GNR652: Assignment 1

Housing Price Prediction using Regression

In this assignment, you are required to implement four variants of multi-variate Regression from scratch. Feel free to reuse code from the previous tutorial session. The dataset is provided as accompanying 'housingPriceData.csv' file where each row represents a data sample i.e., a house. Each house is provided with three attributes, namely, bedrooms, bathrooms, and sqft_living. The aim here is to predict the price of each house using these three attributes.

1. Linear and non-linear Regression

   a. Linear Regression

      To start off, we want to learn a linear model between the features and prices of the following type.

$$h_\beta(X) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

   b. Polynomial Regression

      Next, in this subproblem, we want to learn a non-linear model of the following type between the only two features (bedrooms and sqft_area) and the prices.

$$h_\beta(X) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2$$

   Here, $h(.)$ is our hypothesis, β_i's are the model parameters, and x_i's are the features.

2. Regularized Regression

   In these sub-problems we want to learn a linear model between the features and prices and improve the prediction performance using regularization.

   a. Ridge regression

      Ridge Regression or Tikhonov regularization is a method of regularization of ill-posed problems. In essence, it imposes an L_2 regularization on the model parameters. More details about Ridge regression are available in the course slides. For further reference, you may refer to: https://en.wikipedia.org/wiki/Tikhonov_regularization. Below is the cost function of Ridge regression:

$$J_{Ridge} = \sum_{i=1}^{n} (h_\beta(x_i) - y_i)^2 + \alpha \sum_{j=1}^{m} \beta_j^2$$

b. LASSO Regression

LASSO (Least Absolute Shrinkage and Selection Operator) is a another method that performs regularization. In essence, it imposes an L_1 regularization on the model parameters. More details about LASSO are available in the course slides. For further reference, you may refer to: https://en.wikipedia.org/wiki/Lasso_(statistics). Given below is the cost function for LASSO regression:

$$J_{Lasso} = \sum_{i=1}^{n} (h_\beta(x_i) - y_i)^2 + \alpha \sum_{j=1}^{m} |\beta_j|$$

Here, $J(.)$ is our cost function, β_i's are the model parameters, y_i's are ground truth, and x_i's are the features, and α is the regularization parameter. Also, n and m denote the number of samples and number of model parameters respectively.

Along with the learned model parameters, the following two performance metrics for evaluation of each model's performance has to be reported and compared.

1. Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{1}{n} (\hat{y_i} - y_i)^2}$$

2. R^2 Score

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y_i} - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Here, y_i's are the ground truth, $\hat{y}$'s are the predictions and $\bar{y}$ denote the mean y.

Submission: All submissions must follow the following guidelines:

- `All the solutions should be in the Julia programming language.`
- `No packages other than CSV.jl and Plots.jl are allowed.`
- `Submission must be a single zip file.`
- `Submission folder must contain only two folders named: src and data.`
- `All codes must be placed in src, and all data files should be in data.`
- `More details on the submission format will be available on the moodle upload link.`

Rubric: Each problem carries equal weightage. Total marks will be scaled to out of 15 for inclusion in the final performance record.