A Comparison of Various Supervised Machine Learning Algorithms

Matthew Kang mjk095@ucsd.edu

COGS 118A, University of California San Diego

ABSTRACT

There are very few methodological evaluations of supervised machine learning algorithms from the modern era. This is in part, due to the fact that the expansion of the field of supervised machine learning is a relatively recent development. A paper done in 2006 by Caruana & Niculescu-Mizil elaborates on the variety of different supervised machine learning algorithms, and their respective accuracies. This paper seeks to accomplish a similar comparison of machine learning algorithms as that achieved in CNM06, on a smaller scale.

I. INTRODUCTION

The prevalence of data collection in daily life has become widespread in the modern era. Cell phones, laptops, smart TV's and other devices that people tend to use everyday serve as data collection entry points. From these things, there is now a plethora of data to sort through if one wishes to make algorithmic based predictions. Now that the problem of prediction has shifted from "how to collect enough data" to "how to utilize this data", the field of supervised machine learning has become an area of intense focus. An influx of medical data, for example, has led to the creation of many algorithms that can now predict whether a patient has a disease or not, using the contextual data of other medical patients with similar biological attributes. An example of a supervised machine learning algorithm that could be applicable in a situation like that would be K-Nearest Neighbors. Machine learning algorithms have paved the way from lidar technology for self driving cars to facial recognition. But not all algorithms are made equal. Some algorithms solve binary "0 or 1" prediction problems, while others give probabilities of events happening. Furthermore, different real life problems require different algorithms to solve them. In a 2006 paper by Caruana & Niculescu-Mizil, various different supervised machine learning algorithms are tested, compared, and analyzed. CNM06 tests ten different machine learning algorithms, using eight different performance metrics, on eleven different datasets. The validity of these datasets are then compiled and organized into a table. CNM06 has become an extremely influential paper in the field of supervised machine learning. In this paper, the procedures and methods of CNM06 will be replicated, in a simpler fashion. Three different supervised machine learning algorithms will be tested, on three different data sets. The only performance metric being accounted for will be accuracy.

II. METHODOLOGY

2.1 Algorithms Used

Listed below are the three algorithms that will be used and compared throughout this paper. The parameters and parameter specifications will be the same as those in the CNM06 paper. Specific hyperparameters will be tested for and found. Each three algorithms will be tested on three different datasets, for three trials. That means there will be 27 total trials. Each trial will randomly choose 5000 data points within the respective dataset for five cross validation in order to find hyperparameters via gridsearch.

K-Nearest Neighbors :

KNN will be implemented. Distance between points will be measured by the Euclidean Distance. The size of the training set will be 25 k values used. Hyperparameters will be optimized using gridsearch.

Support Vector Machines:

The regularization parameter, C, will vary by factors of ten from 10^-7 to 10^3.

Hyperparameters will be optimized using gridsearch.

Logistic Regression:

The hyperparameter for ridge regression (lambda) will be tested by factors of 10 from 10^-8 to 10^4. This would indicate 14 different hyperparameter settings being tested.

Hyperparameters will be optimized using gridsearch.

2.2 Performance Metrics

The only performance metric being used in this paper will be accuracy. This is the ratio of the number of predicted elements that exist within the set of the true elements. It is accessed using sklearn.metrics.accuracy_score.

2.3 Datasets

Three datasets will be used. They have all been retrieved from the UCI Machine Learning Repository.

Adults

The adult data set represents a variety of census data on around 50,000 adults. The original intended purpose of this data set was to find out if there were factors that would determine whether or not an adult would make more than $50,000 a year. Features include things like age, employment status, race, education, marital status,sex, occupation, country of origin, etc.

Bank

The bank marketing data set represents bank data gathered from around 50,000 adults. It's original intended purpose was for use with bank telemarketing. Each row represents one client of the bank. Things like age, income, occupation, marital status, and loan status are recorded to name a few features.

Cov_type

The cover type data set is from a geological study. It tracks data related to forest cover. 30 meter by 30 meter squares of forest were analyzed in northern Colorado, with features being listed for each square. Things like elevation, slope, soil type, and cover type were recorded, to name a few. There are 40 different soil types, with each soil type being one-hot encoded. I cannot stress how difficult that made things.

Experiment

Process

The basics of this experiment are as follows. We have our three datasets, adult, bank, and cov_type. On each of these three datasets, we will perform KNN, SVM, and Linear regression. Each of these algorithms will be performed three times. So on adult, for example, we will perform KNN, SVM, and Linear regression. Adult KNN will be performed for three trials, adult SVM will be performed for three trials, adult Linear regression will be performed for three trails. Same process applies for bank and cov_type.We will search for hyperparameters through gridsearch optimization. Each algorithm (KNN,SVM,LinReg) will have a different number of hyperparameter settings. In order to find the optimal hyperparameters, 5 fold cross validation will be used on 5,000 randomly selected points. Once the optimal hyperparameters have been chosen, the model will be tested on all the remaining points. (minus the 5000 that were used for hyperparameter optimization).

Table 1 : Mean Test Set Performance for each Algorithm/Dataset Combination

	Mean Accuracy : Adult	Mean Accuracy: Bank	Mean Accuracy for Cov_type
KNN	0.8214	0.8847	incomplete
SVM	0.7867	0.7733	incomplete
LinReg	0.8058	0.88525	incomplete

The best hyperparameters for KNN for the adult data was a k value of 4 with uniform weights.

The best hyperparameters for KNN for the bank data was a k value of 8 with uniform weights.

The best hyperparameters for SVM for the adult data was a C value of 0.1 with a linear kernel.

The best hyperparameters for SVM for the bank data was a C value of 10^-7 with a linear kernel.

The best hyperparameters for LinReg for the adult data was a C value of 1 with penalty being L2.

The best hyperparameters for LinReg for the adult data was a C value of 10^-8 with penalty being L2.

Conclusion

Different algorithms have different hyperparameters with much testing needing to be done in order to find the optimal hyperparameters. The abundance of algorithms that exist today with which programmers and analysts can apply to data means that drawing conclusions from data is now a matter of choosing the right algorithm. Different algorithms for different situations. This paper, in its attempts to replicate the 2006 paper by Caruna and Niculescu, has showed that different supervised machine learning algorithms will perform either better or worse when faced with different data and hyperparameters.

References

Caruana, Rich., & Niculescu-Mizil , Alexandru. *An Empirical Comparison of Supervised Learning Algorithms.* Department of Computer Science, Cornell University, Ithaca

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.