Journal of Machine Learning Research 1 (2019) INSERT PAGE STUFF Submitted 9/19; Published DATE PUBLISHED

Learning with Naive Bayes

Andrew Kirby ANDY-KIRBY@LIVE.COM
Kevin Browder BROWDERKEVIN54@GMAIL.COM
Nathan Stouffer NATHANSTOUFFER1999@GMAIL.COM
Eric Kempf ERICKEMPF123@GMAIL.COM

Editor:

Abstract

Keywords:

1. Introduction

Naive Bayes is a simple algorithm and a nice introduction into machine learning. It works surprisingly well at classification learning. A classification algorithm takes a given set of classes and attributes, and determines the probability of that a given class has said attributes. The reason Naive Bayes is so simple is due to the very powerful assumption that our attribute set is conditionally independent. This assumption, although probably untrue for most sets, is what allows the Naive Bayes classifier to do what it does, and do it relatively well for how simple it is. In this paper we go over our process of training the Naive Bayes classifier on various data sets, and the results of each trial.

The paper is organized as follows. Section 2 goes over the problem we had to solve. Section 3 takes a look at the actual classification algorithm. Section 4 goes over our design process and each step we took to get our end results.

2. Problem Statement

The problem is a classification problem. Given an input file that contains examples (each example consists of a list of attributes and an associated classification), the task is to implement a learning algorithm that is trained to classify examples. The algorithm is called Naive Bayes. The performance of the learning algorithm will be evaluated by two metrics and 10-fold cross validation. Specifically, the metrics are accuracy and mean squared error.

2.1 Variables

The independent variable is whether the data is scrambled or not. Scrambling is described as follows. First, 10% of the attributes in a given data set are randomly selected. Then, within each attribute, the values are randomly swapped between examples. Now the data set is scrambled. The dependent variable is how well the algorithm performs.

2.2 Hypothesis

The hypothesis is that scrambling a given data set will not significantly change the performance of the Naive Bayes Algorithm. In essence, scrambling renders 10% of attributes useless. Any pattern that existed before scrambling is no longer discernible within those attributes. However, there remains 90% of the data that persists with the original pattern. Naive Bayes chooses a classification based on the relative probability that a given example is in a class. The order of the relative probabilities will stay constant even when the original pattern is lost in 10% of the attributes.

3. Algorithm

-It be what it do

4. Experimental Design

4.1 Set Up

4.2 Tuning

Add stuff about bin size, what else?
Experimenting w/ using different attributes

4.3 Final Parameters

-Bin size
-Number of attributes used
-Which attributes used?

5. Results

6. Summary

References