A Comparison of Supervised Learning Algorithms

Ashley Yu Jia Chen YJC029@UCSD.EDU

Cogs 118A Professor Tu

March 20 2020

Abstract

In Cogs 118A, many supervised learning algorithms were covered. We study the 2-class classification problem and the different loss functions, then analyzed each of the classifier's performances. To apply what has been learned in the course, we present an empirical study on four algorithms. The comparison of methods is between random forests, decision trees, K-nearest neighbors, and logistic regression. We examine the impact of varying splits between training and testing data for a variety of datasets. Cross validation is also performed to find the optimal hyperparameters. Overall, our performance criteria is based on the training and testing accuracy for each classifier, where we rank algorithms by its testing averages.

1. Introduction

Based on the studies performed by Caruana and Niculescu-Mizil, this paper also performs an empirical analysis, on a smaller scale, of four supervised learning algorithms evaluated by accuracy performance (Caruana and Niculescu-Mizil, 2006). This project was carried out in order to compare and contrast different learning algorithms. It was carried out with multiple datasets to increase the variance and test each's performance. Each algorithm has different ways of computing the weights for classification but have similar performances. First, cross validation was performed in addition to further increase the accuracy of the

classifiers. After the experiments ran, the results were slightly more accurate when the training set was large. Random forest do seem to perform better on average, depending on the dataset.

2. Methods

2.1. Learning Algorithms

This section describes the hyperparameters for each algorithm tested during cross validation (if possible) to fit each dataset best, and other parameters inputted. The numbers chosen were based on the study mentioned earlier by Caruana and Niculescu-Mizil. Every algorithm was provided by the Scikit-Learn package. These are the four models used in the study:

Logistic Regression (LOGREG): This algorithm creates a softened loss function for convex optimization. The C range tested was {0.00000001, 0.0000001, 0.000001, 0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000}.

KNN: K nearest neighbors is a non-parametric estimator that can be used for classification problems by picking the k nearest points to the query point. The nearest neighbors tested were {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23}.

Random Forests (RF): This algorithm combines tree predictors dependent on sampled random vectors, an ensemble method. The max features tested were {1, 2, 4, 6, 8, 12, 16, 20} and max depth was {1, 2, 3, 4, 5}. The forest had 1024 trees.

Decision Tree (DT): This non-parametric model recursively splits so the leaf nodes hold the classifications. The depths tested was {1, 2, 3, 4, 5}.

2.2. Performance Metrics

The classification accuracy was used as the main performance metric to determine the power of each learning algorithm. Specifically, training, testing, and validation errors were recorded. One performance metric is enough for this small scale experiment.

Table 1. Description of Problems

Problem	Attributes	Data Set Size
IRIS	6	150
SHROOM	23	8123
WINE	12	1599
INCOME	15	32561

2.3. Data Sets

We tested the algorithms on four binary classification problems. The data sets were retrieved from UCI's machine learning repository. SHROOM contained all categorical values, so we converted each value with one-hot encoding. Edible mushrooms were classified as positive, poisonous ones were negative. The second dataset, IRIS, contained three flower groups with negative values to Setosa and Versicolor, and positive is Virginica. If the iris target is greater than 1.5, then it is categorized as Virginica, and the other two otherwise. WINE divides red wines by a rating system for high quality wines with an aggregate score $>$ 6.5 (1) or low quality with $<=$ 6.5 (-1). All the features are categorical and weighted together created the score for the wines. The fourth data set, INCOME, determines the income of a person based on multiple factors, with $>$50k a year classified as a positive value and less than that being negative. Some missing, repeated, or irrelevant data was dropped to keep the classification algorithm consistent. Also, categorical data was converted into numbers. Table 1 summarizes the problem characteristics.

