华南理工大学

South China University of Technology

The Experiment Report of Machine Learning

SCHOOL: SCHOOL OF SOFTWARE ENGINEERING

SUBJECT: SOFTWARE ENGINEERING

Author:
Lilong Huang

Supervisor:
Han Huang

Student ID：
201720144962

Grade:
Graduate

December 9, 2017

Logistic regression, linear classification and stochastic gradient descent

Abstract—Stochastic Gradient Decreasing Because of its high speed, slow convergence, it is widely used for complex optimization problems. In this report, we mainly use different optimization methods to update the model parameters, and analyze the advantages and disadvantages of various update formulas through comparative experiments, which can guide and reference future optimization problems.

I. INTRODUCTION

The main purpose of this experiment is to contrast and understand the relationship between gradient descent and stochastic gradient descent. Comparing and understanding the difference and connection between logistic regression and linear classification, we can further understand the principle of SVM and implement it on larger data.

Linear regression uses the Housing data in LIBSVM Data, which contains 506 samples, each of which has 13 attributes and is split into training sets, validation sets.Linear classification uses australian data from LIBSVM Data, which contains 690 samples, each with 14 attributes. Please download the scaled version and divide it into training set and verification set. The experimental code and drawing are completed on jupyter.

In the second part,we will first exhibit the loss function of logistic regression and linear classification as well as the derivative function. Moreover, we will also illustrate the equations of various optimized methods and compare the difference between each other. In the third Experiments part, for intuitively observing, we will provide some experimental results about the mentioned methods. In the finally part, we will draw some conclusions about the whole report.

II. METHODS AND THEORY

In this section, we mainly introduce the experimental procedures and methods used in the experiment, including loss function, using gradient information, different optimization methods to update model parameters.The loss function of logistic regression is:

$$L_{reg} = -\frac{1}{N}\sum_{i=1}^{N} y^{(i)} log(h(x^{(i)})) + (1-y^{(i)}) log(1-h(x^{(i)})), \quad (1)$$

where $h(x^{(i)}) = \frac{1}{1+e^{-w^T x}}$

The corresponding gradient formula is:

$$\frac{\partial L_{reg}}{\partial w} = \frac{1}{N}\sum_{i=1}^{N}(h(x^{(i)}) - y^{(i)}) * x_j^{(i)}$$

The loss function of Support Vector Machine (SVM) is:

$$L_{cls} = \frac{\|w\|^2}{2} + C\frac{1}{N}\sum_{i-1}^{N} max(0, 1 - y_i(w^T x_i + b)) \quad (3)$$

The corresponding gradient formula is:

$$s_i = \begin{cases} 0, & 1 - y_i(w^T x_i + b) \leq 0 \\ 1, & 1 - y_i(w^T x_i + b) > 0 \end{cases}$$

$$\frac{\partial L_{cls}}{\partial w} = w + CX^T(y \odot s)$$

The Nesterov's Acceleration Gradient Algorithm (NAG) is an improved version of the Gradient Algorithm (GD). Nesterov first proposed in 1983. It has been shown that the NAG algorithm is the best method for all gradient based (or first order) algorithms. However, the original NAG algorithm can only deal with the smooth convex optimization problem. The latest development is extending NAG to a wider range of convex optimization problems.NAG optimization update formula is:

$$\begin{aligned} v_t &= \gamma v_{t-1} + \eta \nabla_\theta L(\theta - \gamma v_{t-1}) \\ \theta &= \theta - v_t \end{aligned}$$

Adadelta is an extension of Adagrad that seeks to reduce its aggressive, monotonically decreasing learning rate. Instead of accumulating all past squared gradients, Adadelta restricts the window of accumulated past gradients to some fixed size w. Adadelta optimization update formula is:

$$\begin{aligned} g_t &= \nabla L(\theta_{t-1}) \\ G_t &= \gamma G_t + (1-\gamma) g_t \odot g_t \\ \Delta\theta_t &= -\frac{\sqrt{\Delta} + \epsilon}{\sqrt{G_t + \epsilon}} \odot g_t \\ \theta_t &= \theta_{t-1} + \Delta\theta_t \\ \Delta_t &= \gamma \Delta_{t-1} + (1-\gamma) \Delta\theta_t \odot \Delta\theta_t \end{aligned}$$

