Foundations of Data Science Assignment-1  
Documentation

# Brief description of the model

## Bivariate Polynomial Regression

As the dataset has two independent variables (Strength and Temperature) and one dependent variable (Pressure), bivariate polynomial regression is an appropriate choice of model to predict the value of the dependent variable.

Bivariate Polynomial regression is a type of regression analysis in which the association between the independent variables x1 and x2 and the dependent variable y is represented as an nth degree polynomial in x1 and x2. The generic equation representing a bivariate polynomial equation of order n is given by:

y= b0 + b11x1 + b21x2 + b12x12 + b22x22 + b13x13+ b23x23+ ... + b1nx1n +b2nx2n

# Data pre-processing

As the scale of the independent variables is different, standardisation was carried out on these features before model fitting in order to bring them onto the same scale. Through the use of standardisation, we bring the values of the mean to be equal to zero and standard deviation to be equal to one.

Standardisation was carried out using the following formula:

The dataset was then shuffled and a train-test split of 70:30 was introduced in the dataset.

# Algorithms

## Gradient Descent

Gradient Descent is an optimisation algorithm through which a cost function can be minimised by iteratively moving in the opposite direction of the slope of the said function at a given point. Gradient Descent starts from a randomly selected point and iteratively goes down the slope to approach the local minima of the function, where the gradient, or the slope, is found to be approximately, or sometimes exactly, zero. The updating of the set of parameters is done using the expression:

<insert expression here>

Here, α is a hyper-parameter known as the learning rate, which determines the size of the steps taken in the process of updating the weights. A larger learning rate indicates that we take bigger strides towards the minimum, but it also means there is a chance of overshooting as well as oscillation around the minimum.

The hyperparameter values used for Gradient Descent are listed below:

Learning Rate:   
Epochs:  
Precision: ?

## Stochastic Gradient Descent

Stochastic Gradient Descent is a variant of the regular Gradient Descent described above which trades a little bit of performance for speed. Instead of updating after processing each data point, Stochastic Gradient Descent picks up a random point to ascertain the value of the loss function. Therefore, the set of parameters are updated using the expression:

<insert SGD expression>

α, like in Gradient Descent, is a hyper-parameter representing the learning rate. The hyperparameter values used for Stochastic Gradient Descent are listed below:

Learning Rate:   
Epochs:  
Precision: ?

Results

The regression models were fitted to the given data via the two methods listed above for varying values of the degree of the polynomial. The RMSE (Root Mean Square Error) values for each were found as such:

<table>

<observations?>

The most optimal model is obtained when the degree is <insert degree>

Surface Plots

<insert plots>

<comments on overfitting>

# Regularisation

Regularisation is a technique that is used to avoid the problem of overfitting by rewarding simpler fitting functions over complex ones. It works by introducing a penalty term to the cost function. Two prominent types of regularisation techniques are Ridge Regression and Lasso Regression.

Lasso Regression

Here, the sum of the absolute magnitudes of the weights is added as a penalty term to the cost function, to give us the modified cost function:

<insert equation>

This is also known as L1 regularisation. In lasso regression, the coefficients of the features which are less important get shrunk down to zero, thereby also acting as a feature selection technique.

Ridge Regression

Here, the sum of the squares of weights is added as a penalty term to the cost function, to generate the modified cost function to be:

<insert equation>

This is also known as L2 regularisation.

Results

The minimum training and testing error obtained post-regularization for each model and regression technique are listed as follows:

<insert table>

Plot of RMSE vs log λ

<insert plot>

From the graph as well as the data listed above the optimal model is obtained when λ = <insert value>.

Comparing models pre- and post-regularization