NYCU Introduction to Machine Learning, Homework 1

110550014, 吳權祐

The screenshot and the figures we provided below are just examples. The results below are not guaranteed to be correct. Please make sure your answers are clear and readable, or no points will be given. Please also remember to convert it to a pdf file before submission. You should use English to answer the questions. After reading this paragraph, you can delete this paragraph.

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

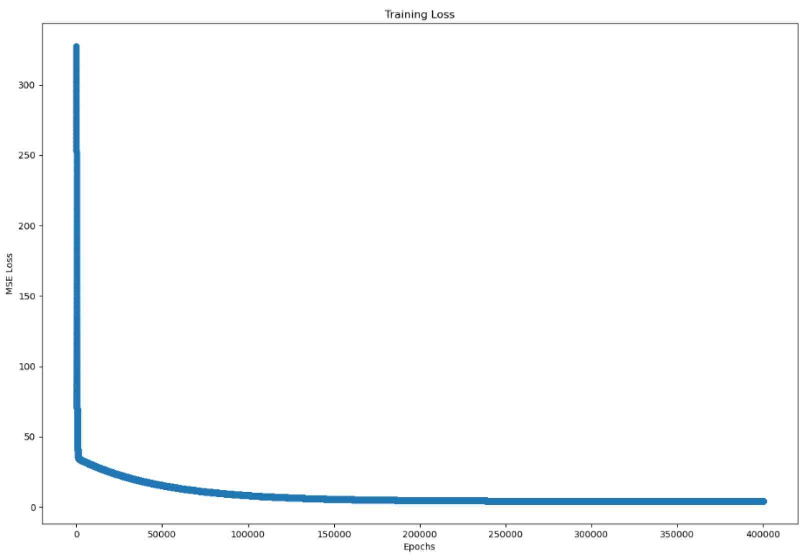
2. (0%) Show the learning rate and epoch (and batch size if you implement mini-batch gradient descent) you choose.

```
LR.gradient_descent_fit(train_x, train_y, lr=0.0001925, epochs=400000)
```

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

```
Gradient Descent Solution
Weights: [2.84688411 1.01462342 0.44711261 0.18416944], Intercept: -33.20581206434713
```

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 determines the speed of gradient descent. If the learning rate is too small, the descent may be excessively slow, or it may encounter difficulties. On the other hand, if the learning rate is too large, it can lead to gradient exploding.

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

   1. The learning rate is too high : May lead to gradient exploding and cause the weights bec oming overflow.
   2. There are flat or plateau regions in the loss function : This case would cause the converg ence speed to become slow, or more seriously, the algorithm to get stuck

3. (15%) Is mean square error (MSE) the optimal selection when modeling a simple linear re gression model? Describe why MSE is effective for resolving most linear regression probl ems and list scenarios where MSE may be inappropriate for data modeling, proposing alter native loss functions suitable for linear regression modeling in those cases.

   1. MSE is the optimal selection when modeling the simple linear regression model. Becaus e it aligns with the principle of Maximum Likelihood Estimation (MLE) under the assumpt ions of normally distributed errors.
   2. If there are some outliers in the data, which means the errors deviate from the normal di strbution principle, the MSE loss function may not be effective. The Huber loss function m ay be more suitable in this case.

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 performan ce? Choose one of the following options: "Yes, it will always improve," "No, it wil l 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

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

4.2.1 This would weaken the regularization effect and may causing the problem that the model fit to all features. And get a overfitting model which could not predict the true value in the test dataset.
4.2.2 This would result in excessive regularization. And make the model less effective at capturing the features, lead to a underfitting model which is worse to predict the value ,too.