NYCU Introduction to Machine Learning, Homework 1

109350008, 張詠哲

Part. 1, Coding (50%):

(10%) Linear Regression Model - Closed-form Solution

1. **(10%) Show the weights and intercepts of your linear model.**

```
Closed-form Solution
Weights: [2.85817945 1.01815987 0.48198413 0.1923993 ], Intercept: -33.78832665744869
```

(40%) Linear Regression Model - Gradient Descent Solution

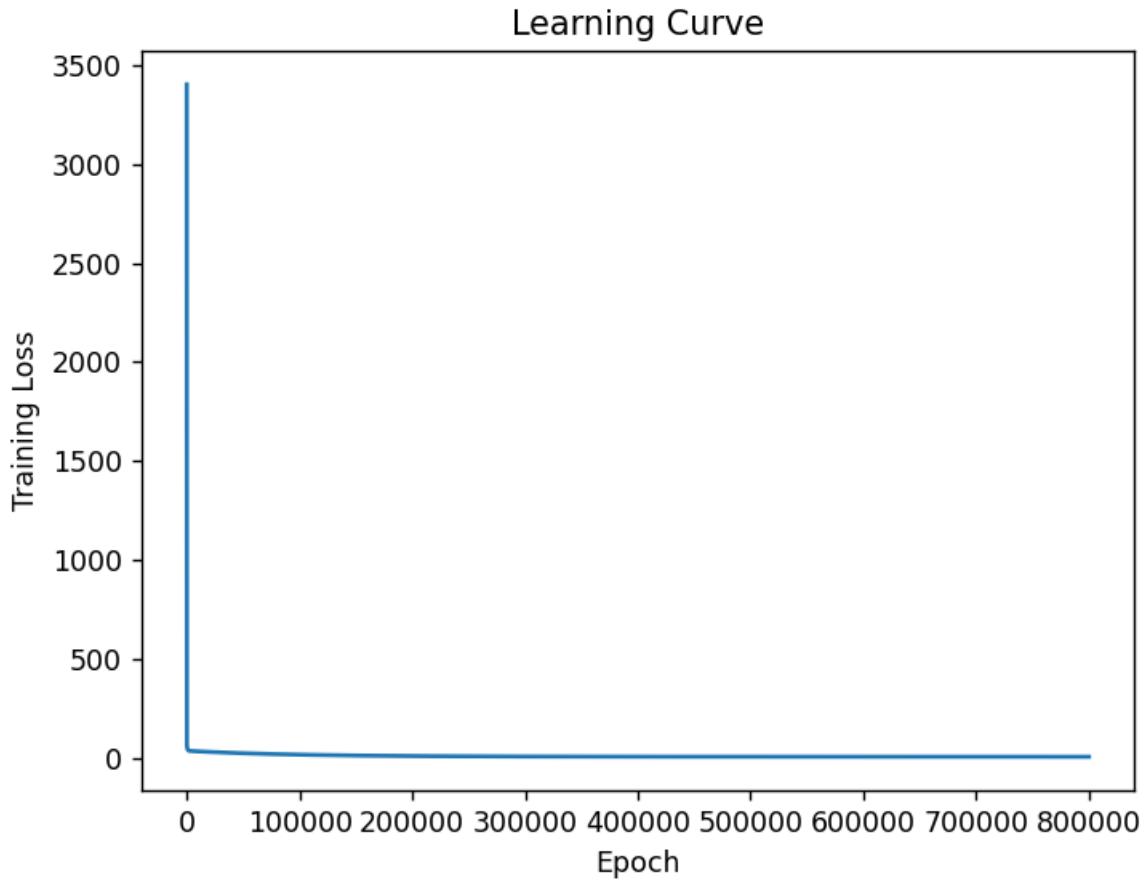
2. **(0%) Show the learning rate and epoch (and batch size if you implement mini-batch gradient descent) you choose.**

```
LR.gradient_descent_fit(train_x, train_y, lr=1e-4, epochs=800000)
```

3. **(10%) Show the weights and intercepts of your linear model.**

```
Gradient Descent Solution
Weights: [2.84865294 1.01517722 0.45257341 0.18545821], Intercept: -33.29703250166867
```

4. **(10%) Plot the learning curve. (x-axis=epoch, y-axis=training loss)**



5. **(20%) Show your error rate between your closed-form solution and the gradient descent solution.**

```
Error Rate: 0.1%
```

Part. 2, Questions (50%):

1. **(10%) How does the value of learning rate impact the training process in gradient descent? Please explain in detail.**

   The learning rate has a significant impact on the convergence, speed, and stability of the model during the training process. Both too large and too small values can lead to poor model performance.

