An Empirical Comparison of Supervised Learning Algorithms

Billy Sudirdja bsudirdja@ucsd.edu

Abstract

In our COGS 118A class we learned about a lot of different types of classifiers that use algorithms to try and achieve the goal of assigning a class to an input based on its features. In this evaluation we will be comparing the results of Logistic Regression, Random Forests, and K-Nearest Neighbors on three datasets from the UCI ML Repository.

1. Introduction

There are many different types of machine learning classifiers in today's age and these three were chosen since they were all importable from the sklearn library. The sklearn library is an extensive library with a lot of tools to do machine learning in Python and we will be using their library for consistency purposes. If they were to be implemented by myself I would feel more inclined to optimize the classifiers that are more familiar to me so I left that out of the equation. Also they have great tools to deal with large datasets like a partition function. For our three datasets we will be partitioning them into three different categories. We are going to split the dataset into a 20/80 split where 20% of the data gets randomly selected to be a part of the training dataset while the other 80% is used for testing. This split between the dataset is then done with a 50/50 split and a 80/20 split on all three datasets. The datasets are from the UCI ML Repo and we are using the Iris dataset (Fisher,R), Adult dataset (Becker,Barry and Kohavi,Ronny), and the Rice dataset (Cammeo and Osmancik). These three datasets are all easy to work with since they can just be imported from the ucimlrepo library in python. They are also beginner friendly since they all deal with binary classification. The adult dataset takes a look at features of adults such as their age, education, and capital-gain. We will be only taking the features that are numerical for simplicity sake. The data is paired with the resulting target vector of income which is binary with the classes being “>50K” and “<=50K”. The Iris dataset is a dataset we’ve been using throughout the quarter and is a small dataset with only 4 features that has information about a flower's sepal and petal length and width. The resulting target vector is then categorical type of either Iris Setosa, Iris Versicolour, or Iris Virginica. The Rice dataset looks at the physical features of individual rice grains such as the length and glossiness. The data’s target vector is split into two classes as either from the Cammeo species or the Osmancik species.

2. Method

An Empirical Comparison of Supervised Learning Algorithms

Billy Sudirdja bsudirdja@ucsd.edu

Different classifiers will have different ways of implementing and optimizing them but for this experiment these are the parameters given to the datasets. We will be running the three classifiers and checking their accuracies for the three partitions and taking the best accuracy for each of the trials. For logistic regression we will not be adding in any bias since we are not concerned with any fine tuning for the recall or precision since we are just comparing the results. For the random forest we will be using an n_esimators as 100. This means that the number of trees in the forest will be 100. Generally speaking the larger the number of trees in the forest the better the classifier does but again we do not want to skew the results of the classifiers. For our KNN or nearest neighbor classifier we went with the default value of n_neighbors as 5.

An Empirical Comparison of Supervised Learning Algorithms

Billy Sudirdja bsudirdja@ucsd.edu

3. Experiment

Adult Dataset

Classifier	Train/Test split	Training Acc	Validation Acc	Testing Acc	Hyperparameters
Logistic Regression	20/80	0.534	[0.53045348 0.52912274 0.53378378 0.52825553 0.53378378]	0.530	
Logistic Regression	50/50	0.532	[0.53045348 0.52912274 0.53378378 0.52825553 0.53378378]	0.532	
Logistic Regression	80/20	0.532	[0.53045348 0.52912274 0.53378378 0.52825553 0.53378378]	0.531	
Random Forests	20/80	0.661	[0.52267376 0.51919337 0.52405815 0.5250819 0.5250819]	0.500	
Random Forests	50/50	0.628	[0.52267376 0.51919337 0.52405815 0.5250819 0.5250819]	0.516	
Random Forests	80/20	0.614	[0.52267376 0.51919337 0.52405815 0.5250819 0.5250819]	0.524	
K Nearest Neighbors	20/80	0.545	[0.52543761 0.51489405 0.5299959 0.53194103 0.53276003]	0.500	
K Nearest Neighbors	50/50	0.546	[0.52543761 0.51489405 0.5299959 0.53194103 0.53276003]	0.516	
K Nearest Neighbors	80/20	0.517	[0.52543761 0.51489405 0.5299959 0.53194103 0.53276003]	0.524	

