ECGR 4105: Intro to Machine Learning HW 2

Axel Leon Vasquez

Student ID: 801182414

September 16, 2024

Link to GitHub

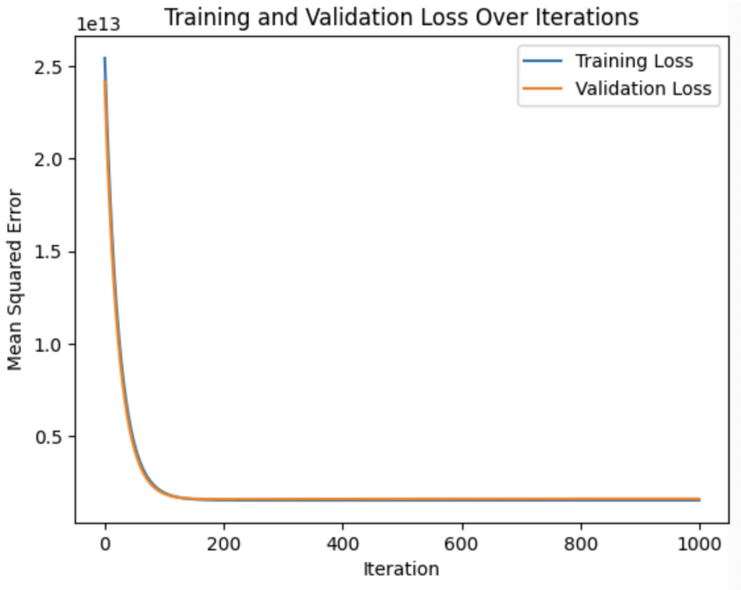
Problem 1a

1. **Introduction:**
   This task aimed to develop a linear regression model using gradient descent from scratch to predict housing prices. The model is trained using a set of input variables including:
   - Area (sqft)
   - Number of bedrooms
   - Number of bathrooms
   - Number of stories
   - Parking spaces available

2. **Methodology:**
   The model was developed using a selected set of input variables: area, bedrooms, bathrooms, stories, and parking. The dataset was standardized and split into training (80%) and validation (20%) sets, with a bias term added to account for the intercept in the linear regression model. Gradient descent was implemented with a learning rate of 0.01 over 1,000 iterations, allowing the model to iteratively update its parameters by minimizing the loss function.

3. **Results:**
   The following graph shows the training and validation loss over 1,000 iterations. The losses were calculated using the Mean Squared Error (MSE) for the training and validation sets. The final optimal values for the model parameters were found to be:

Input Variable	Theta
Area	729,932.49
Bedrooms	79,337.73
Bathrooms	641,263.77
Stories	463,853.77
Parking	287,330.72



4. **Summary of Findings**
   The analysis developed a gradient descent algorithm to predict housing prices based on features such as area, bedrooms, bathrooms, stories, and parking. The training and validation losses consistently decreased over 1000 iterations, indicating effective learning and generalization of the model. The final optimal parameters revealed that area (729,932.49) and bathrooms (641,263.77) significantly influence housing prices, while stories (463,853.77), bedrooms, and parking also contribute, albeit to a lesser extent. These findings suggest that key features, particularly areas and bathrooms, should be prioritized in real estate evaluations.

