COSC 5010-03 Practical Machine Learning Fall 2023 Algorithm Selection Report

Michael Elgin

October 11, 2023

1 Introduction

Regression and classification are two of the most important tasks in supervised machine learning. This exercise compares the performance of various regression and classification models on the wine quality dataset[1]. 4 regression models are evaluated, and 6 classification models are evaluated

2 Dataset Description

The wine quality dataset is standard tabular data. There are 2 datasets for each color of red and white. Both contain 11 features, all of which are continuous. The target is a discrete value which is the assigned quality of the wine. For the regression tasks, the white wine dataset is used to evaluate models. For classification tasks, both datasets are concatenated, with the target being changed to be the color of the wine, with 0 being assigned to red and 1 to white. In both cases of classification and regression, the datasets are split into training and testing sets, with 20% for testing and 80% for training.

3 Experimental Setup

All computation is done with the Python programming language. Scikit-learn is used to construct models with the one exception of pygam to created a generalized additive model for regression. Pandas and numpy are used to load and preprocess the data along with scikit-learn's train_test_split function.

For regression, performance is defined by rounding the regression number predicted by the model to the nearest whole number, then the accuracy is defined as the amount of predictions that matched the assigned wine quality (also discrete) divided by the total number of samples in the test set. More formally:

$$GE_{Regression\ model}(\hat{f}, \boldsymbol{X}_{test}, \boldsymbol{y}_{test}) = \frac{\sum_{i=1}^{|\boldsymbol{X}_{test}|} l_{0,1}(\left\lfloor \hat{f}(\boldsymbol{X}_{test,i}) \right\rceil, \boldsymbol{y}_{test,i})}{|\boldsymbol{X}_{test}|}$$

[1] https://archive.ics.uci.edu/dataset/186/wine+quality

With the accuracy then defined as:

$$Acc_{Regression\ model} = (1 - GE_{Regression\ model}) * 100$$

For classification, the error performance is defined similarly:

$$GE_{Classification\ model}(\hat{f}, \boldsymbol{X}_{test}, \boldsymbol{y}_{test}) = \frac{\sum_{i=1}^{|\boldsymbol{X}_{test}|} l_{0,1}(\hat{f}(\boldsymbol{X}_{test,i}), \boldsymbol{y}_{test,i})}{|\boldsymbol{X}_{test}|}$$

$$Acc_{Classification\ model} = (1 - GE_{Classification\ model}) * 100$$

4 Results

Table 1: Regression model accuracies

	Accuracy
Baseline (mode guessing)	41.735%
Linear Model	48.163%
Decision Tree	57.959%
Random Forest	63.673%
Generalized Additive Model	50.816%

Table 2: Classification model accuracies

	Accuracy
Baseline (mode guessing)	76.077%
Support Vector Classifier	93.538%
Logistic Regression	98.692%
Decision Tree	98.231%
K-Nearest Neighbor	94.923%
Naive Bayes	97.615%
Random Forest	99.538%

In both regression and classification, the random forest algorithm performs the best.

5 Code

The associated code is in Alg_Selection.ipynb