Comparison of Supervised Learning Algorithms

Jeffrey Lau j8lau@ucsd.edu

COGS 118A Professor Fleischer

December 16, 2020

Abstract

Many supervised learning algorithms were covered in COGS 118A, including 2-class classification problems and different loss functions. Here, an empirical study was conducted on three supervised learning algorithms to analyze each of the classifier's performances. Random forests, logistic regression, and K-nearest neighbors were used for the comparison of the three algorithms. Grid search with five fold cross validation was performed to select the best hyperparameters. Accuracy scores were used as the performance metric of the model performance.

1. Introduction

Inspired by Caruana and Niculescu-Mizil's study, this project also performed an empirical study comparing learning algorithms evaluated by accuracy performance (Caruana and Niculescu-Mizil, 2006). Multiple datasets were used to increase variance and better test the performance of each algorithm. In order to conduct a project similar to the one performed by Caruana and Niculescu-Mizil, the three datasets used in this project are ADULT, COV_TYPE, and LETTER, which are also used in the study performed by Caruana and Niculescu-Mizil. These three datasets are large; each has over ten thousand samples, which is adequate to perform five fold cross validation. The algorithms used in this project are random forests, K-nearest neighbor, and logistic regression, which are also used by Caruana and

Niculescu-Mizil. Each of these algorithms uses different parameters to compute the weights for classification, allowing variation in performance. Cross validation was performed to find the best parameters for each algorithm to further improve the accuracy of the classifiers. The results of the project show that random forests is a better performing model than K-nearest neighbor and logistic regression.

2. Methods

2.1 Methods

This section describes the hyperparameters for each algorithm used in cross validation. Most of the hyperparameters were chosen based on those used by Caruana and Niculescu-Mizil in their study, except for the hyperparameters in K-Nearest Neighbors. In their study, "we use 26 values of K ranging from K=1 to K=|trainset|" (Caruana and Niculescu-Mizil, 2006). However, those hyperparameters might lead to over-fitting. Therefore, linear space of 26 values of K ranging from 1 to 500 were used instead.

Logistic regression (LOGREG): This algorithm is best for convex optimization. The regularization parameter, C, ranges from $10^{-8}\ to\ 10^{4}.$

KNN: K nearest neighbors could be used for classification problems by picking the k nearest points. The nearest neighbors used with the KNN algorithm were done with 26 linear spaced values of K ranging from 1 to 500. Such procedure results in {1, 21, 42, 63, 84, 104, 125, 146, 167, 188, 208, 229, 250, 271, 292, 312, 333, 354, 375, 396, 416, 437, 458, 479, 500}

Random Forests (RF): This algorithm has a feature for splitting, and the features tested were {1, 2, 4, 6, 8, 12, 16, 20}. The forests had 1024 trees.

2.2 Performance Metrics

The classification accuracy score was used as the performance metric in this project. Since the datasets used in the project do not have imbalanced classes, the classification accuracy score would suffice.

Table 1 Description of Problems

Problem	Attributes	Train Size	Dataset Size
ADULT	14/105	5000	32561
COV_TYPE	54	5000	581012
LETTER	16	5000	20000

2.3 Data Sets

The algorithms were tested on three binary classification problems. The data ADULT, COV_TYPE, LETTER were retrieved from UCI's machine learning repository. ADULT determines the income of an individual based on multiple factors, with income greater than 50K a year classified as positive and income less than or equal to 50K a year classified as negative. Since there are many categorical features in ADULT, one-hot encoding was used to transform the data into numerical values, consequently expending attributes of the dataset from 14 to 105. The data COV_TYPE predicts forest cover type based on multiple factors. After finding which forest cover type occurs the most, the data was used to classify forest cover type 2 as positive and other cover type as negative. The third dataset, LETTER, identifies letter based on multiple factors. As mentioned in (Caruana and Niculescu-Mizil, 2006), classifying letter O as positive and other alphabets as negative would lead to an unbalanced problem. Therefore, dataset LETTER was used to classify letter A-M as positive and other letters as negative. Table 1 above summarizes the characteristics of the problems.

3. Experiments

Three trials were performed for each combination of dataset and algorithm, resulting in a total of 27 trials in this project. In each trial, 5000 random samples were drawn for training, and the remaining samples were used for testing. 5-fold cross validation was then performed using grid search on the training data of the 5000 random samples to find the best hyperparameters to use. This approach allows the results to be more reliable and fair. Using the best hyperparameters, each algorithm was trained on the whole training set, and was used to predict on the test set. The next procedure was to get the accuracy score of the training accuracy and the test accuracy. As a result, there are 27 best estimators, training accuracy, and test accuracy from 27 total trials. The two sample t-tests were performed to compare between each pair of algorithms and between each pair of algorithms for the different datasets.

