Introduction to Machine Learning, Programming Assignment 1

Comparing the Performance of two Basic Classification Algorithms

Julie P. Garcia

Johns Hopkins University

Abstract

In this paper, we will compare the performance of two basic classification algorithms, Naïve Bayes and the Winnow-2 algorithm. The Winnow-2 algorithm is a classification machine learning algorithm for use on linearly separable data. Naïve Bayes is also a classifier algorithm that uses probabilistic models to learn on a set of data. Both algorithms were implemented in Python and run on three publicly available datasets that are commonly used for learning.[1,2,3] The models were tuned for best performance and, in general, the results show that Naïve Bayes performs better than the Winnow-2 Algorithm on all three datasets.

Problem Statement

In this paper, I compare the Winnow-2 Classifier algorithm to the Naïve Bayes classifier algorithm. Because it uses a probabilistic model, I expect Naïve Bayes to perform better than the Winnow-2 Algotihm

Algorithms Implemented

Both algorithms are simple classifier algorithms, however, Naïve Bayes uses a probabilistic model and Winnow-2 is a teaching algorithm that uses a threshold theta and a set of weights for each feature to predict the class.

Winnow-2 Algorithm

The winnow-2 algorithm initializes an array of weights, whose length is the number of features in the dataset and each is initialized to zero to start. A threshold, θ, is set based on analysis of the data. A second hyperparameter, α, and is used to promote or demote the weights based on a correct or incorrect answer, respectively.[4] The algorithms is then implemented as follows:

1. Each sample in the dataset is fed into the algorithm.
2. A prediction is made by comparing the sum of the weights multiplied by the feature value with θ.
3. The prediction is compared to the actual classifier. If correct ,the value is promoted by multiplying by α. If incorrect, the value is demoted by dividing by α. The new weights are returned and used in the next evaluation.

Naïve Bayes Algorithm

The Naïve Bayes algorithm uses a probabilistic model to classify samples of data.[5] The algorithm was implemented as follows:

1. The model is constructed by finding P(C=0) and P(C=1) in dataset, where C is the classifier, and then P(F=1|C=0), P(F=0|C=0), P(F=1|C=1) and P(F=0|C=1) for each sample in the dataset.
2. Predictions are made my by iterating through each sample in the dataset and multiplying P(C=0) by the probabilities for each feature in the sample to get the total probability for that combination, given C=0 and given C=1. If the P(C=0) > P(C=1) the prediction is 0, otherwise it is 1.

Experimental Approach

Both algorithms were run on three datasets. For each of the three the datasets, the Winnow-2 algorithm hyperparameters were tuned by first created boxplots for the classifiers and the sums of each row to see if there was a clear threshold. The data classifiers were one hot encoded and missing data removed. The hyperparameter θ was tuned based on these charts. The hyperparameter α was tuned by starting with 2 and changing it until the best results were obtained.

The Naïve Bayes algorithm was run by one hot encoding the datasets to meaningful values based on the boxplots created and run for each dataset. All percentages recorded below, were based on 10 runs of splitting test and training data randomly, and different each time. The average of each percent accuracy was recorded below.

Results

The Winnow-2 algorithm performed better on the breast dataset with 96.9% accuracy on the test set, while Naïve Bayes resulted in 95.2% accuracy for the breast test data. For the iris dataset, the Winnow-2 algorithm only resulted in 74% accuracy on the test data, while Naïve Bayes gave 90% accuracy. This dataset was made into a two class problem by only evaluating "Iris-virginica" or "not Iris-virginica". For the votes dataset, Winnow-2 performed at 89% accuracy, while Naïve Bayes performed at 91%.

Summary

Neither of these algorithms performed that well, but the Naïve Bayes algorithm performed better for two out of the three datasets. The Winnow-2 algorithm performed extremely poorly on the iris dataset as the data does not seem to be linearly classifiable.

References

1. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original% 29

2. https://archive.ics.uci.edu/ml/datasets/Iris

3. https://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records

4. http://www.cs.tau.ac.il/~mansour/ml-course-10/scribe4.pdf

5. https://towardsdatascience.com/introduction-to-naive-bayes-classification-4cffabb1ae54