CS422/622- HW 3

Regression Task with Linear Regression Model Implementation

In HW3, you will develop a Regression pipeline using the Linear Regression model. By completing this task, you'll gain hands-on experience with data preprocessing, model training, and evaluation—key steps in machine learning. Follow the guidelines below to ensure structured and successful implementation.

1. Dataset

- Download data from https://archive.ics.uci.edu/ml/datasets/auto+mpg, carefully check the dataset information on the website.

2. Data Preprocessing

- **Feature Selection**: Choose the most relevant features that contribute to the classification task.
- **Normalization**: Normalize the features if necessary to ensure they are on the same scale, especially if they have different units or ranges.
- **Missing Values**: Address any missing data by either removing or imputing the samples with missing values.

3. Model Training

- **Coefficients**: Compute the optimal set of coefficients for the model that minimizes prediction error (use RMSE defined below).

4. Testing and Model Evaluation

- **RMSE**: Evaluate the model's performance using the following formula describing Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \mathbf{X_i b})^2}{N}},$$

where y_i is a ground truth value of the i-th test data, $\mathbf{X_i b}$ is a prediction from the linear regression model, and N is a number of test data.

- **10-fold cross validation:** Perform 10-fold cross validation and fill out the following table, calculating a coefficient for each feature then RMSE.

	Feature 1	Feature 2	Feature 3	...	Feature j	RMSE
Fold 1						
Fold 2						
Fold 3						
⋮						
Fold 10						

Report

- Your report should provide a clear and structured explanation of each step, from dataset selection to evaluation. Discuss not only what you did, but why you made specific choices.
- Include test codes and execution results along with screenshots of your results for better clarity and visualization.
- Describe what you learned from this assignment—focusing on insights about Linear Regression, data preprocessing, or any challenges faced during implementation.
- Feel free to describe anything else you want, such as additional observations, interesting findings, or how you might approach the task differently in the future.
- **There are no restrictions on using external libraries,** but make sure to document the libraries you used and their purposes in the report.

Additional requirements for graduate students in CS 622

Graduate students are expected to:

- Use **Ridge Regression** to limit the unexpected growth (overfitting) of coefficients.

Submission instructions:

You must submit the followings to UNLV WebCampus:

1. A report file
2. Source code file(s)
   - Must be well organized (function name, indentation, …)
   - **You need to upload the python text file (*.py.txt). Simply add ".txt" to the py extension. Don't upload jupyter notebook files**

You must submit the files SEPERATELY. DO NOT compress into a ZIP file. If you fail to provide all required information or files, you may be given zero score without grading.

Once you submit, Webcampus will perform similarity check for your submission and show you the result. Your similarity score must be lower than 50% unless something essential is described in the report. Otherwise, (the score -50%) will be deducted. Detecting any attempts to bypass the similarity check may result in receiving zero points.