Problem 1b

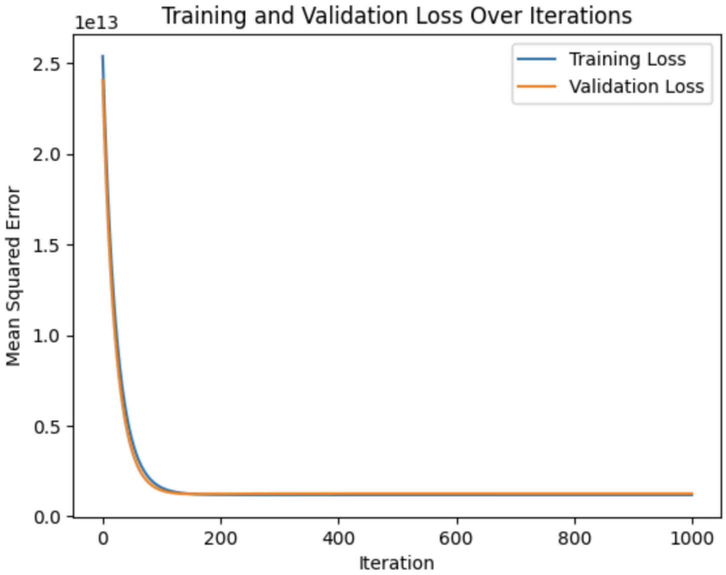
1. **Introduction:**
   This task aimed to develop a linear regression model using gradient descent from scratch to predict housing prices. The model is trained using an expanded set of input variables, including:
   - Area (sqft)
   - Number of bedrooms
   - Number of bathrooms
   - Number of stories
   - Main road accessibility
   - Guestroom availability
   - Basement presence
   - Hot water heating
   - Air conditioning
   - Parking spaces
   - Preferred area

2. **Methodology:**

The input variables were standardized along with the housing price target variable. The dataset was split into training (80%) and validation (20%) sets, with a bias term added to the features. Gradient descent was implemented with a learning rate of 0.01 over 1,000 iterations.

3. **Results:**
The final theta values for the linear regression model, along with the corresponding input variables, are presented in the table below:

Input Variable	**Theta**
Area	592965.81
Bedrooms	95825.54
Bathrooms	593538.86
Stories	343100.13
Main Road	237368.13
Guestroom	132563.88
Basement	139134.95
Hot Water Heating	222898.82
Air Conditioning	428014.59
Parking	253089.41



4. **Summary of Findings**

The analysis developed a gradient descent algorithm to predict housing prices using a comprehensive set of features, including area, bedrooms, bathrooms, stories, and various amenities. Training and validation losses decreased consistently over 1000 iterations, indicating effective learning. The final parameters showed that area (4,795,729.21) and bathrooms (593,538.86) significantly influence housing prices, with other features contributing to a lesser extent. Compared to Problem 1a, which had a more limited feature set, the broader model in Problem 1b revealed different magnitudes for parameter influences, highlighting the importance of including additional variables for improved predictive accuracy.

Problem 2a

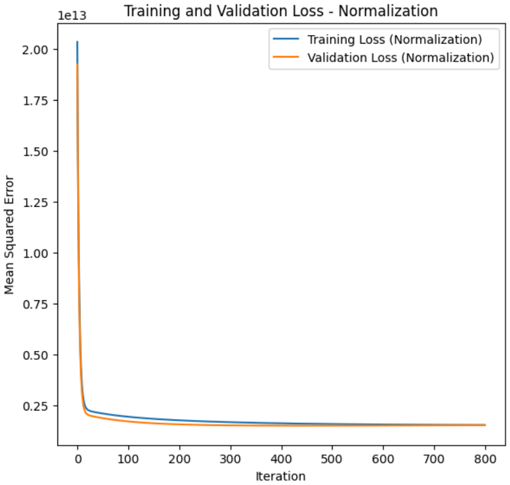
1. **Introduction:**
   This task involved repeating the analysis from Problem 1a to predict housing prices while implementing input normalization and standardization as part of the preprocessing. The objective was to assess how these scaling methods affect model performance by training the linear regression model with both approaches and comparing the results.

2. **Methodology:**
   The dataset was split into training (80%) and validation (20%) sets. Two separate scaling methods were applied: input standardization using StandardScaler and input normalization using MinMaxScaler. A bias term was added to the features, and gradient descent was executed with a learning rate of 0.05 for 800 iterations. The training and validation losses were recorded for each scaling method, enabling a comparison of model performance.

