NYCU Introduction to Machine Learning, Homework 1

[111550108], [吳佳諭]

Part. 1, Coding (60%):

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

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

```
2024-09-26 17:38:50.932 | INFO     | __main__:main:102 - LR_CF.weights=array([2.8491883 , 1.0188675 , 0.4
8562739, 0.1937254 ]), LR_CF.intercept=-33.8223
```

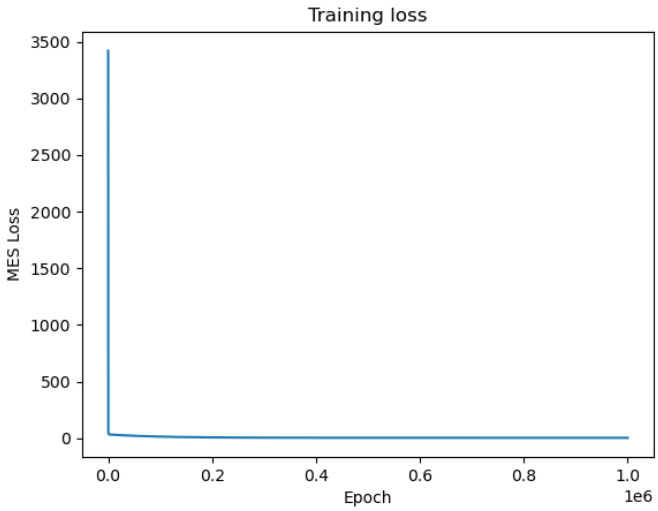
(40%) Linear Regression Model - Gradient Descent Solution

2. (10%)
   - Show the hyperparameters of your setting (e.g., learning rate, number of epochs, batch size, etc.).
   - Show the weights and intercepts of your linear model.

```
losses = LR_GD.fit(train_x, train_y, learning_rate=1e-4, epochs=1000000)
```

```
2024-09-26 17:42:04.310 | INFO     | __main__:main:109 - LR_GD.weights=array([2.84575631, 1.01779039, 0.4
7489042, 0.19126502]), LR_GD.intercept=-33.6441
```

3. (10%) Plot the learning curve. (x-axis=epoch, y-axis=training loss)

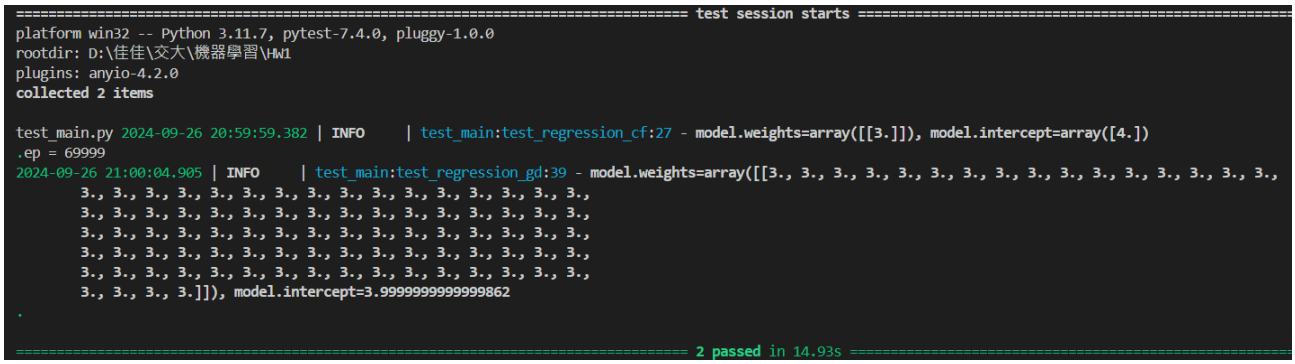


4. (20%) Show your MSE.cf, MSE.gd, and error rate between your closed-form solution and the gradient descent solution.

```
2024-09-26 17:42:04.407 | INFO     | __main__:main:118 - Mean prediction difference: 0.0238
2024-09-26 17:42:04.408 | INFO     | __main__:main:123 - mse_cf=4.1997, mse_gd=4.1978. Difference: 0.047%
```

(10%) Code Check and Verification

5. (10%) Lint the code and show the PyTest results.



Part. 2, Questions (40%):

1. (10%) How does the presence of outliers affect the performance of a linear regression model? How should outliers be handled? List at least two methods.

Outliers will make the model's performance bad. In order to fit the outliers, the model cannot perfectly fit the normal data. That is, the model always wants to minimize the MSE, so it will get closer to the outliers to fit them, making it not able to fit the normal data perfectly.

We can use the following ways to handle outliers.

(1) Remove the outliers from the data set.
By removing the outliers, the model will only have to learn from the normal data, thus being able to fit the data perfectly.

(2) Use np.clip().
We can set an upper bound and lower bound for the data. If the data is out of the range, it will be modified to the upper bound or lower bound.

(3) Scale the data.
Scaling and transforming the data to reduce the impact of outliers, such as normalizing or taking log.

