Lab Assignment II

This lab exercise will help students to gain in depth knowledge about when linear and ridge regression is used.

Exploring Linear and Ridge Regression with Cross-Validation

Objective: In this project, students will explore the concepts of Linear Regression and Ridge Regression using real-world datasets from online repositories. The goal is to understand the strengths and weaknesses of both models and explore methods to improve their performance, including cross-validation for model evaluation.

Project Tasks:

1. Dataset Selection

- Students should select a dataset from online repositories such as Kaggle, UCI Machine Learning Repository, or any online repository.
- The dataset must include both numerical and categorical features, with a continuous target variable suitable for regression.

2. Data Preprocessing

- Handle missing values using appropriate techniques (e.g., imputation).
- Encode categorical variables using One-Hot Encoding or Label Encoding.
- Scale numerical features to ensure better model performance.
- Split the dataset into training and testing sets.

3. Model Training and Evaluation

- Train a Linear Regression model and analyse its performance.
- Train a Ridge Regression model with different values of alpha.
- Implement k-fold Cross-Validation to evaluate both models and reduce variance in performance estimates.

4. Analysis and Comparison

- Compare the models using Mean Squared Error (MSE) and R^2 score.
- Analyse cases where Linear Regression overfits and where Ridge Regression helps prevent overfitting.

5. Improving Model Performance

- Experiment with different alpha values in Ridge Regression and observe their effect.
- Discuss practical applications where each model would be more suitable.

Expected Outcomes:

- A detailed video discussing the findings, including model performance, strengths and weaknesses, and potential improvements.
- A Python implementation with code and visualizations demonstrating the results including Linear and Ridge Regression and the role of cross-validation in model evaluation.

Submission Guidelines:

1. Video Presentation
   - Students must create a video presentation of maximum 5 min explaining their project.
   - The video should cover relevant code details, insights, and key findings.
   - Each team member must present their individual contribution.
   - The video should include a live recording of each presenter with their face overlayed on the slides while explaining their part.
2. GitHub Repository
   - Submit a GitHub link containing the full project.
   - The repository must include:
     o Clean and well-commented code
     o README file explaining the project, dataset, and steps taken and insights
3. Individual Contributions
   - Clearly highlight the contributions of each team member in both the GitHub repository and the video.
   - Each person should explain their part in the project during the video presentation.
4. Submission Format
   - Submit the final video link along with the GitHub repository link.
   - Ensure the video is clear, well-structured, and professional.

Important note on this lab exercise:

- The deadline for the submission: **Week 9, Sunday 19th October**. The link for the submission will be enabled in NTULearn.
- Only one person from the group need to submit the github link and video link/ video. If you are submitting any links, please make sure that it is given the access right to view by us.
- *We believe and trust every NTU student can behave and act responsibly or participate fairly when working on this exercise (NTU Student Code of Conduct - Truth, Trust, & Justice).*