Machine Learning
Assignment 3: Regression

R00159222 - Zachary Dair
zachary.dair@mycit.ie

Extra Information:

The assignment was done in python 3.8, uses the following libraries:

Pandas, numpy, matplotlib, and sklearn.

Table of contents:

If anything is unclear, please let me know, I apologise that the documentation is on the shorter side, my comments in the code should explain everything.

Task 1: Pre-Processing

The function loadData loads the contents of the diamonds.csv and extracts the unique types for each cut, color and clarity. (line 8-32)

Loop through combinations extracting the features and targets of those with more that 800 data points (line 120-132)

Task 2: Model Function

Trivariate polynomial model function (line 35-44)

Correct coefficient size calculation function (line 47-56)

Task 3: Linearization

Linearization function (line 60-73)

Model value calculation (line 62)

Jacobian calculation (line 66-72)

Task 4: Parameter Update

Normal Matrix calculation (line 80)

Regularisation and residual calculation (line 82)

Solving the final equation (line 86-87)

Task 5: Regression

Creates initial coefficient vector with zeros (line 94)

Iterative Procedure (line 96-104)

Task 6: Model Selection

Kfold initialisation (line 135)

Loop through each dataset calculating the best degree (line 137 - 189)

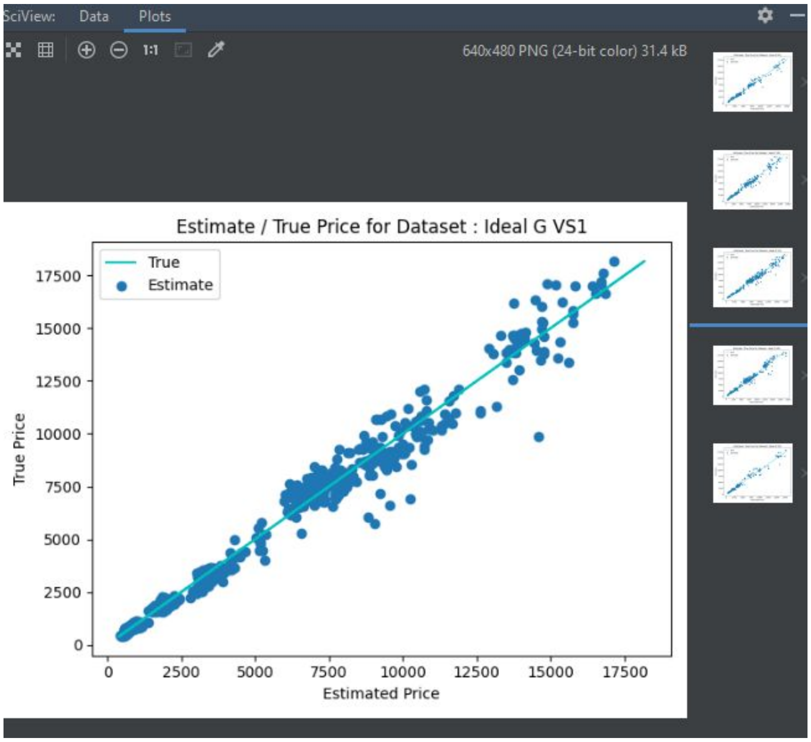
Using the mean price difference for each degree, we can determine the best degree for each dataset (line 163-189)

Determine the best degree per dataset (line 191-202)

Task 7: Visualisation

Loop through each dataset calculating the model parameters, using the corresponding optimal degree value (line 204-230)

Plot the estimated vs actual prices (line 219-228)



Here the estimate price VS actual has been plotted

Each plot also is named corresponding to the dataset (cut, color, clarity) combination

From the above plot, we can see that there is a somewhat linear distribution, where the estimated prices are quite close to the actual values axis. We can also see that there is a certain degree of variance between the actual and estimated values. This variance displays the accuracy of the model, we can conduct further analysis on this data by calculating the covariance.