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:MengDan Zheng

Supervisor:Mingkui Tan

Student ID：201530613795

Grade:Undergraduate

December 15, 2017

Linear Regression, Linear Classification and Gradient Descent

Abstract—Motivation of Experiment is :

①Compare and understand the difference between gradient descent and stochastic gradient descent.

②Compare and understand the differences and relationships between Logistic regression and linear classification.

③Further understand the principles of SVM and practice on larger data.

I. INTRODUCTION

The experiment is divided into two parts, the first part is to deal with classification problem using logistic regression . The second part is to deal with the classification problem using svm linear classification. Both of these parts use seven gradient optimization algorithms, including SGD, Momentum, NAG, Adagrad, RmsProp, Adadelta and Adam. We will make the comparison of the effects of logistic regression and svm on classification, and the comparison of these optimization algorithms.

II. METHODS AND THEORY

First, about logistic regression, the loss function is:
j_i(w)=∑_i_n
(y_i*log(1/e**(-w.T*x_i))+(1-y_i)*log(1-1/e**(-w.T*xi))
The gradient of loss function about w is：
grad_w=(1/e**(-w.T*x)-y)*x.
There is one thing to note is the original y is in the set of {-1,1}.The formulation above is with the precondition that y is in the set of{0,1}.So we should convert the y from -1 to 0 before starting the computation of loss and gradient.

About linear classification, the loss function is:
j(w,b)=1/2 w.T*w + c*∑_i_n (max(0,1-yi*(w.T*xi+b)))
The gradient of loss function about w , b is：
If 1-yi*(w.T*xi+b)>=0 :
g_wi=-y_i*x_i;
g_bi=-y_i;
if 1-yi*(w.T*xi+b)<0 :
g_wi=0;
g_bi=0;
grad_w=w+ c / n * ∑_i_n (g_wi);
grad_b=c / n * ∑_i_n (g_bi)
We can let b be a element of w, so we should insert a column with all one to x data. Then the loss function and gradient of linear classification is:
Loss function :
j(w)=1/2 w.T*w + c*∑_i_n (max(0,1-yi*(w.T*xi)))
Gradient of w:
If 1-yi*(w.T*xi)>=0 :
g_wi=-y_i*x_i;
if 1-yi*(w.T*xi)<0 :
g_wi=0;
grad_w=w+ c / n * ∑_i_n (g_wi);

Second, about those gradient optimization algorithms. I learned the fomulation from the blog:
SGD:

$$J(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i}^{n} J_i(\boldsymbol{\theta})$$

$$\mathbf{g}_t \leftarrow \nabla J_i(\boldsymbol{\theta}_{t-1})$$
$$\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \eta \mathbf{g}_t$$

Momentum:

$$\mathbf{g}_t \leftarrow \nabla J(\boldsymbol{\theta}_{t-1})$$
$$\mathbf{v}_t \leftarrow \gamma \mathbf{v}_{t-1} + \eta \mathbf{g}_t$$
$$\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \mathbf{v}_t$$

NAG:

$$\mathbf{g}_t \leftarrow \nabla J(\boldsymbol{\theta}_{t-1} - \gamma \mathbf{v}_{t-1})$$
$$\mathbf{v}_t \leftarrow \gamma \mathbf{v}_{t-1} + \eta \mathbf{g}_t$$
$$\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \mathbf{v}_t$$

AdaGrad:

$$\mathbf{g}_t \leftarrow \nabla J(\boldsymbol{\theta}_{t-1})$$
$$G_t \leftarrow G_t + \mathbf{g}_t \odot \mathbf{g}_t$$
$$\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \frac{\eta}{\sqrt{G_t + \epsilon}} \odot \mathbf{g}_t$$

RMSProp:

$$\mathbf{g}_t \leftarrow \nabla J(\boldsymbol{\theta}_{t-1})$$
$$G_t \leftarrow \gamma G_t + (1-\gamma)\mathbf{g}_t \odot \mathbf{g}_t$$
$$\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \frac{\eta}{\sqrt{G_t + \epsilon}} \odot \mathbf{g}_t$$

AdaDelta:

