NYCU Introduction to Machine Learning, Homework 1

Deadline: 23:59, Oct. 10 (Tue), 2023

Part. 1, Coding (50%):

For this coding assignment, your task is to implement linear regression only using numpy. It's important to emphasize that no other libraries are allowed for implementing your model. If you use any other library, you will receive zero points for this assignment.

You are required to implement linear regression by closed-form solution and gradient descent.

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

Requirements:

- Implement Linear Regression by closed-form solution.

Criteria:

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

```
Closed-form Solution
Weights: [2.85817945 1.01815987 0.48198413 0.1923993 ], Intercept: -33.78832665744904
```

(40%) Linear Regression Model - Gradient Descent Solution

Requirements:

- Update your weights and intercept by gradient descent (you can implement mini-batch gradient descent or stochastic gradient descent if you want).
- Use MSE (Mean Square Error) as your loss function.

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

(Note: You must use this provided function, do not multiply any constant factor such as 1/2)

- Tune the learning rate and epoch hyper-parameters (and batch size if you implement mini-batch gradient descent) to make your testing MSE loss as closed as the closed-form solution.

(next page)

Criteria:

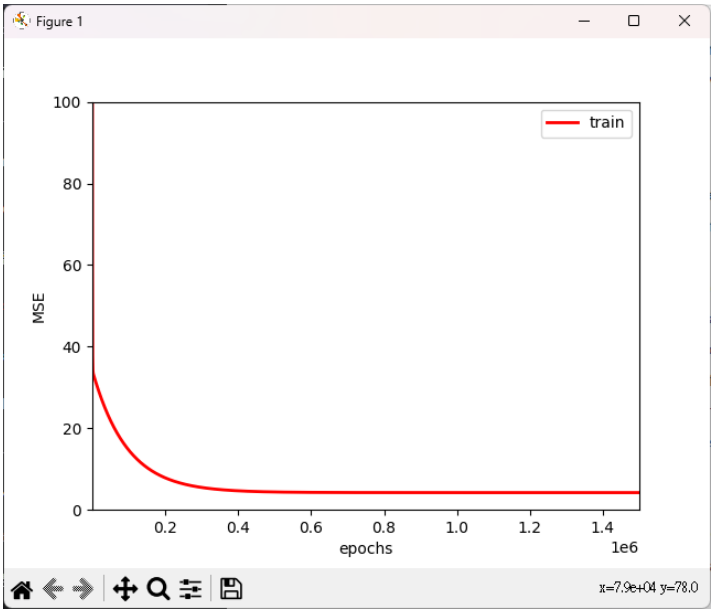
2. (0%) Show the learning rate and epoch (and batch size if you implement mini-batch gradient descent) you choose.

```
LR.gradient_descent_fit(train_x, train_y, lr=0.0001, epochs=1500000)
```

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

```
Gradient Descent Solution
Weights: [2.85794536 1.01808658 0.48126141 0.19222874], Intercept: -33.77625395685295
```

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: -0.0%
```

note: error rate: (gradient_descent_loss - closed_form_loss) / closed_form_loss * 100

Points	**error rate**
20	**< 0.5%**
15	**< 1%**
10	**< 3%**
5	**< 5%**
0	**>= 5%**

(next page)

Part. 2, Questions (50%):

1. (10%) How does the value of learning rate impact the training process in gradient descent? Please explain in detail.
   Learning rate impact the speed of the model converges. If the rate is too large, the model may pass the minimum continuously and can not converge to the minimum point. On the other hand, if the rate is to small, it just like we walk in very small steps and we need to take lots of time to reach the same target.

2. (10%) There are some cases where gradient descent may fail to converge. Please provide at least two scenarios and explain in detail.
   Inappropriate learning rate: details is in question 1.
   Several minima: If the dunction has multiple minima, the algorithm might converge to a local minimum instead of the global minimum.

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.
   Yes.
   Advantages of MSE:
   mathematical simplicity: only needs to calculate the means and the difference, and the squared terms can ensure that negative and possitive errors would not impact each others.
   convex: MSE loss function is convex, which means it has a single minimum.

   Where MSE may be inappropriate and its alternative plan:
   outliers: due to the squaring of errors, the outliers will cause big impact to the MSE. Mean absolute Error (MAE), using absolute value instead of squared, is less sensitive to outliers.
   prediction might be bounded in a range: MSE doesnot consider the bounds or constraints on the prediction. Beta Regression is designed for modeling percentages or proportions, and it constrains the predition within the [0, 1] range.

(next page)

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, $\lambda=10^{-100}$ or $\lambda=0$

The penalty on coefficients is negligible. It has almost no regulariztion effect, and also may lead to overfitting. The model behaves similar to a standard linear regression model without regularization.

Regularization with small lambda can decrease bias. It allows the model to capture more complex patterns in the data.

It also increases variance by allowing the model to fit the training data more closely, capturing both the underlying patterns and the noise in the data.

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

It can lead to over-regularization, where the model is excessively penalized, and drives many coefficients to exactly zero. Only a handful of features with the most significant impact on the target variable will have non-zero coefficients.

Regulariztion with high lambda can increase bias. It simplifies the model by shrinking the coefficients, which may lead to underfitting, occuring when the model is too simplistic to capture the underlying patterns in the data.

It also reduces variance by preventing the model from fitting the noise in the training data, making the model more stable and lesssensitive to small fluctuations in the traing dataset.