華南理工大學

South China University of Technology

The Experiment Report of *Machine Learning*

SCHOOL: SCHOOL OF SOFTWARE ENGINEERING

SUBJECT: SOFTWARE ENGINEERING

Author:
Lizhao Liu

Supervisor:
Mingkui Tan

Student ID:
201730683109

Grade:
Undergraduate

November 10, 2019

Linear Regression, Ridge Regression and Gradient Descent

***Abstract*—In this report, we solve the house price prediction problem in Housing Dataset by using closed form solution(CFS) optimized linear regression(LR), CSF optimized ridge regression(RR), and LR optimized by Gradient Descent(GD) and its variants e.g. Stochastic Gradient Descent(SGD) and Mini Batch Gradient Descent(MBGD).**
We perform experiments on four aspects:
1. Comparision between CFS optimized LR and CFS optimized RR.
2. Comparision between GD, SGD, MBGD in terms of convergence and time-cost.
3. Tuning learning rate in MBGD.
4. Comparision between CFS optimized and MBGD optimized LR.

I. Introduction

LINEAR Regression is the core of machine learning. Based on it, many machine learning algorithm, such as RR, Logistics Regression, SVM are developed. Even in many deep learning model, such as Covolutional Neural Network(CNN), Long Short Term Memory(LSTM), Gated Recurrent Unit(GRU), LR(or Linear Layer) is the basic component. So it is very important to fully exploit the insight of LR. To achieve that goal, In this experiment, we conduct many experiments on CSF optimized LR, CSF optimized RR, MBGD optimized LR by solving the house price predicting problem.

II. Methods and Theory

In this part, we first define the Linear Regression, Mean Squared Error(MES) Loss function. Then we solve the closed form solution of Linear Regression and its variant Ridge Regression. At last we define gradient descent algorithm, and its variants schocastic gradient descent and mini batch gradient descent.

A. Linear Regression

Given dataset $D = \{x_i, y_i\}_{i=1}^{m}$, where $x_i \in \mathbf{R}^n$ and $y_i \in \mathbf{R}$. Liner Regression Model is parameterized by (W, b), where $W \in \mathbf{R}^n$ and $b \in \mathbf{R}$ and the form of it is:

$$y_i = x_i^T W + b \tag{1}$$

We can write it in vertorized form:

$$y = X^T W \tag{2}$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}, \quad \boldsymbol{W} = \begin{pmatrix} b \\ W_1 \\ W_2 \\ \vdots \\ W_n \end{pmatrix},$$

$$X = \begin{pmatrix} 1 & \mathbf{x}_1^{\mathsf{T}} \\ 1 & \mathbf{x}_2^{\mathsf{T}} \\ \vdots & \vdots \\ 1 & \mathbf{x}_m^{\mathsf{T}} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1n} \\ 1 & x_{21} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{m1} & \cdots & x_{mn} \end{pmatrix}.$$

B. Mean Squared Error

Given the prediction of Linear Model $\hat{y} = X^T W$ and the Ground truth y, the MSE loss function is:

$$L(\hat{y}, y) = \frac{1}{2m} \Sigma_{i=1}^{m} (\hat{y}_i, y_i) = \frac{1}{2m} (\hat{y} - y)^T (\hat{y} - y) \tag{3}$$

C. Closed form Solution

The best parameter W^* can be obtained by

$$W^* = \arg\min_{W} L(X^T W, y) \tag{4}$$

So, we set

$$\frac{\partial L(\hat{y}, y)}{\partial W} = 0 \tag{5}$$

We can get

$$W^* = (X^T X)^{-1} X^T y \tag{6}$$

D. Ridge Regression

In Equation 6, when $m > n$, $X^T X$ is not invertible. Ridge Regression tackle this problem by adding a small term. So the closed form solution for Ridge Regression is

$$W^*_{ridge} = (X^T X + \lambda \mathbf{I})^{-1} X^T y \tag{7}$$

Where $\mathbf{I}$ is identity matrix.
In this way, the loss function for Ridge Regression is

$$L_{ridge}(\hat{y}, y) = \frac{1}{2m} (\hat{y} - y)^T (\hat{y} - y) + \frac{1}{2} W^T W \tag{8}$$

In another point of view, Ridge Regression can be seen as a regularized form of Linear Regression.

E. Gradient Descent

Not all problem can be derived a closed form solution, which is often the case in machine learning. So instead of getting the best parameters W^* immedeatly, we can update it incremently until W gets closed enough to W^*. That is the basic idea of Gradient Descent Algorithm. Gradient Descent Algorithm can be found in Algorithm 1.