An Empirical Comparison of Supervised Learning Algorithms

Billy Sudirdja bsudirdja@ucsd.edu

Iris Dataset

Classifier	Train/Test split	Training Acc	Validation Acc	Testing Acc	Hyperparameters
Logistic Regression	20/80	1.0	[0.96666667 1.0 0.93333333 0.96666667 1.0]	0.966	
Logistic Regression	50/50	0.933	[0.96666667 1.0 0.93333333 0.96666667 1.0]	1.0	
Logistic Regression	80/20	0.975	[0.96666667 1.0 0.93333333 0.96666667 1.0]	1.0	
Random Forests	20/80	1.0	[0.96666667 0.96666667 0.93333333 0.96666667 1.0]	0.933	
Random Forests	50/50	1.0	[0.96666667 0.96666667 0.93333333 0.96666667 1.0]	0.986	
Random Forests	80/20	1.0	[0.96666667 0.96666667 0.93333333 0.96666667 1.0]	1.0	
K Nearest Neighbors	20/80	0.966	[0.96666667 1. 0.93333333 0.96666667 1.]	0.933	
K Nearest Neighbors	50/50	0.96	[0.96666667 1. 0.93333333 0.96666667 1.]	0.986	
K Nearest Neighbors	80/20	0.966	[0.96666667 1. 0.93333333 0.96666667 1.]	1.0	

An Empirical Comparison of Supervised Learning Algorithms

Billy Sudirdja bsudirdja@ucsd.edu

Rice Dataset

Classifier	Train/Test split	Training Acc	Validation Acc	Testing Acc	Hyperparameters
Logistic Regression	20/80	0.919	[0.93963255 0.95144357 0.93044619 0.93044619 0.8976378]	0.933	
Logistic Regression	50/50	0.931	[0.93963255 0.95144357 0.93044619 0.93044619 0.8976378]	0.928	
Logistic Regression	80/20	0.931	[0.93963255 0.95144357 0.93044619 0.93044619 0.8976378]	0.925	
Random Forests	20/80	1.0	[0.93569554 0.93700787 0.92125984 0.92388451 0.88451444]	0.924	
Random Forests	50/50	1.0	[0.93569554 0.93700787 0.92125984 0.92388451 0.88451444]	0.919	
Random Forests	80/20	1.0	[0.93569554 0.93700787 0.92125984 0.92388451 0.88451444]	0.925	
K Nearest Neighbors	20/80	0.883	[0.91469816 0.89107612 0.87270341 0.88845144 0.83070866]	0.924	
K Nearest Neighbors	50/50	0.908	[0.91469816 0.89107612 0.87270341 0.88845144 0.83070866]	0.919	
K Nearest Neighbors	80/20	0.914	[0.91469816 0.89107612 0.87270341 0.88845144 0.83070866]	0.925	

An Empirical Comparison of Supervised Learning Algorithms

Billy Sudirdja bsudirdja@ucsd.edu

4. **Conclusion**

These three tables are formatted the same for readability. On the leftmost column we have the type of classifier which is either logistic regression, random forests, or KNN. The next column is followed with the partition of the dataset for testing and training. The next column shows the training accuracy of each of the partitions followed by the validation accuracy. For the cross validation we do a K-fold cross validation with 5 folds. We can take the mean of these 5 folds to obtain a metric for seeing how good our model might be on data it has never seen before. The next column is the testing accuracy. Followed by the hyperparameters. Since we are not finetuning any of these algorithms we do not need to worry about the hyperparameters.

5. **References**

Becker,Barry and Kohavi,Ronny. (1996). Adult. UCI Machine Learning Repository. https://doi.org/10.24432/C5XW20.

Fisher,R. A.. (1988). Iris. UCI Machine Learning Repository. https://doi.org/10.24432/C56C76.

Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E., Scikit-learn: Machine Learning in Python(2011). https://scikit-learn.org/stable/about.html

Rice (Cammeo and Osmancik). (2019). UCI Machine Learning Repository. https://doi.org/10.24432/C5MW4Z.