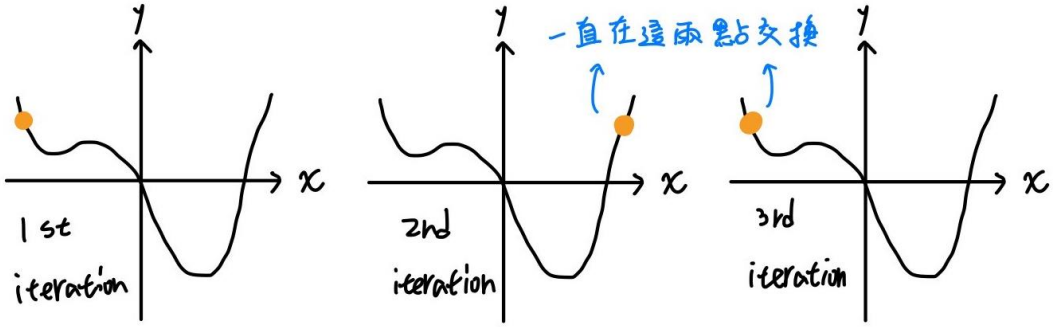
(4) Use other algorithms.
We can use other algorithms that are more robust to the outliers to train the model. For example, robust regression or M-estimators.

2. (15%) How do different values of learning rate (too large, too small…) affect the convergence of optimization? Please explain in detail.

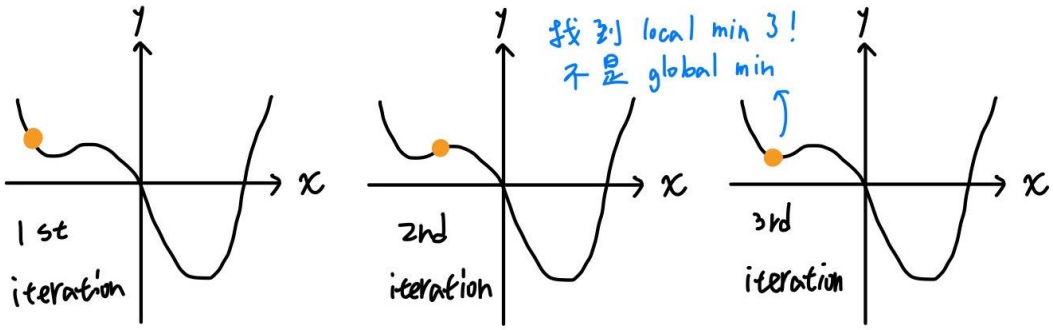
If the learning rate is too big, the model cannot learn very well. It may skip the minimum value and go to the next value. The model will diverge to a super large value in the end and never converge. On the other hand, the model cannot learn very well with a small learning rate, either. It may take too long (too many epochs) to reach the

minimum value. Or it will stop at the local minimum instead of the global minimum. The model's update is too small, so it cannot leave the local minimum, which makes it find the local minimum as the global minimum. Therefore, we need to find a proper learning rate that can give the best result.

Learning rate too large, the orange point will keep moving between the two positions. It will never go to other position, and thus cannot find the minimum point.



Learning rate too small, it may stop at the local minimum because it is too small to leave there, so it cannot find the global minimum either.



3. (15%)

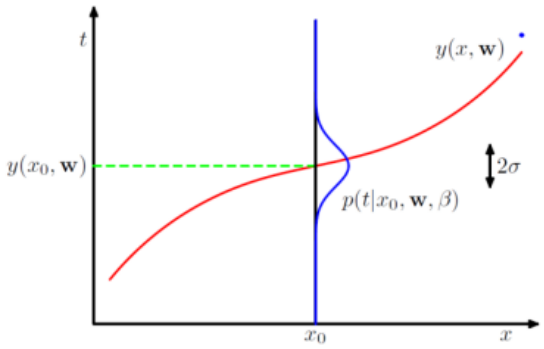
- What is the prior, likelihood, and posterior in Bayesian linear regression. [Explain the concept in detail rather than writing out the mathematical formula.]

  - Prior

    Before we see any data, we can make an assumption to the model, that is the prior function. We may assume it is a zero mean Gaussian function so the weights could be closed to zero, which avoids overfitting.

- Likelihood

  It means how possible an observed data is given. For example, in the following figure, the red curve means the predicted value ($y(x_0, w)$) of given x_0, and the blue curve is the likelihood function. It means the probability of getting this red curve value. The most probable value is at the mean (the value that the green curve is).



(From course slide)

- Posterior

  Posterior is proportional to *prior * likelihood.* After we see some data, we can update it and therefore get the model. It is suitable when we only have little data because it takes both our beliefs (prior) and the observed data (likelihood) into account.

- What is the difference between Maximum Likelihood Estimation (MLE) and Maximum a Posteriori Estimation (MAP)? (Analyze the assumptions and the results.)

  MLE tries to maximize the likelihood function, while MAP maximizes the posterior function. MLE only considers the data, but MAP will combine both observed data and our beliefs. If we have a large dataset, MLE can give a consistent efficient result. However, if our dataset is small, MAP can give a better result because it considers the prior function.