Algorithm 1: Gradient Descent

Input: Training Dataset $D = \{x_i, y_i\}_{i=1}^{m}$, learning rate η, max iteration N, parameter W
Output: Optimized parameter $\hat{W}$

$t \leftarrow 1$
for $t < N$ **do**
  $\hat{y} \leftarrow X^T W$
  $L(\hat{y}, y) \leftarrow \frac{1}{2m}(\hat{y} - y)^T(\hat{y} - y)$
  $\Delta W \leftarrow \frac{\partial L}{\partial W}$
  $W \leftarrow W - \eta * \Delta W$
$\hat{W} \leftarrow W$
return $\hat{W}$

F. Stochastic Gradient Descent

Different from SD which updates after it computes the gradient of parameter over all examples, stochastic gradient descent updates immediately once it computes parameter's gradient from only one example. SGD Algorithm can be found in Algorithm 2. It can be seen in the algorithm that, there is another for loop inside the for loop of gradient descent algorithm.

Algorithm 2: Stochastic Gradient Descent

Input: Training Dataset $D = \{x_i, y_i\}_{i=1}^{m}$, learning rate η, max iteration N, parameter W
Output: Optimized parameter $\hat{W}$

$t \leftarrow 1$
for $t < N$ **do**
  $i \leftarrow 0$
  for $i < m$ **do**
    $\hat{y}_i \leftarrow X_i^T W$
    $L(\hat{y}_i, y_i) \leftarrow \frac{1}{2}(\hat{y}_i - y_i)^T(\hat{y}_i - y_i)$
    $\Delta W \leftarrow \frac{\partial L}{\partial W}$
    $W \leftarrow W - \eta * \Delta W$
$\hat{W} \leftarrow W$
return $\hat{W}$

G. Mini Batch Gradient Descent

Due to the fact that SGD updates the parameter based on only one examples, which it is time-comsuming, mini batch gradient descent updates the parameter based on a batch of examples, typically the batch size is bigger than 1 and smaller than the number of whole examples, e.g. 8, 16, 32... MBGD Algorithm can be found in Algorithm 3.

III. Experiments

A. Dataset

We conduct all the experiments on housing dataset, which has total 506 examples and each example's form is (x, y), where $x \in \mathbf{R}^{13}$ and $y \in \mathbf{R}$. We split the dataset into three part: 272 training examples, 67 validation examples and 167 test examples. X is already scaled between -1 and 1. And we normalized the $X_{train}, X_{validation}, X_{test}$ by using the mean and variance of X_{train}.

Algorithm 3: Mini Batch Gradient Descent

Input: Training Dataset $D = \{x_i, y_i\}_{i=1}^{m}$, learning rate η, max iteration N, parameter W
Output: Optimized parameter $\hat{W}$

$t \leftarrow 1$
for $t < N$ **do**
  $i \leftarrow 0$
  for X_{batch}, y_{batch} *in* D **do**
    $\hat{y}_{batch} \leftarrow X_{batch}^T W$
    $L(\hat{y}_{batch}, y_{batch}) \leftarrow \frac{1}{2}(\hat{y}_{batch} - y_{batch})^T(\hat{y}_{batch} - y_{batch})$
    $\Delta W \leftarrow \frac{\partial L}{\partial W}$
    $W \leftarrow W - \eta * \Delta W$
$\hat{W} \leftarrow W$
return $\hat{W}$

B. Implementation

We implement the Linear Regression, Ridge Regression and Gradien Descent Algorithm using python and mainly rely on the numpy package.

C. Linear Regression and Ridge Regression

In this section, we conduct the experiment of serveral magnitude of λ, which leads to serveral W_{ridge}. We use L_2norm to represent the magnitude of a vector. Given $v \in \mathbf{R}^n$

$$L_2norm(v) = \sqrt{\Sigma_{i=1}^{n} v_i} \quad (9)$$

From Table I, we can see that:

First, as the magnitude of lambda becomes bigger, both MSE_{train} and MSE_{val} become bigger.

Second, as the magnitude of lambda becomes bigger, the magnitude of W_{ridge} become smaller.

The first observation is not often the cases. But the second observation can be explained that as the regularizer λ become bigger, the W is regularized, so that it is smaller.

By comparing Table I and Table II, we can see that, the performance and the weight magnitude is almost the same as the λ is small, but the difference become larger as λ grows.

We also explore the weight mantitude between linear regression and ridge regression under different λ magnitude setting. Specifically, we set log_{10}^{λ} in range from -7 to 3. We can see from Fig 1, as the magnitude of λ growing, the weight magnitude of ridge regression become much smaller than linear regression.