3. Experiments

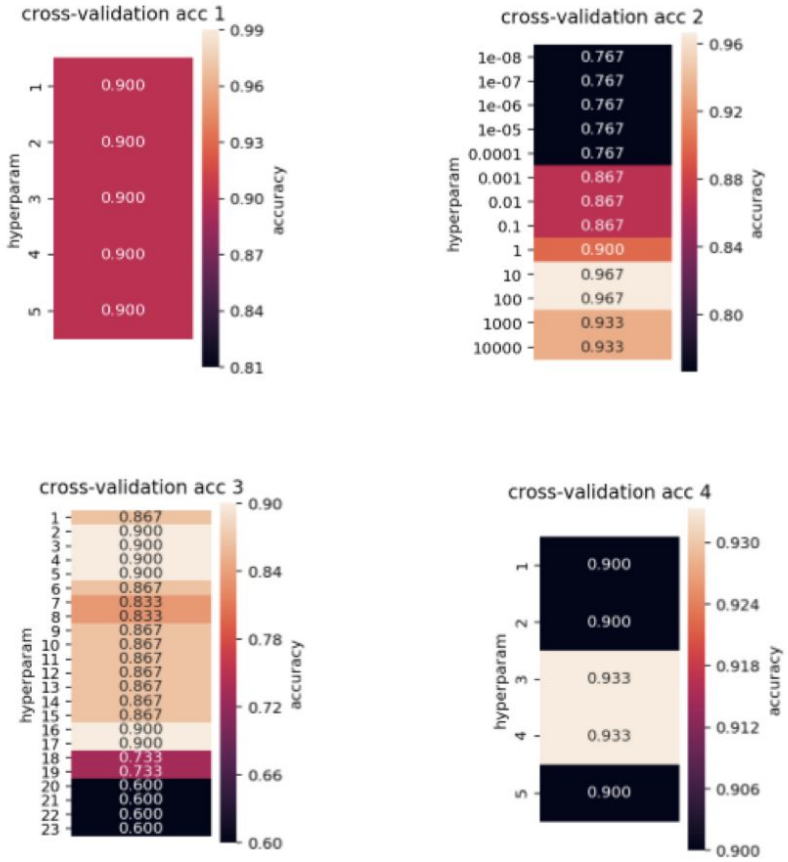

Image 1. Grid Search

For every data set, we ran three partitions, splitting the training and testing data to 20/80, 50/50, and 80/20. This is to see the effects of the training versus testing data ratio. Next, three trials were performed for each split, with the data randomly sectioned for each trial. This way, the results could be more reliable and make sure that every sample has a fair chance to appear. For every dataset's partition, we ran 5-fold cross validation for all four classifiers using grid search to obtain the best hyperparameters to use. Image 1 displays the results from one classifier of one trial.

Table 2. Classification Results

Table 1. Classification results of Random Forests, Decision Tree, Logistic Regression, and K-Nearest Neighbor classifiers on Dataset A, Dataset B, Dataset C and Dataset D. Accuracy averaged over three trials/repeats.

	Dataset A (20/80)		Dataset B (20/80)		Dataset C (20/80)		Dataset D(20/80)		Average
	Training Accuracy	Testing Accuracy	Training Accuracy	Testing Accuracy	Training Accuracy	Testing Accuracy	Testing Accuracy	Training Accuracy	Testing Accuracy
Random Forests	100%	94%	99%	99%	93%	88%	85%	85%	92%
Logistic Regression	97%	96%	100%	100%	88%	87%	80%	80%	91%
K-Nearest Neighborhood	92%	91%	100%	100%	87%	86%	80%	79%	89%
Decision Tree	99%	94%	100%	100%	88%	86%	84%	84%	91%
	Dataset A (50/50)		Dataset B (50/50)		Dataset C (50/50)		Dataset D(50/50)		Average
	Training Accuracy	Testing Accuracy	Training Accuracy	Testing Accuracy	Training Accuracy	Testing Accuracy	Testing Accuracy	Training Accuracy	Testing Accuracy
Random Forests	97%	94%	99%	99%	92%	88%	85%	85%	92%
Logistic Regression	98%	95%	100%	100%	88%	87%	80%	80%	91%
K-Nearest Neighborhood	98%	96%	100%	100%	91%	85%	81%	80%	90%
Decision Tree	98%	94%	100%	100%	87%	86%	84%	84%	91%
	Dataset A (80/20)		Dataset B (80/20)		Dataset C (80/20)		Dataset D(80/20)		Average
	Training Accuracy	Testing Accuracy	Training Accuracy	Testing Accuracy	Training Accuracy	Testing Accuracy	Testing Accuracy	Training Accuracy	Testing Accuracy
Random Forests	97%	98%	99%	99%	92%	89%	85%	85%	93%
Logistic Regression	98%	100%	100%	100%	89%	87%	80%	80%	92%
K-Nearest Neighborhood	97%	98%	100%	100%	89%	85%	80%	80%	91%
Decision Tree	97%	98%	100%	100%	90%	86%	84%	84%	92%

3.1. Performances by Metric

Table 2 displays the classification results organized by a table for each split. Each dataset contains the average testing and training accuracy of the four learning algorithms. For dataset B, the training and testing accuracy are extremely correct, at 100%. Looking at the averages in testing accuracy, the best performing model is random forests with

92-93%. Overall, all the classifiers performed with high accuracy, with small variations depending on the classifier and some datasets affecting the accuracy too.

Table 3. Algorithm Rankings per Problem

Table 3. Rank order of the classifiers in comparison.				
	1st	2nd	3rd	4th
Random Forest	0.5	0.25	0.25	0.0
Logistic Regression	0.5	0.25	0.25	0.0
K-Nearest Neighborhood	0.0	0.25	.75	0.0
Decision Tree	0.25	0.5	0.25	0.0

3.2. Performance by Problem

The results for each classifier ranked by highest average accuracy is displayed in Table 3. Random forest and logistic regression performed well with 0.5 chance of placing first. Decision tree was mainly ranked second and KNN as third. Many classifiers tied for the same rank, so there were no fourth places.

4. Conclusion

Through this experiment, we learned about the performance of four classifiers on four datasets. The results had extremely high accuracy and the algorithms mainly equal each other in classification strength. Dataset B could have skewed the results because the classification was very accurate. Overall, logistic regression and random forest performed the best for the datasets combined. Even though the classifiers performed well on average, the individual results were also dependent on the datasets and the training/testing size. Even though the results don't appear to be affected much by the split sizes, for dataset A, the testing accuracy rises with greater training data. A great deal was learned through testing and next time, larger datasets should be used to increase the precision of the results.

References

Blake and C. Merz. UCI repository of machine learning databases, 1998.

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

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