HY573 - Assignment #2:
Regression with Missing Data

Submission Guidelines

- **Due date: 16/12/2023**
- Submission via e-learn
- Provide one zip file with the written solutions and the code.
  - Provide a report discussing the different topics.
  - Incorporate your code, visualizations, and intermediate outputs in a Jupyter notebook.

Overview

In this assignment, you are required to apply regression analysis techniques to the Boston Housing dataset. You must manually implement these models using foundational libraries like NumPy and Pandas, **not** high-level functions from libraries like scikit-learn. The assignment encompasses several stages of regression modeling, from unconstrained least squares regression to Lasso regression, and approaches for handling missing data based on low-rank matrix completion.

Dataset

The Boston Housing dataset (that can be downloaded from learn (housing.csv)), includes 13 features such as crime rate, property tax rate, average number of rooms, among others, and the target variable is the median value of owner-occupied homes (14th column).

Exercise 1: Unconstrained Least Squares Regression

Objective: Implement a linear regression model using the least squares method.
Tasks:

- Conduct exploratory data analysis to understand the dataset's characteristics, including distributions, and potential outliers.
- Develop a least squares regression model without regularization to predict the median house value (MEDV).
- Analyze the model's performance using the Root Mean Squared Error on a test set.

Exercise 2: Lasso Regression

Objective: Understand the impact of L1 regularization in regression.
Tasks:

- Implement Lasso regression to incorporate feature selection in your model.
- Tune the regularization parameter for optimal model performance.
- Compare the outcomes with the least squares model in terms of feature selection and model performance.

Exercise 3: Handling Missing Data

Objective: Evaluate the effect of missing data on regression models.
Tasks:

- Systematically introduce missing values into the dataset. Consider different missing percentages and patterns.
- Apply the least squares and Lasso regression models to the modified dataset.
- Observe and analyze the impact on model performance and interpretability.

Exercise 4: Matrix Completion

Objective: Implement a method to impute missing data in the dataset.
Tasks:

- Develop and apply matrix completion through nuclear norm minimization.
- Discuss the method's effectiveness in recovering missing data.

Exercise 5: Re-evaluation Using the Completed Dataset

Objective: Assess the performance of regression models on the dataset after matrix completion.
Tasks:

- Reapply least squares and Lasso regression models on the completed dataset.
- Compare and analyze the performance against the models trained on the dataset with missing values.

Deliverables

Comprehensive Report: Submit a single detailed report covering:

- Methodologies, implementations, and results for each exercise.
- Analytical insights and interpretations from each stage.
- Supporting visualizations, code snippets, and statistical analyses.
- Comparative analysis of model performances and methodologies across exercises.