D. Gradient Descent, Stochastic Gradient Descent and Mini Batch Gradient Descent

In this section, we conduct experiments on optimizing the linear regression model by using three gradient descent

TABLE I
DIFFERENT LAMBDAS' IMPACT IN TERMS OF ERROR AND W MAGNITUDE.

log_{10}^{λ}	-2	-1	0	1	2	3
MSE_{train}	10.23	10.23	10.24	10.63	30.37	181.35
MSE_{val}	17.08	17.09	17.16	18.20	43.92	214.55
$L_2norm(W_{ridge})$	0.83	0.82	0.81	0.72	0.50	0.38

TABLE II
LINEAR REGRESSION ERROR AND W MAGNITUDE.

MSE_{train}	10.23
MSE_{val}	17.08
$L_2norm(W)$	0.83

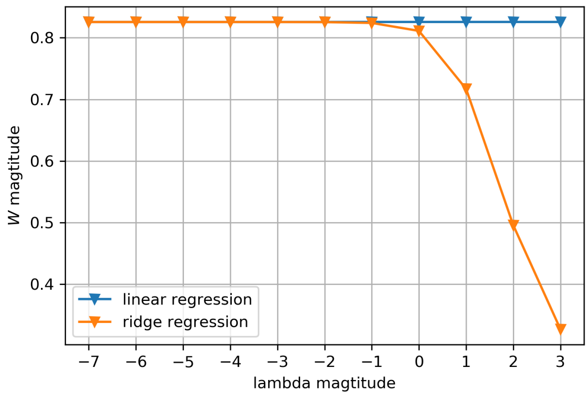

Fig. 1. The magnitude of W_{ridge} under different lambda mgnitude.

algorrithms, and compare them in terms of convergence and time cost.

We first evaluate the convergence of three algorithm, specificly we set batch size in a range from 1 to number of full examples in training set. When the batch size is one, mini batch gradient descent becomes stochastic gradient descent and when the batch size equals to number of full examples, mini batch gradient descent becomes gradient descent. All experiments are with the same learning rate 0.01.

As we can see from Fig 2 and Fig 3, as the batch size become larger, the convergence of the linear regression model bocome slower. The reason is that, in each epoch, the smaller the batch size, the model have more update steps, so the convergence is faster. Especially, the stochastic gradient descent converges within 10 steps, while gradient descent requires nearly 400 steps, there is a 400x margin.

We then mesure the time cost for each batch size.

As illustrated in Fig 4, the smaller the batch size, more time cost is required, because of the explicit for loop in the Stochastic gradient descent and mini batch gradient descent. Especially, the gradient descent requires nearly 1 seconds in one step, while stochastic gradient descent require 60 seconds, there exist a 60x margin.

E. Impact of learning rate

In this section, we evaluate the impact of learning rate in mini batch gradient descent, we set batch size to 64 and epochs to 200. We set learning rate in a range e.g. 1e-4, 1e-3, 1e-2, 1e-1, 1.

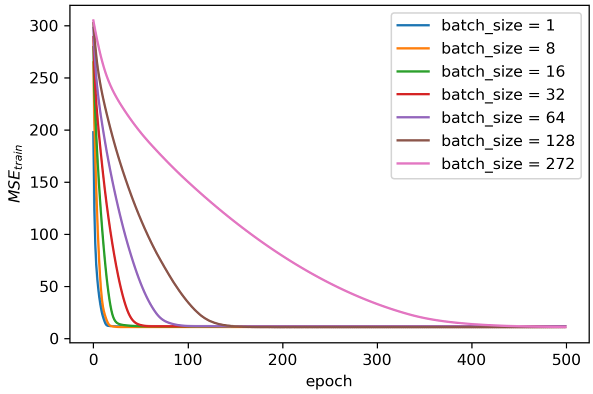

Fig. 2. The MSE of training set under different batch size.

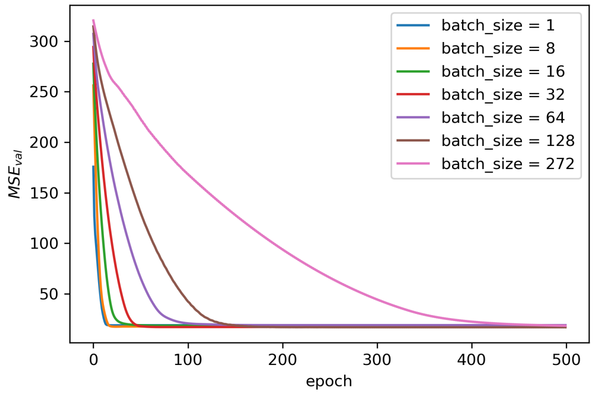

Fig. 3. The MSE of validation set under different batch size.

