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:
Enjun Zhong

Supervisor:
Mingkui Tan

Student ID：201530613818

Grade:
Undergraduate or Graduate

December 12, 2017

Logistic Regression, Linear Classification and Stochastic Gradient Descent

Abstract—Compare and understand the difference between gradient descent and stochastic gradient descent. Compare and understand the differences and relationships between Logistic regression and linear classification.

I. INTRODUCTION

In the exoeriment, we use the Logistic Regression and Linear Classification model to solve the classification problem. And use GD and SGD to train the model in order to compare the differences between different model and algorithm.

II. METHODS AND THEORY

The loss function of Logistic Regression:

$$-\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} \log h_\theta\left(x^{(i)}\right) + (1-y^{(i)}) \log(1-h_\theta\left(x^{(i)}\right))]$$

Derivatives: $-\frac{1}{m}\sum_{i=1}^{m} \alpha(h_\theta\left(x^{(i)}\right) - y^{(i)})x^{(i)}$

The loss function of Linear Regression:

$$\frac{\|w\|^2}{2} + C\sum_{i=1}^{n} max(0,1-y_i(x_i w + b))$$

Derivatives:

$$g_t = \begin{cases} w + C\sum_{i=1}^{n} -x_i^T y_i & 1-y_i(x_i w + b) \geq 0 \\ w & 1-y_i(x_i w + b) < 0 \end{cases}$$

SGD algorithm: NAG

$$g_t = \nabla J(\theta_{t-1})$$
$$v_t = \gamma v_{t-1} + \eta g_t$$
$$\theta_t = \theta_{t-1} - v_t$$

RMSProp:

$$g_t = \nabla \mathrm{J}(\theta_{t-1})$$
$$G_t = \gamma G_t + (1-\gamma) g_t \odot g_t$$
$$\theta_t = \theta_{t-1} - \frac{\eta}{\sqrt{G_t + \epsilon}} \odot g_t$$

AdaDelta

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

Adam

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

III. EXPERIMENT

A. Dataset

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

B. Implementation

Logistic Regression and Stochastic Gradient Descent

1. Load the training set and validation set.
2. Initalize logistic regression model parameters.
3. Select the loss function and calculate its derivation.
4. Calculate gradient G toward loss function from partial samples.
5. Update model parameters using different optimized methods(NAG，RMSProp，AdaDelta and Adam).
6. Select the appropriate threshold, predict under validation set and get the different optimized method loss L_{NAG}， $L_{RMSProp}$， $L_{AdaDelta}$and L_{Adam}.
7. Repeate step 4 to 6 for several times, and drawing graph of L_{NAG}， $L_{RMSProp}$， $L_{AdaDelta}$ and L_{Adam} with the number of iterations.

Linear Classification and Stochastic Gradient Descent

1. Load the training set and validation set.
2. Initalize SVM model parameters.
3. Select the loss function and calculate its derivation.
4. Calculate gradient G toward loss function from partial samples.
5. Update model parameters using different optimized methods(NAG，RMSProp，AdaDelta and Adam).
6. Select the appropriate threshold,predict under validation set and get the different optimized method loss L_{NAG}， $L_{RMSProp}$， $L_{AdaDelta}$ and L_{Adam}.
7. Repeate step 4 to 6 for several times, and drawing graph of L_{NAG}， $L_{RMSProp}$， $L_{AdaDelta}$ and L_{Adam} with the number of iterations.

