# Task 1: Linear Regression

*Machine Learning Assignment*

**Matthew Doherty**
*Rhodes University*
*Lecturer: Dane Brown*
May 13, 2019

*Keywords:* Machine Learning

## Part A

```
Please see 3_4_Linear_Regression.py.
```

## Part B

Generalization error provides a measure of how accurately an algorithm is able to predict output for unseen values, thus generlisation error is minimised when a model does not overfit the data. A means of determining generalization error for the linear regression model is given by a loss function such as Mean Square Error (MSE):

```
R_Squared = 0.760
MSE = 14.996
```

Whereas the generalization error for ridge regression model is given by:

```
R_Squared = 0.763
MSE = 14.775
```

R-squared is a measure of how well outcomes are replicated by the model, or how much of the variance of the dependent variable is explained by the model. Since R-squared for the Ridge Regression is greater than R-squared for Linear Regression, we therefore can conclude that the Ridge regression model better generalizes to unseen data points than the Linear Regression model in this example. A good model should be able to minimize MSE which demonstrates the quality of the estimator. This further confirms the model's efficacy.

## Part C

The predicted plot for the Ridge regression model more closely fits a line to the ground truth data points, that is minimizes mean squared error. Therefore the Linear model overfits its model when training whereas the Ridge regression model introduces a regularization term to the model which allows it to better predict new data points and avoid overfitting the testing set by simplifying the model. Ridge Regression Regression tries to strike a balance between overfitting and underfitting as is expressed visually in the plots.

## Part D

A measure of performance in Linear Regression is Mean Squared Error (MSE). It measures the average of the squares of errors. The closer the value is to 0 the better the quality of the estimator has predicting values.

For the Linear Regression Model:

```
R_Squared = 0.763
MSE = 14.996
```

For the Ridge Regression Model:

```
R_Squared = 0.763
MSE = 14.775
```

Since ridge regression provides a lower MSE value we can conclude that it is a better quality loss function to model the given data at this test sample size. Hence there is an improvement in performance.

## Part E

*Linear Regression*

```
R-squared = 0.760
MSE = 14.996
```

*Ridge Regression*

```
R-squared = 0.763
MSE = 14.775
```

*Lasso Regression*

```
R-squared = 0.701
```

```
MSE = 18.645
```

Lasso Regression has a smaller R-squared value and a larger MSE value than both of the other models. The generalizability of the model to new data is therefore worse due to the lower R-squared value. In addition the performance of the model in this context is worse due to producing a large mean squared error, which speaks to the inability of the model to account for the variance of the data when compared to the other two models in this context.

## Part F

The following are the results from a test size of 0.5.

*Linear Regression*

```
R-squared = 0.690
MSE = 25.175
```

*Ridge Regression*

```
R-squared = 0.684
MSE = 25.632
```

*Lasso Regression*

```
R-squared = 0.690
MSE = 25.175
```

The following are the results from a test size of 0.9.

*Linear Regression*

```
R-squared = 0.676
MSE = 28.430
```

*Ridge Regression*

```
R-squared = 0.671
MSE = 28.430
```

*Lasso Regression*

```
R-squared = 0.676
MSE = 28.014
```

These sets of results ranging from test sizes of 0.1 (previous question), 0.5 and 0.9 show a changing relationship given different sized testing data. In terms of the two loss functions ridge regression has shown itself to produce more a generalizable model when given a small test set. Whereas linear regression is better when given a larger test set to train on in the context of this dataset. Therefore a model is only useful given a context where it excels.

# Task 2

```
Please see file 5_Logistic-Regression_Classifier.py
```