Stone Barrett

CSE 590: Intro to Machine Learning

Homework 1

The goal of this assignment was to take a given set of training data, a given set of testing data, and knowledge of introductory machine learning statistical models and analyze the data using these models. The datasets regarding wine quality were provided in separate files and the four models needing training and analysis for this assignment are k-nearest neighbors, ordinary least squares, ridge, and LASSO regressions. The programs to build and train these models were developed in Python.

Problem 1a: KNN Regression

The K-Nearest Neighbors approach may be the simplest to understand. The placement of a given set of independent variables is considered to find the distance between it and the nearest k entries so that the dependent variable can be calculated.

We begin with verifying the shape of the data being read. This can be seen in Figure 1.1, below.

Text

Description automatically generatedFigure 1.1

Text

Description automatically generatedFigure 1.2

After that, we can see that in Figure 1.2 predictions are made for the testing scores based on the training data for each value of K.

As the above happens, the R^2 scores of the training and testing datasets are calculated and collected. Once every value of K (1-50) has been tested, the results are displayed graphically.

In Figure 1.3 below, we can see that at K’s smallest, the training data is a perfect fit and the testing data is as bad as it can get. This is the most extreme example of overfitting – scores varying by an entire point. As K grows, the training data scores plummet, while the testing data scores grow.

From this display, we can choose the optimal value of K to be K = 20, as past this point, underfitting becomes increasingly more of a problem. It is still relatively underfitting even at its best point, but the model stands strongest at k = 20.

Chart, line chart

Description automatically generated

Figure 1.3

Problem 1b: OLS Regression

Ordinary Least Squares regression is what is considered the traditional method for regression. It is simplest to run because it takes no parameters, but that also means there is no way to test different complexities.

Text

Description automatically generated

Figure 1.4

In Figure 1.4 above, we can see the coefficients and intercept found by OLS. We can also see the training and testing data sets’ scores sitting fairly low. Although the scores are close together, they do resemble an underfitting situation and for this reason OLS regression cannot be considered a good fit for this data.

Problem 1c: Ridge Regression

Ridge regression is a linear regression model that insists any individual attribute of a dataset should not significantly affect the prediction. One goal of Ridge regression is to starkly avoid overfitting.

Text

Description automatically generated

Figure 1.5

In Figure 1.5 above, we can see the scores for training and testing datasets for each alpha value tested. I suspect either this is a very poor model for this data or there is something wrong with the program itself, as all three alpha values yield the same scores and they are all very underfit. Figure 1.6 below shows the coefficient graph given.

Chart, scatter chart

Description automatically generated

Figure 1.6

Problem 1d: LASSO Regression

LASSO, or Least Absolute Shrinkage and Selection Operator, is a regression model that sometimes ignores certain attributes. It should be easier to interpret because those ignored coefficients are 0, exacerbating the effect of stronger features.

Text

Description automatically generatedFigure 1.7

In Figure 1.7 above, we can see the default alpha = 1 and no max iteration increase resulted in scores of 0. After increasing the max iteration count and trying alpha at .01 and .001, we can see that the optimal alpha value here should be considered as .001, though it is still a serious case of underfitting. Figure 1.8 below shows the resulting coefficient graph.

Figure 1.8Chart, scatter chart

Description automatically generated