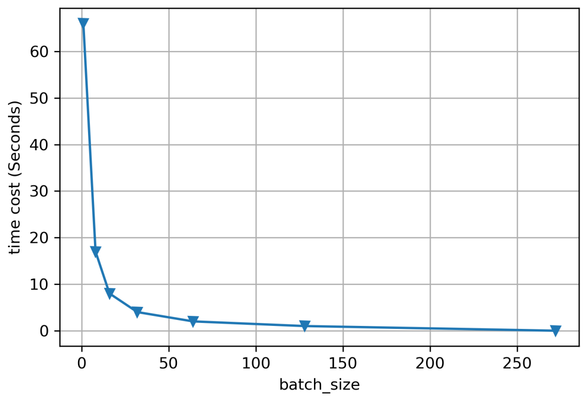

Fig. 4. The time cost of 100 epoches under different batch size.

As presented in Fig 5 and Fig 6, small learning rate tend to make the convergence smoothly, but need to get more epoches to fully converge. However, big learning rate make the convergence very fast, but will not guarateee that the convergence will be near the optimal(or local optimal). And big learning rate make the MSE shaking around after a specific epoch.

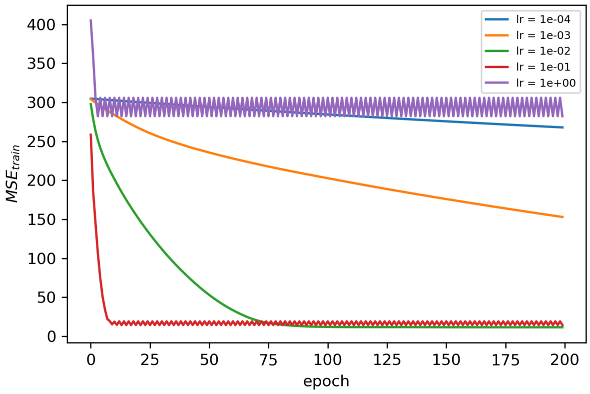

Fig. 5. The MSE in training set under different learning rate.

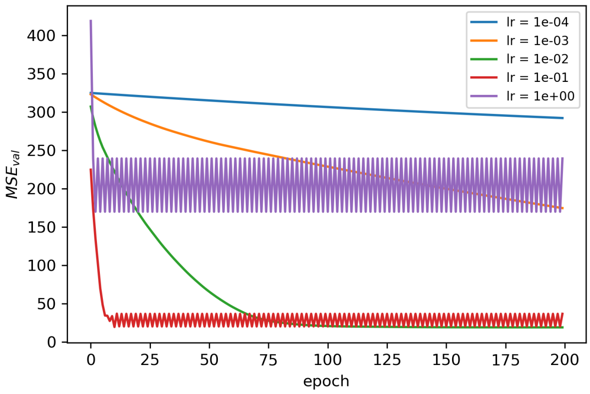

Fig. 6. The MSE in validation set under different learning rate.

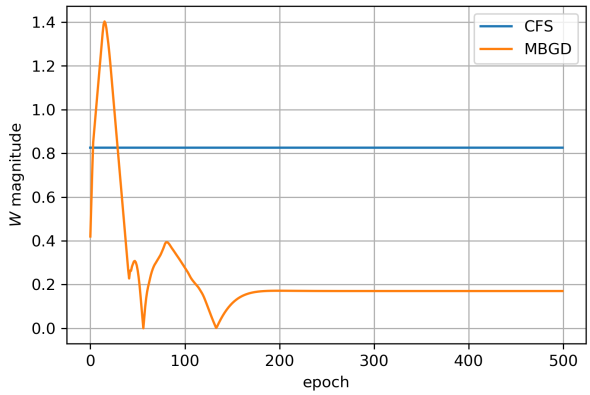

Fig. 7. The magnitude of W_{mbgd} under different epoches..

F. Weight magnitude between CSF and MBGD

In this section, we explore the weight magnitude($L_2 norm$) between CSF and MBGD optimized weight of linear regression.

Then, we experiment on comparing the weight magnitude of linear regression by closed form solution(CFS) and mini batch gradient descent(MBGD). In MBGD, we set learning rate to 0.01 and batch size to 64. We can see that MBGD optimized weight's magnitude is stable after some epoches. But MBGD optimized weight's magnitude has a large margin(e.g. about 0.6) to CFS optimized weight. The reason is still unclear.

IV. Conclusion

In this report, we explore the