Victor MAYAUD
Hichem KHETTAB

Project 3: Ridge Regression

Task 1 :

The RidgeRegression class, defined in ridge.py, is designed to work with any dataset. This class includes five methods:

-**__init__**: Initializes the class. Sets the lambda value and initializes the weights based on user input.

-**fit**: Takes the dataset X and output y as inputs. This method calculates the weights using this formula:

$$\hat{\mathbf{w}} = (\mathbf{X}^T \mathbf{X} + \lambda^2 \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$$

The first coefficient of the identity matrix I is zero to exclude the bias from regularization. In order to compare my result with **sklearn** I just multiply the identity matrix by lambda and not square of lambda.

-**predict**: Returns the predicted output.

-**loss_function**: Calculates the loss function, incorporating regularization.

-**L2_regularization**: Computes the L2 regularization, multiplied by the lambda value.

Task 2 :

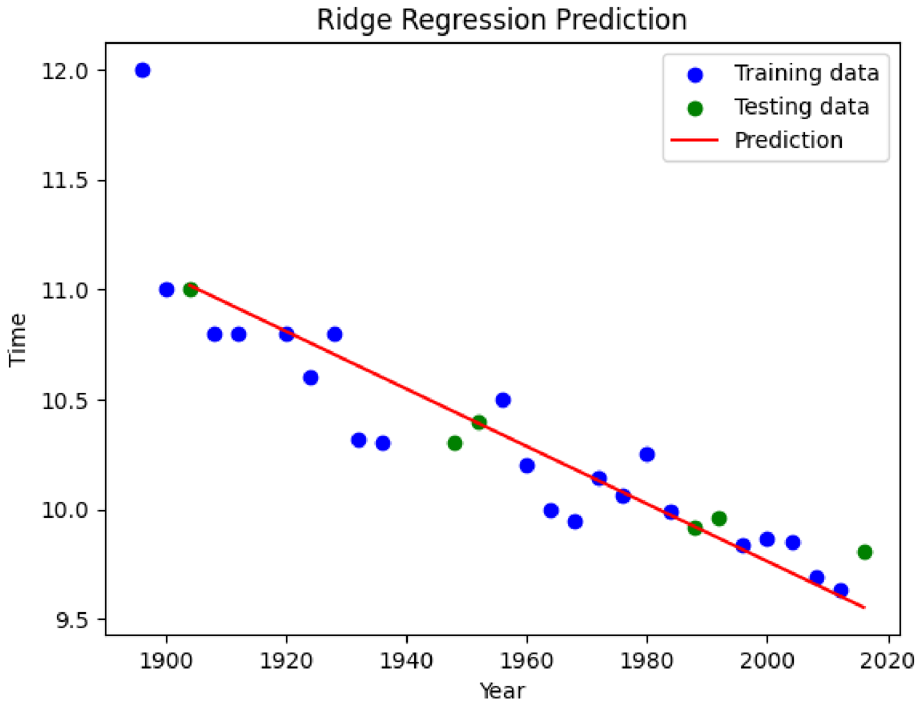
Once the RidgeRegression class is implemented, we can apply ridge regression to datasets like the Olympics 100m. However, determining the most efficient lambda value for optimal results across all datasets is challenging. To identify the best lambda, I performed Leave-One-Out (LOO) cross-validation, particularly suitable for small datasets. LOO cross-validation involves:

-Iteratively leaving out one data point from the dataset, using the remaining data for training the model, and then using the excluded data point for validation.
-Repeating this process for each data point in the dataset, thus ensuring that each one is used exactly once as the validation set.
-For each iteration, the model is evaluated based on how well it predicts the left-out data point, typically using a metric like Mean Squared Error (MSE).
-After cycling through all data points, the average of these evaluations is computed to estimate the model's overall performance.
-This procedure is repeated for each candidate lambda value. The lambda yielding the lowest average error is selected as the best one, balancing the model's bias and variance effectively.

After applying my LOO cross-validation to a broad range of lambda values, the optimal lambda identified was 231, resulting in a Mean Squared Error (MSE) of 0.01582 on the test dataset, which was previously partitioned from the initial dataset. Visualizing the prediction line on a graph, labeled as **A-1**, shows a good fit with the dataset. Subsequently, I compared my results with those obtained using scikit-learn's implementation on the same dataset. Remarkably, the outcomes of the best lambda and the mean square errors were identical. This consistency can be attributed to scikit-learn employing a similar cross-validation approach to determine the best lambda value.

Annexe:

A-1:



Ridge Regression Prediction
Training data
Testing data
Prediction
12.0
11.5
11.0
10.5
10.0
9.5
Time
1900
1920
1940
1960
1980
2000
2020
Year