RMSprop is an unpublished, adaptive learning rate method proposed by Geoff Hinton in Lecture 6e of his Coursera Class. RMSprop and Adadelta have both been developed indepen- dently around the same time stemming from the need to resolve Adagrads radically diminishing learning rates. RMSprop optimization update formula is:

$$\begin{aligned} g_t &= \nabla L(\theta_{t-1}) \\ G_t &= \gamma G_t + (1-\gamma) g_t \odot g_t \\ \theta_t &= \theta_{t-1} - \frac{\eta}{\sqrt{G_t + \epsilon}} \odot g_t \end{aligned}$$

Adaptive Moment Estimation (Adam). Adaptive Moment Estimation (Adam) is another method that computes adaptive learning rates for each parameter. Adam optimization update formula is:

$$\begin{aligned} g_t &= \nabla L(\theta_{t-1}) \\ m_t &= \beta_1 m_{t-1} + (1-\beta_1) g_t \\ G_t &= \gamma G_t + (1-\gamma) g_t \odot g_t \\ \alpha &= \eta \frac{\sqrt{1-\gamma^t}}{1-\beta^t} \\ \theta_t &= \theta_{t-1} - \alpha \frac{m_t}{\sqrt{G_t + \epsilon}} \end{aligned}$$

Because the batch gradient descent method requires all training samples when updating each parameter, the training process becomes abnormally slow as the number of samples increases. Stochastic Gradient Descent (SGD) is proposed to solve the problem of batch gradient descent method. Stochastic gradient descent method randomly selects one or more (but not all) training samples to participate in the calculation of loss function, The loss function for each sample is deflected by theta to obtain the corresponding gradient to update theta:

$$\theta_j^{'} = \theta_j + (y^i - h_\theta(x^i)) x_j^i$$

III. EXPERIMENT

In this part, we mainly introduce the specific process of experiment and experimental results.

A. Dataset Experiment uses a9a of LIBSVM Data, including 32561/16281(testing) samples and each sample has 123/123 (testing) features.

B. Implementation The steps of our experiments are as the following:

1) Load the training set and validation set of a9a.

2) Initialize logistic regression or SVM model parameters with normal distribution.

3) Define the loss function and calculate its derivation.

4) Compute the gradient with respect to the weight using different optimized method (SGD, NAG, RMSProp, AdaDelta and Adam).

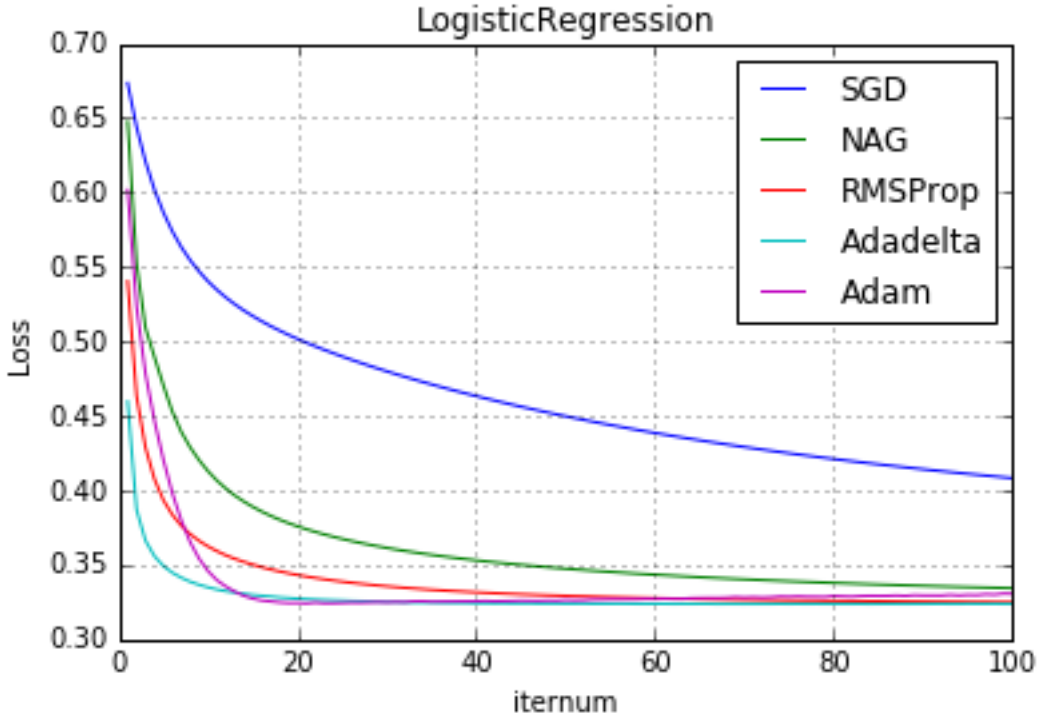
5) Using gradient descent to update the weight.