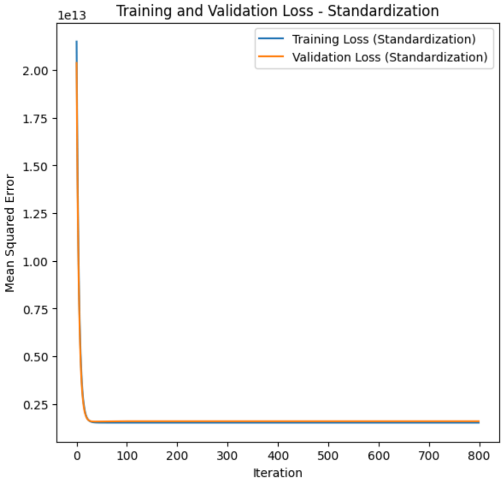
3. **Results:**
   The final training and validation losses for both scaling methods were as follows:

Scaling Method	**Final Training Loss**	**Final Validation Loss**
Normalization	1,542,771,702,725.91	1,532,969,136,956.66
Standardization	1,522,069,682,826.88	1,596,534,488,482.43

   a. **Normalization:**



b. Standardization:



4. **Summary:**
This task focused on comparing training loss between standardization and normalization, indicating that lower training loss typically reflects better model fit. However, significantly higher validation loss could suggest overfitting. Standardization generally performs better for algorithms like gradient descent, especially when features have different units. In contrast, normalization is advantageous for datasets with a specific range and minimal extreme outliers.

Problem 2b

1. **Introduction:**
This task aimed to investigate the impact of input normalization and standardization on the training of a linear regression model to predict housing prices. The analysis involved training the model using two distinct preprocessing techniques, with the dataset split into training (80%) and validation (20%) sets. The binary features such as main road access and amenities were converted into numerical values for effective modeling.

2. Methodology:

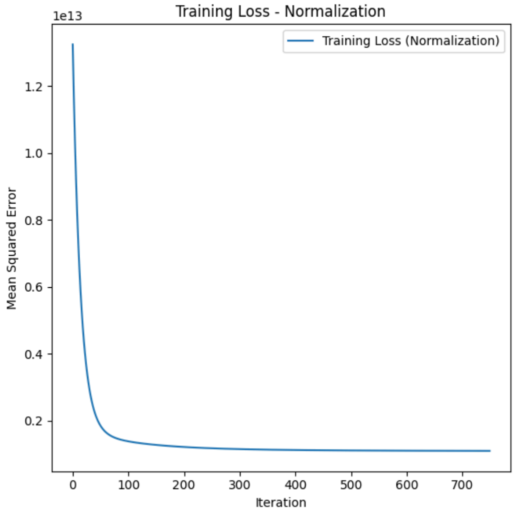
The dataset was preprocessed by applying binary mapping to categorical features and then splitting it into training and validation sets. Two separate trainings were conducted using gradient descent: one with input normalization and another with input standardization. The model was trained over 750 iterations and a learning rate of 0.10.

3. Results:

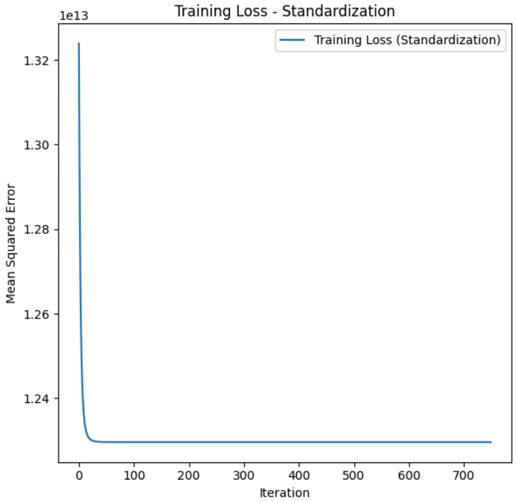
The final training and validation losses for both scaling methods were as follows:

Scaling Method	**Final Training Loss**	**Final Validation Loss**
Normalization	2,206,383,216,211.09	2,128,972,116,184.02
Standardization	24,593,080,356,214.45	23,752,561,176,710.90

a. Normalization:



b. Standardization:



