Homework 1 Report

CSE590-12 Machine Learning

Jonathan Loyd

The following experiments serve as an introduction to machine learning concepts, and as an introduction to implementing them. The following experiments were conducted on given training and testing data for Wine Quality that was provided. The data provided was already split into training and testing data so there was no testing conducted on splitting the data.

The first model built was a K-Nearest Neighbors (KNN) Regression model. The KNN model is a simple algorithm that stores the training dataset and makes predictions for new data by finding the nearest neighboring points. In this model, the main parameter is K, which can be varied to determine what quality the wine is from the data set based on the K nearest neighbors. The testing for KNN involves creating a model with varying K values. The K values tested are the integers from 1 to 20 inclusive.

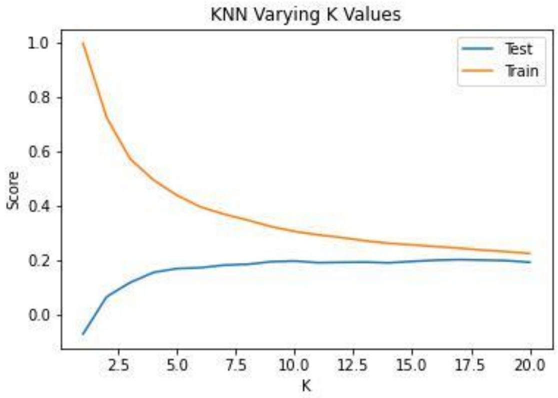

Figure 1: KNN Model with Varying K Values

From Figure 1, it is apparent that K in the range of 1 to 20 has an example of overfitting at K equal to 1, where training data has a score of 1, and the testing data has an extremely low score.

Table 1: K Values from 10 to 20

K	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Test Score	0.1995	0.1935	0.1948	0.1956	0.1930	0.1982	0.2026	0.2048	0.2031	0.2011	0.1950

Table 1 indicates that a K value of 17 gives the highest testing score accuracy, which indicates that a K value of 17 would be the optimal value for the given Wine Quality dataset.

The next experiment was to build and train an Ordinary Least Squares (OLS) Regression model. OLS is the simplest linear method for regression, and its goal is to minimize the mean squared error between predictions and true regression targets for the dataset by using the best parameters for w and b, which correspond to the coefficients and intercept. It should also be noted that this model has no parameters that can be modified by the user, so it is extremely simple, but will not allow the user to vary model complexity. The train and test scores for the model are as follows:

Training set score: 0.2761

Test set score: 0.3367

The OLS Regression model has poor accuracy but will have the same test score as Ridge Regression and LASSO Regression models when they have small alpha values. Out of the models, it is a better choice than the KNN Regression model in terms of test set accuracy, and it has the same performance as Ridge Regression and LASSO Regression at their best performance. These models also indicate that a more complex model may perform better on the data, so testing with additional models is necessary before determining whether OLS Regression is a good choice for the Wine Quality dataset.

After the OLS Regression model was built, a Ridge Regression model was built and trained. The Ridge Regression model has the benefit of controlling complexity over the simple OLS Regression model. Ridge Regression chooses coefficients (w) that predict well on the training data and minimize all instances of w. This constraint can help to regularize the data and avoid overfitting. By default, the constraint that the user can control, alpha, is equal to 1. A larger alpha value will force coefficients to move to 0, and this may improve generalization at the cost of decreasing the performance of the training set. The goal for this experiment is to vary alpha and determine the result it has on the training and testing accuracy of the model.

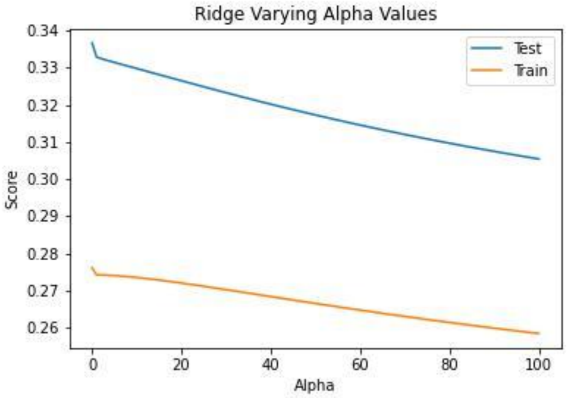

Figure 2: Ridge Regression Model and Test scores for Varying Alpha Values

Figure 2 shows that the model is likely underfitting as alpha increases, and there is an instance of underfitting at alpha equal to 100. There does not appear to be an example of overfitting with alpha in a range from 0 to 100. The graph highlights that as alpha increases, the testing and training performance decreases. The following trend along with the low accuracy for training data (capping at 0.3367) means that it is likely that this model is not complex enough for the Wine Quality dataset. A minimum value of alpha, such as 0 appears to be the best alpha value for this dataset since it gives the best performance for both training and testing data over any other alpha value. The minimum alpha value for Ridge gives the same accuracy as OLS, which is because when alpha is minimum, there are no constraints, which mimics OLS.

The last experiment is on the Least Absolute Shrinkage and Selection Operator (LASSO) Regression model. Like the Ridge model, the LASSO Regression model also restricts coefficients to be their minimum values, but Ridge uses L2 regularization, while LASSO uses L1. LASSO does this by allowing the model to entirely ignore some features through automatic feature selection. The user can vary the number of features used through the alpha variable. The default for alpha is 1, where only 1 feature is used, and as alpha decreases closer to 0, more features will be used. The final experiment tests the effect of decreasing alpha from 1 to 0 on the Wine Quality dataset.

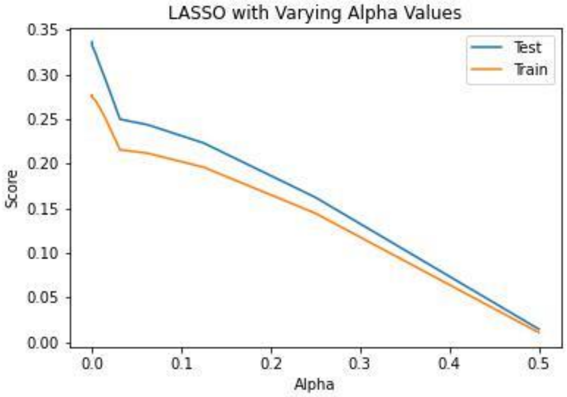

Figure 3: LASSO with Varying Alpha Values

The training and testing performance on the dataset decreases as alpha increases. There are no clear examples of overfitting in this dataset, but there are examples of underfitting as alpha decreases, with the most apparent example from Figure 3 at alpha equal to 0.5. Like the Ridge model, the optimal alpha value is 0, with a training accuracy of 0.3367. Again, increasing alpha simplifies the model and makes performance worse.

All the experiments have indicated that Linear Regression Models are going to perform better at their peak than a KNN regression model will on the Wine Quality dataset. The OLS, LASSO and Ridge models have the same peak performance with a test accuracy score of 0.3367. That means that OLS by itself, or LASSO or Ridge with minimum alpha values are going to be the best performing models for the Wine Quality dataset out of the four that were tested. These models still seem to be too simple for the model, which is indicated by the continuous decline in performance as alpha increases for both LASSO and Ridge.