Scrambled Data and the Naive Bayesian Model

Logan Ladd logan.ladd@ecat1.montana.edu
Gianforte School of Computing
Montana State University
Bozeman, MT 59717-3880, USA

Asher Worley asher.worley@ecat1.montana.edu
Gianforte School of Computing
Montana State University
Bozeman, MT 59717-3880, USA

Editor: Logan Ladd

Abstract

This paper describes our usage of the Naive Bayesian model to learn five different data sets, with the goal of determining whether scrambling the data changes how accurate the model is. The data sets came in a variety of forms, some boolean, others discrete, and still others continuous. We first preprocessed the data sets to get rid of the unknown and corrupted values. We then discretized the continuous data using a binning method. Following that, we ran the data through the Naive Bayesian model with a ten-fold cross validation method, and then tuned our hyperparameters. Using this as our control, we then scrambled the data and ran the experiment again. We found that, on average, the unscrambled data performed 0.94% better than the scrambled data, and that the unscrambled data had, on average, a mean squared error 0.36 higher than the scrambled data. However, the naturally discrete data sets performed much better than the naturally continuous data sets when comparing the scrambled data to the unscrambled data. In fact, they performed better by a factor of ten. This implies that noise has a significant impact to naturally continuous data sets when using the Naive Bayesian model.

1. Introduction

The Naive Bayesian model is, as the name suggests, a naive yet useful model in machine learning. The model works by calculating probability vectors based on the frequency that features appear in the data. This generally works well for data sets that require little to no nuance in analyzing, and requires data sets that are fairly accurate in their classifications. However, one problem that modern data sets often have is misclassified and noisy data. This has the potential to undermine the usefulness of the Naive Bayesian model. Our goal is to determine whether noisy data can be used to the same effect as clean data.

We hypothesise that there is little impact to the Naive Bayesian model when up to 10% of the features are scrambled.

By scrambled, we mean that the feature values have been altered so that they no longer reflect reality.

1.1 Data Sets

We used five data sets in our experiment. The data sets had between 47 and 699 data points, with a maximum of 17 missing attribute values. Of these five data sets, two of them contain continuous data.

Data Set	# of Features	# of Missing Attribute Values	Continuous
breast-cancer-wisconsin[5]	699	16	No
glass[2]	214	0	Yes
house-votes-84[4]	435	17	No
iris[1]	150	0	Yes
soybean-small[3]	49	0	No

We first preprocessed the data in order to fill in the missing attributes. Each missing attribute was filled in by a fixed value of 2. The continuous data sets were then discretized by splitting the values into five equal bins.

1.2 The Model

The Naive Bayesian model works by calculating probability vectors. First, for the training set we need to count how many of each class we have, and divide that number by the total number of data points in the training data. This vector is $\vec{Q}$.

$$\vec{Q} = \frac{c_i}{N}$$

Where c_i is the number of data points that have the i'th class, and N is the number of data points in the training set. Then for each possible attribute value and for each class we count the total number of attribute values that belong to that class, add one and divide by the total number of data points in that class, plus the number of possible attribute values. This gives us matrix F.

$$F_{ij} = \frac{x_{jc} + 1}{c_i + d}$$

Where x_{jc} is the number of data points that have attribute value x_j and are of the i'th class, and d is the number of possible attribute values. Then, in order to classify a new data point, we calculate a new matrix F, where x_j is the j'th feature of our new data point. We then multiply F by Q using matrix multiplication. This gives us a probability vector $\vec{P}$.

$$\vec{P} = F \cdot \vec{Q}$$

In order to classify the data point, we note the maximal value in the vector $\vec{P}$, P_i. This tells us that the most probable classification for the data point is the i'th class.

1.3 Hyperparameter Tuning

The only hyperparameter that required tuning was the number of bins to use when discretizing the continuous data. Since our hyperparameter space was only one-dimensional, it was fairly easy to tune. We used a brute force mechanism, testing everywhere from two to ten bins. We found that using more than five bins provided little improvement to the model, so we used five in our final experiment.

Data Set	2-Bin Accuracy	2-Bin Mean Squared Difference
glass[2]	69.1	5.0
iris[1]	70.7	7.8
	10-Bin Accuracy	10-Bin Mean Squared Difference
glass[2]	20.7	45.2
iris[1]	92.7	0.8

1.4 Results

After performing an experiment on both the scrambled data and the unscrambled data, we found that the model performed about the same on both sets.

Data Set	Average Accuracy	Mean Squared Difference
breast-cancer-wisconsin[5]	93.6	3.8
breast-cancer-wisconsin-scrambled	93.9	3.3
glass[2]	80.8	2.9
glass scrambled	78.0	3.2
house-votes-84[4]	90.1	6.1
house-votes-84-scrambled	89.9	4.2
iris[1]	92.7	0.7
iris-scrambled	90.7	1.0
soybean-small[3]	79.5	0.9
soybean-small-scrambled	79.5	0.9

We found that the unscrambled data performed 0.94% better on average than the scrambled data. Interestingly, the mean squared difference was lower by 0.36 on average for the scrambled data than for the unscrambled data. This implies that the model was much more confident with the unscrambled data, even when it was incorrect. Introducing noise into the equation likely made outliers in the data sets more common, thus lowering the confidence levels of the model. Also of note is the fact that both of the naturally continuous data sets had a lower mean squared difference before scrambling, whereas the naturally discrete data sets had a higher mean squared difference before scrambling.

It appears that the data sets that were impacted the most by the scrambling were the data sets that were naturally continuous. These data sets, *glass* and *iris*, performed 2.8% and 2.7% better than their scrambled counterparts respectively. The average of the difference between the scrambled and the unscrambled naturally discrete data sets is 0.17%.

The fact that this is ten times ten times lower than the average of the naturally continuous data sets suggests that noise hinders the Naive Bayesian model on naturally continuous data much more than on naturally discrete data. This is supported by the differences noted above in the mean squared differences. Unfortunately, more data is required to conclusively accept or reject the hypothesis, however this initial data suggests that the addition of noise to the data doesn't have significant impact to the Naive Bayesian model.

2. Summary

The noise introduced very little variance in the accuracy of the discrete data sets, however the continuous data sets were effected much more by the introduction of the noise. This suggests that the process of discretization, specifically the binning method that we used, does not capture the nuanced information that is present within continuous data sets. Among the naturally discrete data sets, there was very little difference in accuracy between the scrambled and unscrambled data, less than 0.2%. This implies that noise mitigation may not be very important when dealing with data sets with categorical or discrete features. Further experimentation is required to confirm this. One such experiment should include adding more noise and testing to find the rate at which the model deteriorates.

There was little to no evidence that the hypothesis is false, however the difference in accuracy isn't significant enough to accept or reject the hypothesis. More experimentation with varying data sets is required to provide more conclusive evidence.

References

[1] R.A. Fisher. *UCI Iris Data Set*. 1988. URL: `https://archive.ics.uci.edu/ml/datasets/Iris`.

[2] B. German. *UCI Glass Identification Data Set*. 1987. URL: `https://archive.ics.uci.edu/ml/datasets/Glass+Identification`.

[3] R.S. Michalski. *UCI Learning by being told and learning from examples: an experimental comparison of the two methodes of knowledge acquisition in the context of developing an expert system for soybean desease diagnoiss*. 1980.

[4] *UCI Congressional Voting records*. 1984. URL: `https://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records`.

[5] Dr. WIlliam H. Wolberg. *UCI Wisconsin Breast Cancer Database*. 1991.