The parameters of algorithm:

NAG: γ=0.9 η=0.01 $v_t = 0$ w=0

RMSProp: γ=0.9 η=0.01 $G_0 = 0$

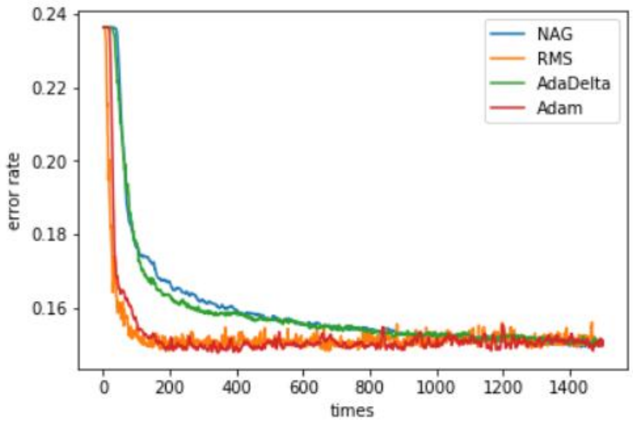
AdaDelta: γ=0.99 Δ_t= 0 $G_0 = 0$

Adam: $\beta_t = 0.9$ $\gamma = 0.95$ $\eta = 0.01$ $m_t = 0$ G_0=0

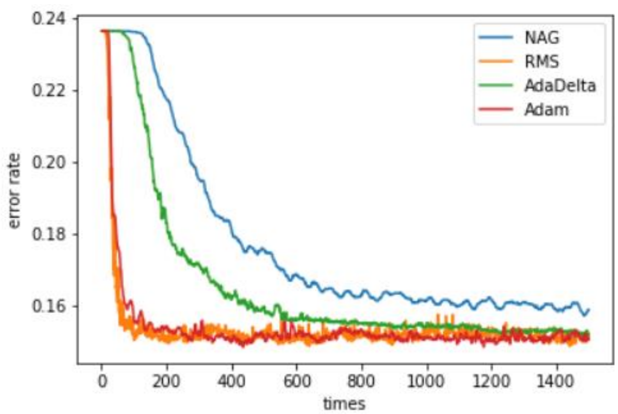
	NAG	RMS	AdaDelta	Adam
Linear Regression loss	0.1503	0.1510	0.1504	0.1510
Linear Classification loss	0.1588	0.1518	0.1518	0.1506

IV. CONCLUSION

Logistic Regression Loss:



Linear Classification Loss:



NAG with η=0.01 decline slower than other algorithms

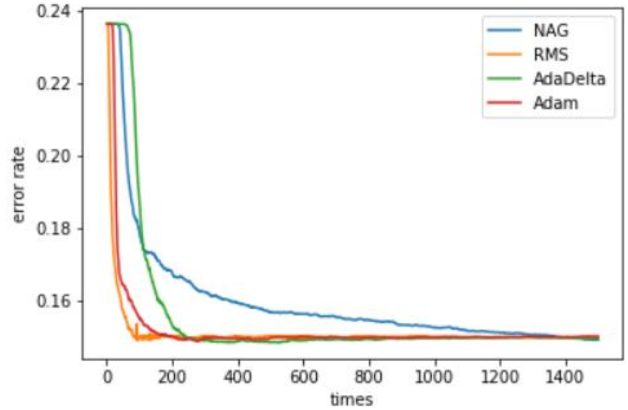
RMS with η=0.01 will decline quicker than NAG but the curve will get most fluctuations.

AdaDelta can adjust its learning rate itself. So the parameter γ will not influence the error rate when reaching the convergence. I try γ=1 an 0 and it’s error rate finally reach the similar value (maybe I try too few values). Maybe this prove that it can adjust the learning rate.

Adam also decline quickly. But different parameter may reach convergence in different value.

According to the Loss curve, the Adam algorithm is the best alogrithm because it has few fluctuations and reach convergence quick.

Compare to GD algorithm, the learning curve of SGD get more fluctuations because it may choose noise to compute gradient while the learning curve of GD decline smoothly.



The picture is the learning curve of Logisitc Regression with GD. We can see the line decline smoothly.

Although SGD will make many fluctuations, its speed of reaching convergence is quicker than GD. But in this experiment, it is hard to perform the efficiency of SGD as you see the learning cueve of GD, in fact, even reaching convergence quicker.Besides, the runing time of GD and SGD is closed . I think the dataset is not large enough to perform the efficiency of SGD.