6) Repeat step (4) and (5) for several times until conver- gence.

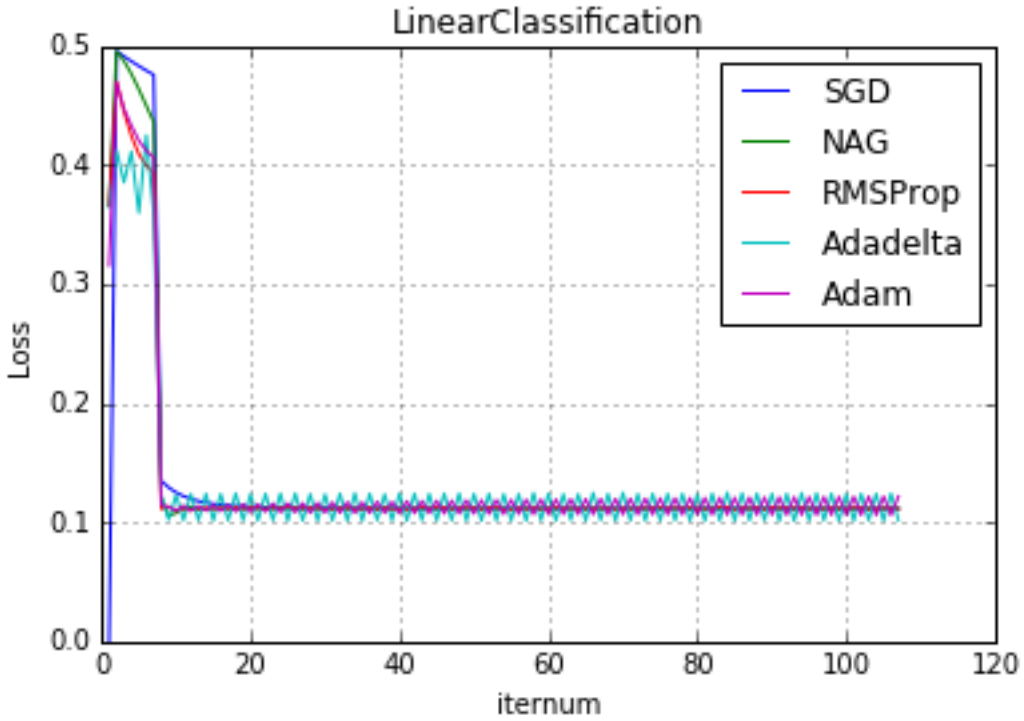
We will use a variety of optimization update parameter formula were used in both logistic regression and

linear classification task.

In logistic regression, This is the final experimental results:



And the linear classification's results:



As can be seen from the figure ,with the increase of time, the loss of the four algorithms are gradually reduced, and it can be clearly seen that the four algorithms achieve lower Loss than SGD.

IV. CONCLUSION

In this reprot, we introduct some variants of SGD algorithm, which aims to compare the difference between the four algorithms. For esaier to understand, we try to conduct some experiments and visualize the results of these optimized methods in two tasks logistic regression and linear classification.The results show that all the variants of SGD better than the original in some ways, convergence is also significantly faster.