CYBR 486 - Lab #4: Polynomial Regression, Ridge, and Lasso Regularization

Overview

This lab focuses on building, training, and evaluating different regression models, including linear, polynomial, and regularized models (Ridge and Lasso), using Python and scikit-learn. You will work with a housing dataset to predict housing prices. Key steps include splitting the data, building and training various regression models, and evaluating their performance using metrics like Root Mean Square Error (RMSE) and R² score.

Objectives

1. Split the dataset into training and testing subsets (80/20).
2. Train a linear regression model on the training set.
3. Build a polynomial regression model with a specified degree.
4. Train a Ridge regression model using cross-validation to select the best regularization parameter.
5. Compare the performance of the models using RMSE and R² score.

Prerequisites

1. Python 3.x installed on your machine.
2. Required Python libraries:
   * scikit-learn
   * pandas
   * numpy To install the required libraries, run the following command:

bash

Copy code

pip install scikit-learn pandas numpy

Dataset

* File: housing.csv (or the dataset you are using)
* The dataset contains various features of houses (e.g., number of rooms, property age, etc.) and the corresponding prices. Ensure the dataset is placed in the same directory as the code file before running the script.

Usage Instructions

1. Clone the Repository  
   Clone this repository or download the code file and dataset to your local machine.
2. Run the Script  
   Execute the Python script (polynomial\_regression\_lab.py) using the command:

bash

Copy code

python polynomial\_regression\_lab.py

1. Outputs
   * Training and test dataset shapes will be printed in the console.
   * Model evaluation metrics (RMSE and R² score) will be displayed for each model (linear, polynomial, Ridge).
2. Save Results (Optional)  
   The evaluation metrics can be saved to a CSV file named model\_evaluation\_results.csv in the same directory.

Code Structure

1. Step 1: Import necessary libraries (scikit-learn, pandas, numpy, etc.).
2. Step 2: Load and preprocess the dataset (check for nulls, separate features and target).
3. Step 3: Split the data into training and testing subsets.
4. Step 4: Train the linear regression model and make predictions.
5. Step 5: Train the polynomial regression model for different degrees and evaluate.
6. Step 6: Train the Ridge regression model using cross-validation and evaluate.
7. Step 7: Compare the performance of the models using RMSE and R² score.

Evaluation Metrics

1. Root Mean Squared Error (RMSE): A measure of the average magnitude of the errors, where lower values indicate better fit.
2. R² Score: The proportion of variance explained by the model, with values closer to 1 indicating a better model fit.

Observations

* Linear Regression: Captures simple relationships but may struggle with more complex patterns in the data.
* Polynomial Regression: Models more complex relationships but may lead to overfitting at higher degrees.
* Ridge Regression: Regularizes the polynomial model, preventing overfitting and maintaining strong performance.