   - **When the learning rate is too small:**
     It results in small updates to the weights, which can help the model converge to the minimum of the loss function more precisely. However, the drawback is that smaller updates require a larger number of iterations to converge (consuming more resources). It can also increase the risk of converging to a local minimum solution.

   - **When the learning rate is too large:**
     It speeds up convergence and may escape from local minimum solutions. However, excessively fast convergence can lead to large jumps in the search points (oscillations), preventing the model from converging effectively.

2. **(10%) There are some cases where gradient descent may fail to converge. Please provide at least two scenarios and explain in detail.**

   - **When the learning rate is too large:**

   A too-large learning rate leads to excessively large steps in parameter updates. During each iteration, the search point may overshoot the global minimum, rebound, oscillate back and forth, and ultimately fail to converge to either a local or global minimum. This prevents stable convergence.

   - **Ill-Conditioned Cost Function:**

   An ill-conditioned cost function refers to a situation where the cost function exhibits regions that are either very steep or very flat. As a result, some areas may have extremely large gradients, while others have very small ones. This can lead to a tendency to diverge away from the minimum in steep regions and slow convergence in flat regions

3. **(15%) Is mean square error (MSE) the optimal selection when modeling a simple linear regression model? Describe why MSE is effective for resolving most linear regression problems and list scenarios where MSE may be inappropriate for**

data modeling, proposing alternative loss functions suitable for linear regression modeling in those cases.

In simple linear regression, MSE (Mean Squared Error) is often considered one of the best choices for modeling. This is because of its advantages:

- **Ease of Interpretation:**
  MSE measures the average of the squared differences between predictions and ground truth, making it easy to understand.

- **Continuity and Differentiability:**
  MSE is a continuous and differentiable function, which makes it easy to implement and compute during gradient descent.

- **Maximum Likelihood Estimate:**
  In certain situations, MSE can be viewed as the result of a maximum likelihood estimate, giving it an advantage in statistical inference.

However, there are cases where MSE may not be suitable:

- **Extreme Outliers:**
  When dealing with extreme outliers, MSE can be heavily influenced by these points since it computes the squared average differences. An alternative is to use the Huber loss function, which uses squared loss for small errors and linear loss for large errors, making it robust to outliers.

- **Differing Data Point Importance:**
  When data points have varying importance, and different weights need to be assigned to them, MSE may not be flexible enough. Weighted Least Squares (WLS) can be used as it adjusts the MSE formula to weight errors based on data point importance.

- **Heteroscedasticity:**
  In cases where the variance of errors is not constant but depends on the magnitude of predictions, MSE may not be suitable. This situation violates one of the fundamental assumptions of linear regression (constant error variance). Alternatives include using WLS or the logarithmic loss.

4. **(15%) In the lecture, we learned that there is a regularization method for linear regression models to boost the model's performance. (p18 in linear_regression.pdf)**

$$E_D(\mathbf{w}) + \lambda E_W(\mathbf{w})$$

4.1. (5%) Will the use of the regularization term always enhance the model's performance? Choose one of the following options: "Yes, it will always improve," "No, it will always worsen," or "Not necessarily always better or worse."

Not necessarily always better or worse.

4.2. We know that λ is a parameter that should be carefully tuned. Discuss the following situations: (both in 100 words)

4.2.1. (5%) Discuss how the model's performance may be affected when λ is set too small. For example, λ=10^(-100) or λ=0

When λ is set too small, the regularization term's effect is virtually insignificant. This situation poses a high risk of overfitting, where the model fits the training data so closely that it becomes overly sensitive to noise and fluctuations, impairing its capacity to generalize to new, unseen data. While the model may exhibit strong performance on the training dataset, its performance on the test dataset can deteriorate significantly, indicating a lack of generalization ability. Moreover, under the influence of a very small λ, the model tends to revert to a conventional linear regression model, forfeiting the advantages conferred by regularization. Therefore, the selection of an appropriate λ assumes paramount importance in striking the right balance between model complexity and the prevention of overfitting, ultimately facilitating robust generalization performance.

4.2.2. (5%) Discuss how the model's performance may be affected when λ is set too large. For example, λ=1000000 or λ=10^100

When λ is set too large, the impact of the regularization term becomes substantial. The model becomes highly constrained, with its weights being forcefully compressed toward values close to zero. This results in an excessive simplification of the model, often causing it to degenerate into a model that essentially predicts a constant or average value. Consequently, the model loses its ability to effectively capture the essential features within the data, leading to a condition known as underfitting. Underfitting results in the model's inability to provide meaningful and accurate predictions. Thus, the choice of λ should strike a balance between regularization strength and model complexity for optimal performance.