Michael Marshall *Iris Data Set.* 1988.

[4] Jeff Schlimmer *Congressional Voting Records Data Set.* UCongressional Quarterly Almanac, 98th Congress, 2nd session 1984, 1987.

[5] Michalski,R.S., Doug Fisher *Soybean (Small) Data Set.* International Journal of Policy Analysis and Information Systems, 1980.

[6] Travis E, Oliphant *A guide to NumPy.* Trelgol Publishing, 2006.

[7] Wes McKinney *Data Structures for Statistical Computing in Python.* Proceedings of the 9th Python in Science Conference, 51-56, 2010

[8] Kevin P. Murphy *Naive Bayes classifiers.* 2006.