The Experiment Report of Machine Learning

SCHOOL: SCHOOL OF SOFTWARE ENGINEERING

SUBJECT: SOFTWARE ENGINEERING

Author:
Xiaoxuan Peng

Supervisor:
Mingkui Tan

Student ID：
201730683321

Grade:
Undergraduate

October 24, 2019

Linear Regression and Stochastic Gradient Descent

Abstract—This experiment intends to use closed-form solution and Stochastic Gradient Descent(SGD) for the linear regression problem.

I. Introduction

Linear regression model is the basic model for machine learning. We can use closed-form solution to find out the parameter **w** and stochastic gradient descent to find the best parameter **w** to minimize the loss function. In this experiment, I implement these two methods on a housing dataset.

II. Methods and Theory

A. Linear Regression Model

Consider function $f(\mathbf{x};\mathbf{w})$ with parameters $\mathbf{w} \in \mathbb{R}^m$ and $b \in \mathbb{R}$, also with input $\mathbf{x}$ where $x_j \in \mathbb{R}$ features for j from 1 to m.

The Model Function is:

$f(\mathbf{x}; b, \mathbf{w}) = \mathbf{w}^T\mathbf{x} + b$

B. Loss Function

In this experiment, I use least squared loss function:

$$\mathcal{L}_D(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$

The goal is use this loss function to find the best parameter **w :**

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \mathcal{L}_D(\mathbf{w})$$

C. Closed-form Solution

$$\frac{\partial \mathcal{L}_D(\mathbf{w})}{\partial \mathbf{w}} = -\mathbf{X}^T\mathbf{y} + \mathbf{X}^T\mathbf{X}\mathbf{w} = 0 \Rightarrow \mathbf{X}^T\mathbf{X}\mathbf{w} = \mathbf{X}^T\mathbf{y} \Rightarrow \mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Solve for optimal parameter **w :**

$$\mathbf{w}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \arg\min_{\mathbf{w}} \mathcal{L}_D(\mathbf{w})$$

D. Stochastic Gradient Descent

We use $\mathbf{d} = -\frac{\partial \mathcal{L}_D \mathbf{w}}{\partial \mathbf{w}}$ as the direction of optimization.

Gradient(vector of partial derivatives):

$$\frac{\partial \mathcal{L}_D(\mathbf{w})}{\partial \mathbf{w}} = \begin{bmatrix} \frac{\partial \mathcal{L}_D(w_1)}{\partial w_1} \\ \frac{\partial \mathcal{L}_D(w_2)}{\partial w_2} \\ \cdot \\ \cdot \\ \cdot \\ \frac{\partial \mathcal{L}_D(w_m)}{\partial w_m} \end{bmatrix}$$

For L2-Loss, the gradient is:

$grad\ \mathcal{L}(\mathbf{w}) = \mathbf{X}^T(\mathbf{X}\mathbf{w} - \mathbf{y})$

III. Experiment

A. Dataset

This dataset is a scaled housing dataset in LIBSVM Data with 506 samples and each sample has 13 features. The features are all scaled from -1 to 1 so we don't need to normalize the data.

B. Implementation

(1) Initialization

The dataset is split into two parts, 80% for training set and 20% for validation set, we don't generate test set here.

I use random normal distribution to initialize parameter **w :**

$$\mathbf{w} \sim N(\mu = 0, \sigma^2 = 1)$$

(2) Parameters

a) There are no parameters in closed-form solution.

b) There are two hyperparameters in stochastic gradient descent. The detailed information about these two hyperparameters is listed below.

Table 1-Hyperparameters in SGD

Parameters	Value
Learning rate	0.0008
Number of epochs	500

(3) Results

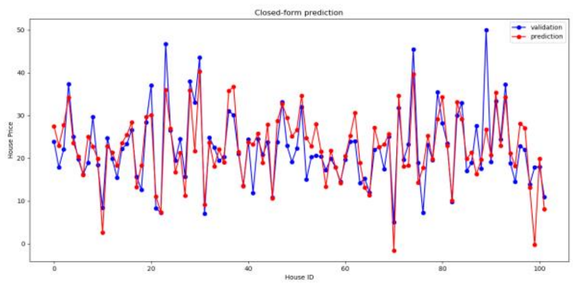
a) Closed-form solution

The loss values are listed below. Figure 1 listed below shows the closed-form prediction on validation set.

Table 2-Loss values result

Items	Loss value
Initial training set	270.285
Trained training set	11.863
Trained validation set	13.604

Figure 1-Closed-form prediction on validation set



b) Stochastic Gradient Descent

The loss values are listed below. Figure 2 listed below shows the loss values on both training set and validation set. Figure 3 listed below shows the L2 norm of the $\mathbf{w}_k$ and $\mathbf{w}^*$ where $\mathbf{w}_k$ means the $\mathbf{w}$ after k-th epoch and $\mathbf{w}^*$ means the optimized $\mathbf{w}$ solved by the closed-form solution**.**

Table 3 Loss values result(SGD)

Items	Loss value
Initial training set	282.8
Trained training set	11.456
Trained validation set	15.432

Figure 2-Loss values in SGD

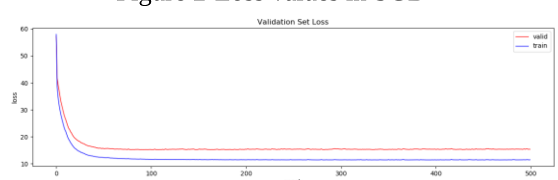
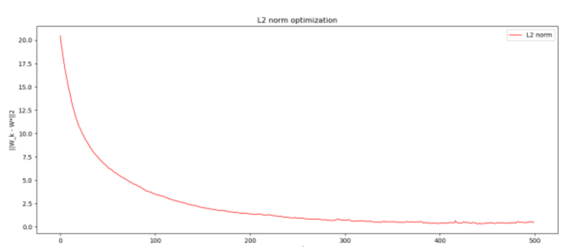


Figure 3-L2 norm of $\mathbf{w}_k$ and $\mathbf{w}^*$



If I use a larger learning rate parameter, for example, learning rate is 0.05,, the figures similar to Figure 2 and Figure 3 are listed below. The results are not converged.

Figure 4-Loss values in SGD(large learning rate)

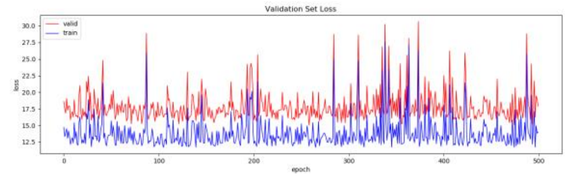
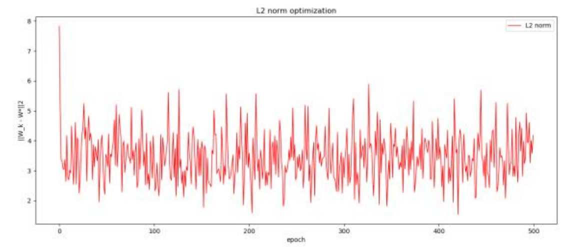


Figure 5-L2 norm of $\mathbf{w}_k$ and $\mathbf{w}^*$(large learning rate)



IV. CONCLUSION

Closed-form solution and Stochastic Gradient Descent are two basic methods of linear regression. As for Closed-form solution, there is a problem if the matrix is not invertible. As for SGD, we can minimize loss function well. Also, we must choose a proper hyperparameter, for example, if we use a large learning rate(LR)in this experiment, the result is not converged.