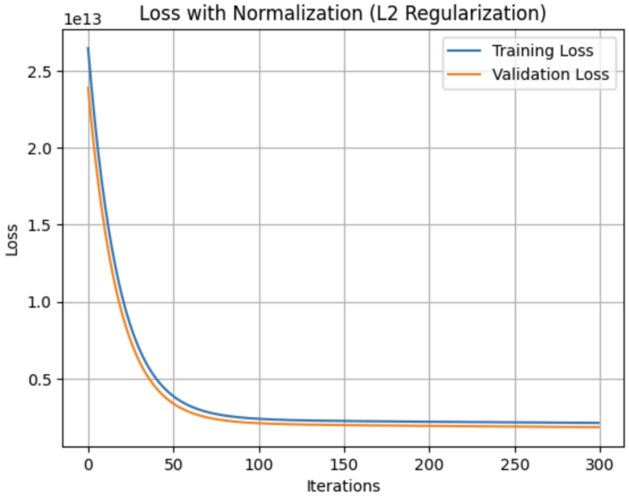
4. **Summary of Findings**
   The analysis highlighted that normalization produced significantly lower training and validation losses compared to standardization. Specifically, the final training loss with normalization was 2,206,383,216,211.09, while standardization yielded a training loss of 24,593,080,356,214.45. These results suggest that normalization is more effective for this dataset, allowing the model to learn better from the input features without being influenced by the magnitude of different scales

Problem 3a

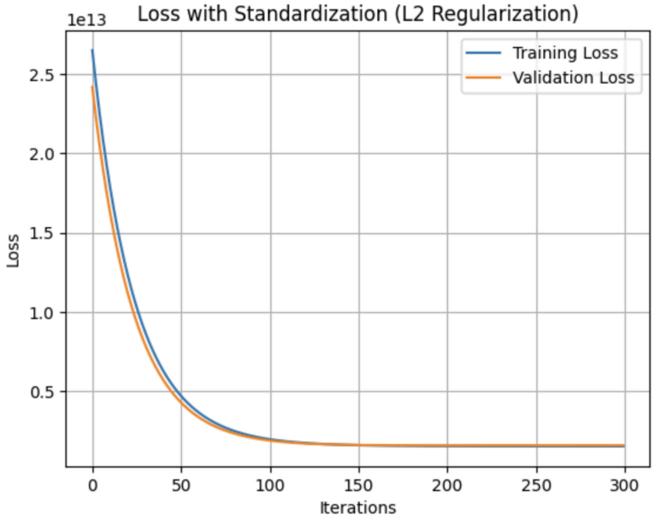
1. **Introduction:**
   In this task, the objective was to analyze the impact of L2 regularization on the training and validation losses of a linear regression model for predicting housing prices. The model was trained using two different scaling methods: normalization and standardization. Regularization was incorporated into the loss function to prevent overfitting.

2. **Methodology**
   The dataset was preprocessed by using both normalization and standardization techniques, with the features transformed accordingly. The training data was split into training (80%) and validation (20%) sets. L2 regularization was added to the cost function, modifying the gradient descent logic while keeping the evaluation loss computation unchanged. The model was trained over 300 iterations with a learning rate of 0.01, and a regularization strength (λ) of 5 was applied.

3. **Results:**
   a. **Normalization with L2 Regularization:**



   b. **Standardization with L2 Regularization:**



4. **Summary of Findings**
   a. **Comparison with Problem 2a:**
   In Problem 2a, without regularization, the model experienced a substantial risk of overfitting, indicated by the widening gap between training and validation losses. However, the introduction of L2 regularization in this task mitigated that risk, leading to more balanced training and validation losses across both scaling methods.
   b. **Conclusion:**
   Both normalization and standardization methods yielded effective results with the inclusion of L2 regularization. While both methods reduced overfitting risks, the model using normalization exhibited slightly better performance in terms of lower validation loss. The analysis underscores the importance of regularization in enhancing model robustness, particularly when handling datasets prone to overfitting.

