UASimpleClassifier: An Implementation of Naive Bayes's Classifier, Probability density functions, and Cumulative distribution function

Ivan Welborn
Computer and Information Sciences Department
University of Arkansas – Fort Smith

February 23, 2023

Abstract

The UASimpleClassifier algorithm takes in a Data Set which will result in a trained model that will be able to predict what class the features belong in with an accuracy of 80.50%

1 Introduction

The UASimpleClassifer algorithm will take in a file with five features, with those features it will calculate the probabilities of each thing happening given each possible class. It will then construct a model that will be able to predict the class it belongs in with an accuracy of 80.5%

2 Background

Naive Bayes is based on conditional probability, and following from Bayes theorem, for a document d and a class c, it is given as $\mathrm{P}(\mathrm{c}\|d) = \frac{P(d\|c)P(c)}{P(d)}$ [1]
Naive Bayes has been found to be very useful in the field of medical science for diagnosing heart patients due to it not being a hard model to build, and not having a complicated iterative parameter estimation. [2]

3 Specification

The main technique I use throughout this problem is the Naive Bayes formula, and Probability density functions. The formula's for each

Naive Bayes

$$\hat{c} = \mathrm{argmax}_{j\ldots k} \frac{Pr(C_j)\Pi_{i=1}^{n} Pr(f_i \mid C_j)}{\sum_{k}^{|C|} Pr(C_k)\Pi_{i=1}^{n} Pr(f_i \mid C_k)}$$

Probability Density Function

$$P(x_i \mid y) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

4 Implementation

My main algorithms are the probability calculations I use, which is just taking the amount of features given a zero or a one class, and dividing that by the total amount of zero or one classes.

Another algorithm that is used extensively is the Naive Bayes Algorithm, this is used to calculate the probability of each outcome, we use it to guess which class each set of data will belong in.

While the Naive Bayes algorithm is running, the program also calls the CDF Function which will output the probability that a given number is less than our inputted number.

5 Evaluation

The data set that was used in my program was severely biased towards a zero class, which skewed my results heavily making most of the model's guesses zeros. My algorithm worked for 80.50% of the data, however because the learning data set was so biased, it made it nearly impossible for my model to guess a higher percentage

6 Conclusions

Data being biased like it was in our data set will severely impact our models efficiency, it is important that the data we take in is equal and unbiased towards any side. Naive Bayes is good with data that is independent however if the data was no longer independent it would struggle to be as accurate as other algorithms.

References

[1] Sarkar, Subhajit Dey, et al. “A Novel Feature Selection Technique for Text Classification Using Naïve Bayes.” ISRN Otolaryngology, Jan. 2014, pp. 1–10. EBSCOhost, https://doi.org/10.1155/2014/717092.

[2] ”Heart Diseases Detection Using Naive Bayes Algorithm” ISSN 2248-7968, 9, September 2015. https://www.ijiset.com/vol2/v2s9/IJISET_V2_I9_54.pdf