3.1. Performances by Metric

The classification accuracy was used to evaluate the performance of the classifiers. Table 2 displays the average of classification accuracy, which is the test accuracy mentioned above, of the three datasets and three trials for each algorithm. Of the 27 trials, random forests was the best performing model with 87.4%. K-nearest neighbors was the second best performing model with 84.2%, while logistic regression was the worst performing model with 71.3%.

Table 2 Classification Evaluation

Model	Mean of 9 Test Scores (ACC)
Logistic Regression	0.713
KNN	0.842
Random Forest	0.874

For table 3, the 2 sample t-test was used to determine whether mean accuracy of algorithm A and of algorithm B differs significantly or is it due to chance. ttest_rel from scipy library was used in this scenario because the mean accuracies of the two algorithms being compared were thought to be similar.

Logistic regression produces a mean accuracy that is significantly different from K-nearest neighbors with the p-value of 0.0076 and cohen's d of -1.56. Logistic regression also produces a mean accuracy that is significantly different from random forest with the p-value of 0.0003 and cohen's d of -2.31. K-nearest neighbor produces a mean accuracy that is significantly different from random forest with the p-value of 0.0213 and cohen's d of -0.42.

Table 3 2 sample t-tests across 3 algorithms

Model	Mean of 9 Test Scores (ACC)
Logistic Regression vs KNN	0.0076
Logistic Regression vs Random Forest	0.0003
KNN vs Random Forest	0.0213

3.2. Performances by Problem

Similar to evaluating performance using metrics, the classification accuracy was used to evaluate the performance of the classifiers across different datasets. Table 4 displays the average accuracy of the three trials for each dataset using different algorithms.

Table 4 displays that K-nearest neighbors performed slightly better on COV_TYPE than random forest did. Aside from that, random forest generally performed with the highest classification accuracy. Overall, logistic regression was the worst performing model for all three datasets.

Table 4 Mean classification accuracy of each algorithm by different dataset

Model	Cov_type	Adult	Letter
Logistic Regression	0.726	0.797	0.615
KNN	0.958	0.785	0.784
Random Forest	0.949	0.851	0.821

In table 5, the 2 sample t-test was used to determine whether mean accuracy of algorithm A and of algorithm B differs significantly or is it due to chance for 3 different datasets. ttest_rel from scipy library was used in this scenario because the mean accuracies of the two algorithms for the different datasets being compared were thought to be similar.

Logistic regression produces a mean accuracy that is significantly different from K-nearest neighbors with the p-value of 0.00024 in COV_TYPE, p-value of 0.01231 in ADULT, and p-value of 0.00012 in LETTER.

Logistic regression also produces a mean accuracy that is significantly different from random forest with the p-value of 0.00005 in COV_TYPE, p-value of 0.00189 in ADULT, and p-value of 0.00034 in LETTER.

K-nearest neighbor produces a mean accuracy that is significantly different from random forest with the p-value of 0.00543 in COV_TYPE, p-value of 0.00122 in ADULT, and p-value of 0.03260 in LETTER.

Table 5 2 sample t-test of each algorithm by dataset

Model	Cov_type	Adult	Letter
Logreg vs KNN	0.00024	0.01231	0.00012
Logreg vs RF	0.00005	0.00189	0.00034
KNN vs RF	0.00543	0.00122	0.03260

4. Conclusions

This study examined the performance of three classifiers, logistic regression, random forests, and K-nearest neighbors, on three datasets, ADULT, COV_TYPE, LETTER. The results of this study showed two separate performances of the classifiers with the mean classification accuracy used as the performance metric. In one of the performances where the average of classification accuracy of the combination of datasets and trials for each algorithm was used, random forests had the highest average, and K-nearest neighbors had the second highest average. Logistic regression was the worst performing model. For the other performance where the average accuracy of the three trials for each dataset was used, random forests generally performed with the highest mean classification accuracy. Logistic regression was the worst performing model for all three datasets. The 2 sample t-test was used on both performances to determine whether the mean accuracy of

algorithm A and B differs significantly or not. In both performances, all p-values were less than the threshold, $p<0.05$, demonstrating that random forests performed better than K-nearest neighbors, and K-nearest neighbors performed better than logistic regression; none of these scenarios happened by chance.

A great deal was learned through this study. In future works, more algorithms could be used in grid search to better find the optimal classifier. Other performance metrics such as RMS, ROC, APR, FSC, etc should be used to further understand the performance of learning algorithms.

References

Blake, C., & Merz, C. (1998). UCI repository of machine learning databases.

Caruana, Rich, and Alexandru Niculescu-Mizil. (2006). An Empirical Comparison of Supervised Learning Algorithms. *Proceedings of the 23rd International Conference on Machine Learning (ICML2006).*

Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.