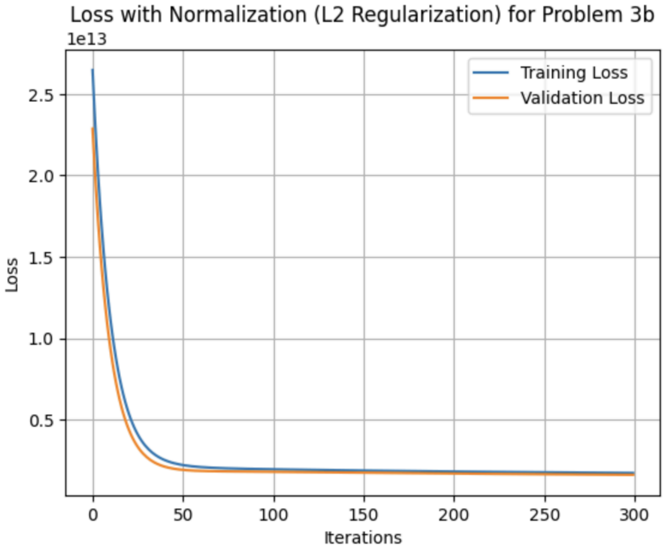
Problem 3b

1. **Introduction:**
   In this task, the goal was to analyze the impact of L2 regularization on the training and validation losses of a linear regression model for predicting housing prices, extending the analysis from Problem 2b. The model utilized multiple input features, and regularization was integrated into the loss function to mitigate overfitting.

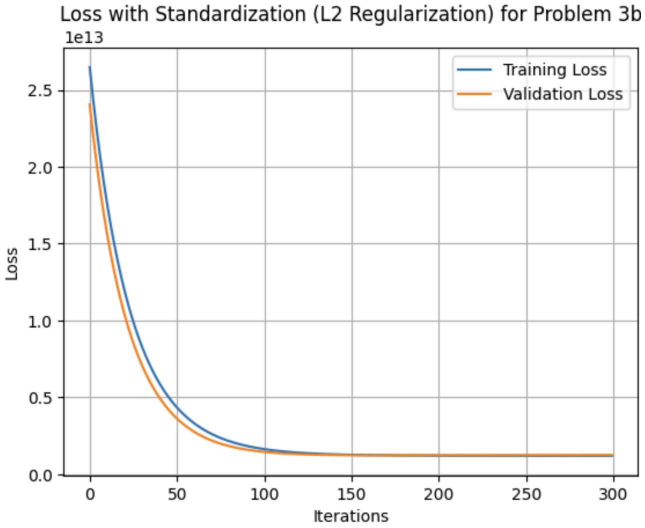
2. **Methodology:**
   The dataset was split into training and validation sets, using features including area, bedrooms, bathrooms, stories, and binary variables indicating the presence of amenities. Both normalization and standardization techniques were applied to the input features, and L2 regularization was included in the cost function. The model was trained over 300 iterations with a learning rate of 0.01 and a regularization strength (λ) of 5.

3. **Results:**

   a. **Normalization with L2 Regularization:**



   b. **Standardization with L2 Regularization:**



4. **Summary of Findings**

   a. **Comparison with Problem 2b:**

   In Problem 2b, the model did not use regularization, which resulted in a noticeable gap between training and validation losses, indicating overfitting. In contrast, the introduction of L2 regularization in this task helped reduce that gap, leading to a more stable model with lower validation loss and improved generalization capabilities. The results highlight the effectiveness of regularization in preventing overfitting in linear regression models.

   b. **Conclusion:**

   Both normalization and standardization methods resulted in improved training and validation losses with the incorporation of L2 regularization. While both methods were effective, normalization displayed a slight edge in terms of final validation loss. This analysis underscores the critical role of regularization in enhancing model performance, especially when dealing with complex datasets that may lead to overfitting.