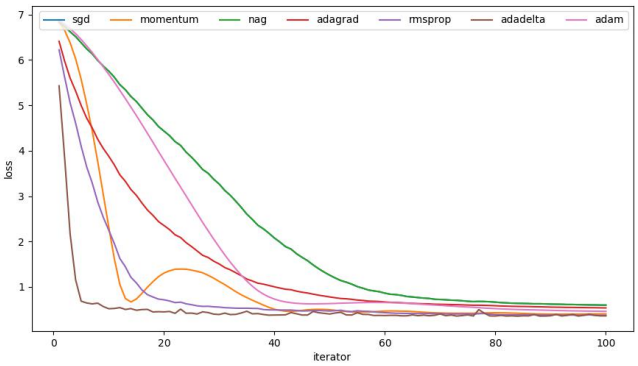
$$\mathbf{g}_t \leftarrow \nabla J(\boldsymbol{\theta}_{t-1})$$
$$G_t \leftarrow \gamma G_t + (1-\gamma)\mathbf{g}_t \odot \mathbf{g}_t$$
$$\Delta\boldsymbol{\theta}_t \leftarrow -\frac{\sqrt{\Delta_{t-1} + \epsilon}}{\sqrt{G_t + \epsilon}} \odot \mathbf{g}_t$$
$$\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} + \Delta\boldsymbol{\theta}_t$$
$$\Delta_t \leftarrow \gamma\Delta_{t-1} + (1-\gamma)\Delta\boldsymbol{\theta}_t \odot \Delta\boldsymbol{\theta}_t$$

Adam:

$$\mathbf{g}_t \leftarrow \nabla J(\boldsymbol{\theta}_{t-1})$$
$$\mathbf{m}_t \leftarrow \beta_1 \mathbf{m}_{t-1} + (1-\beta_1)\mathbf{g}_t$$
$$G_t \leftarrow \gamma G_t + (1-\gamma)\mathbf{g}_t \odot \mathbf{g}_t$$
$$\alpha \leftarrow \eta \frac{\sqrt{1-\gamma^t}}{1-\beta^t}$$
$$\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \alpha \frac{\mathbf{m}_t}{\sqrt{G_t + \epsilon}}$$

III. EXPERIMENT

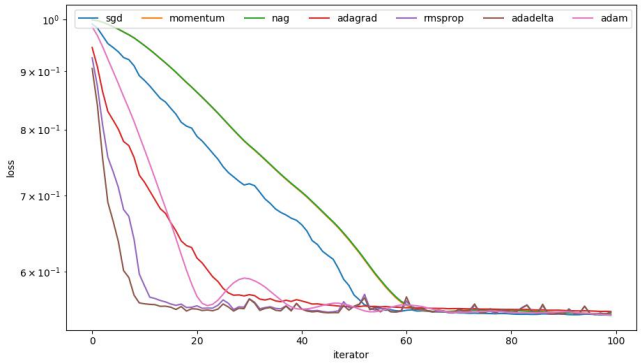
First part is about logistic regression. I used seven gradient optimization algorithms to update the w. The chart of the change of loss is:



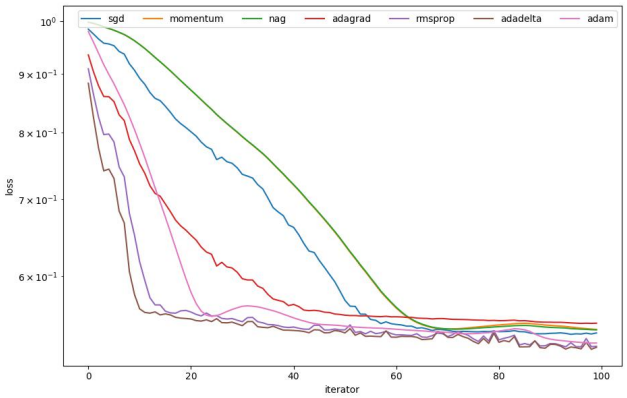
In this section, I set the original learning rate of SGD, Momentum, NAG, RmsProp and Adam are all 0.03. The original rate of adagrad is 0.06 because I found it is too slow to convergence .So in my experiment, adagrad is the slowest algorithm. And adadelta doesn't need original rate.

From the chart, we can get that nag and sgd are very closed.They are the second slowest algorithm to convergence followed by adam, rmsprop, momentum. Adadelta is the fastest algorithm.

Second part is svm linear classifier. In this part I achieved it in two way. First, with the variable w and b, The chart of the change of loss is:



Second, with the only variable w, The chart of the change of loss is:



In this two way, the curves are very similar. I set the original learning rate of Momentum, NAG, RmsProp and Adam are all 0.0005. The original rate of adagrad is 0.01 and rate of sgd is 0.005. The reason is same as the one mentioned above. So the result is so closed with the first part. Adagrad is slowesr,followed by sgd. NAG and Momentum are very closed. Adadelta, Adam and Rmsprop are all very fast.

As the comparison of logistic regression and svm on classification, I compare them by the accuracy of classification. The classification accuracy of linear regression is about ,and the worst is 0.8304772434125668 . The classification accuracy of svm is about 0.7637737239727289. This indicates logistic regression has a better performance than linear svm on this dataset.

IV. CONCLUSION

From this experiment, It can be concluded that logistic regression has a higher classification accuracy than linear svm. What's more, The order of convergence rate of seven different optimization algorithms from slow to fast are: AdaGrad, SGD, NAG, Momentum, Adam, RMSProp, Adadelta. But the faster the algorithm, the effect is not necessarily better. It should be adjusted according to different circumstances.