A Brief Comparison of Machine Learning Algorithms

Joel Villanueva

University of California, San Diego

COGS 118A: Supervised Machine Learning Algorithms

Professor Tu

16 December 2023

Abstract

With machine learning algorithms growing exponentially in professional use, it is key to understand the benefits and shortcomings of some of the most popular methods. Support vector machines, logistic regression classification and random forest algorithms are all widely used, so a comparison among their performances on datasets would prove to be useful. Accuracies for each algorithm will be explored in datasets of varying sizes in order to successfully highlight their individual strengths and weaknesses.

Introduction

With an ever increasing interest in machine learning throughout the tech industry, an overview and comparison of a few of the more popular algorithms would prove to be useful. Of the more commonly used, support vector machines, logistic regression classification and random forests were chosen to be observed for this research. These three algorithms would be tested on three datasets of different sizes, all of which were taken from the University of Irvine Machine Learning Repository. The topics of each dataset vary greatly, allowing for the features to range in data types, which will help to discern the benefits of using one algorithm over another. The datasets will look at wart treatments using immunotherapy, classification of raisins and cooling load requirements of buildings.

When examining the algorithms' performance on the datasets, their prediction accuracy and resulting error will be the main metric that they are judged on. The accuracy and error are inverse measurements and we are looking to minimize the error while maximizing the accuracy. For the support vector machines, a linear support vector classification is used in addition to a hinge loss function and a L2 normalization penalty. For the logistic regression classifier, a "liblinear" solver is used to train the data in addition to a balanced class weighting. The maximum depth of the random forest classifier is limited to two in order to prevent overuse of memory on the larger datasets.

After training the three algorithms on the three datasets, the random forest algorithm performed the best, followed by the support vector machine and finally the logistic regression classifier.

Data

Three datasets revolving around immunotherapy, classification of raisins and energy efficiency were used to examine the performance of the aforementioned algorithms. The immunotherapy dataset was the smallest, with only 90 instances and 8 features. The next largest dataset was the energy efficiency set that discussed the cooling load requirement, with its 768 instances and 8 features. The largest dataset was the one examining raisins, with 900 instances and 8 features. The testing variable of each dataset was turned into a binary variable of either -1 or 1 for ease of use of the algorithms.

Methodology

The datasets were imported from the UCI Machine Learning Repository and made into dataframes through use of the Python Data Analysis Library. Each dataframe was cleaned by renaming columns to be more concise and the test feature for each dataframe was turned into a categorical variable. For the dataframe regarding cooling load requirements, a heating load requirement was also provided, however this extra target variable was removed for clarity.

Each algorithm was partitioned into three different training and testing splits: a 80 percent and 20 percent split, a 50 percent and 50 percent split, and a 20 percent and 80 percent split. All the algorithms and their splits were tested on every dataframe three times to ensure the best and most accurate scores. Finally, each algorithm was cross validated and given a score.

Assessment

After having trained the algorithms on the immunotherapy, raisins and energy efficiency datasets, the random forest algorithm was the most accurate. It averaged an accuracy of 0.83 and cross validation score of 0.82 on the immunotherapy set, an accuracy of 0.86 and cross validation score of 0.85 on the raisins set and a 0.96 accuracy and 0.96 cross validation score on the energy set. Next, the support vector machine scored better, which averaged an accuracy of 0.79 and cross validation score of 0.75 on the immunotherapy set, an accuracy of 0.77 and cross validation score of 0.74 on the raisins set and a 0.76 accuracy and 0.62 cross validation score on the energy set. Finally, the logistic regression classifier scored the worst in terms of the metrics, which veraged an accuracy of 0.7 and cross validation score of 0.68 on the immunotherapy set, an accuracy of 0.72 and cross validation score of 0.83 on the raisins set and a 0.83 accuracy and 0.96 cross validation score on the energy set.

Conclusion

The random forest classifier had a higher overall placement across the three algorithms in terms of higher accuracies and lower classification errors. The linear support vector classifier would follow suit, with the logistic regression classifier having a lower overall placement. The placement of said classifiers is testament to the general flexibility of the random forest algorithm has in terms of predicting data, despite its sole use to classify two-class data throughout this experiment.

References

Çinar,İlkay, Koklu,Murat, and Tasdemir,Sakir. (2023). Raisin. UCI Machine Learning Repository. https://doi.org/10.24432/C5660T.

Khozeimeh,Fahime, Alizadehsani,Roohallah, Roshanzamir,Mohamad, and Layegh,Pouran. (2018). Immunotherapy Dataset. UCI Machine Learning Repository. https://doi.org/10.24432/C5DC72.

Tsanas,Athanasios and Xifara,Angeliki. (2012). Energy efficiency. UCI Machine Learning Repository. https://doi.org/10.24432/C51307.