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:
Chu Tao

Supervisor:
Qingyao Wu

Student ID：
201710106550

Grade:
Graduate

December 15, 2017

# Linear Regression, Linear Classification and Gradient Descent

**Abstract—This experiment has test SGD and four methods (NAG，RMSProp，AdaDelta and Adam). In two classification problems, we can find the either SGD and four other methods get good results, but the range of convergence rate is: SGD, AdaDelta, NAG, RMSProp, and Adam. Although the optimization strategy can solve more complex problems, it has sacrificed the speed of convergence.**

## I. Introduction

This experiment in order to comparing and understanding the difference between gradient descent and stochastic gradient descent. And we can also find the differences and relationships between Logistic regression and linear classification. Further understand the principles of SVM and practice on larger data.

## II. Methods and Theory

### A. SGD

When the amount of data is large, it is difficult to not use the Stochastic gradient descent (SGD) method. SGD is very intuitive, that is to randomly take one or a few data to do a gradient descent, that is,

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

### B. NAG

The core idea of NAG (Nesterov accelerated gradient) is to use Momentum to predict the next step of the gradient, rather than using the current gradient.

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

### C. RMSProp

This algorithm is not too simple. It was Hinton mentioned in the class, not even published. RMSProp is to solve the problem of learning rate of 0 in AdaGrad. To see how easy it is.

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

### D. AdaDelta

AdaDelta can also solve the problem of AdaGrad. Though it is often seen as similar to RMSProp, I feel AdaDelta is more advanced, because it doesn't even set the initial learning speed, and AdaDelta is sometimes relatively slow. Update as follows:

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

### E. Adam

First, Adam makes use of the advantages of AdaGrad and RMSProp on sparse data. The correction of the initialized deviation also makes the Adam better.

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

All methods update the gradient according to the progressive delay strategy, although the speed of updating is getting slower and slower, but the stability is stronger and stronger.

## III. Experiment

There are the results of the two experiments. these contain:

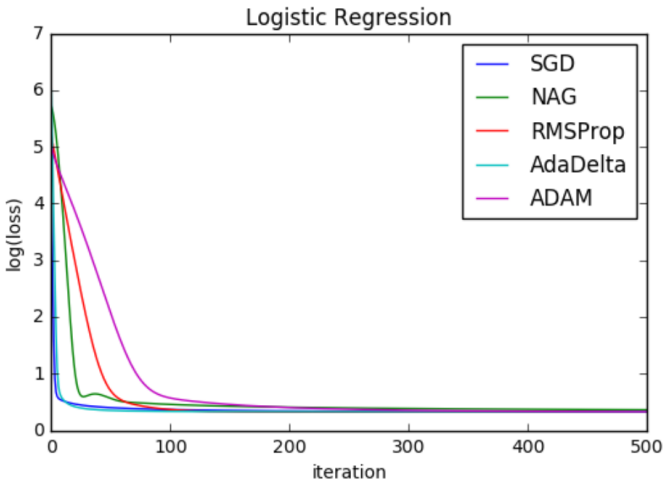


Fig. 1. The loss of SGD and four methods in logistic regression.

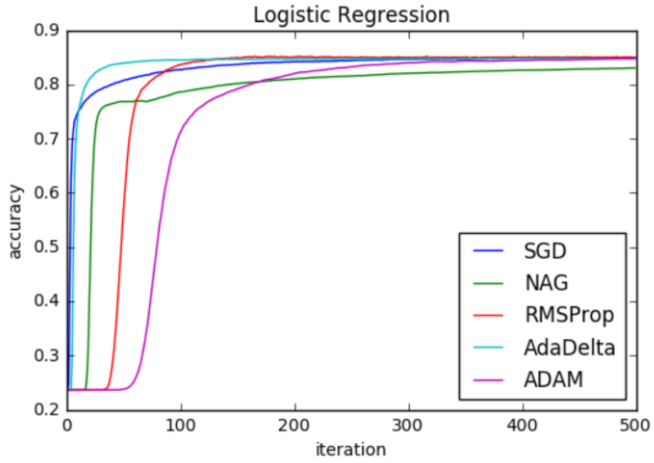


Fig. 2. The accuracy of SGD and four methods in logistic regression.

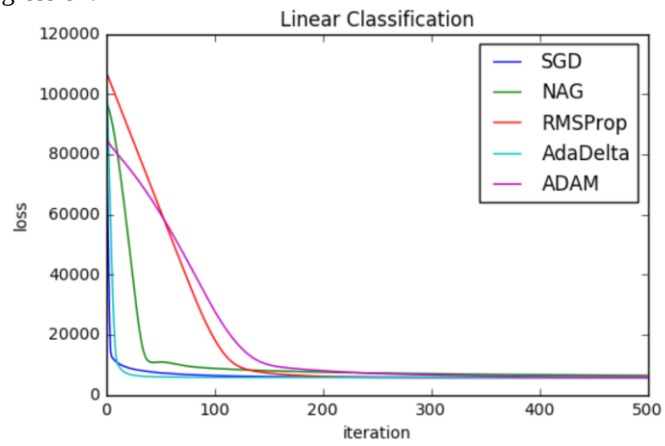


Fig. 3. The loss of SGD and four methods in linear classification.

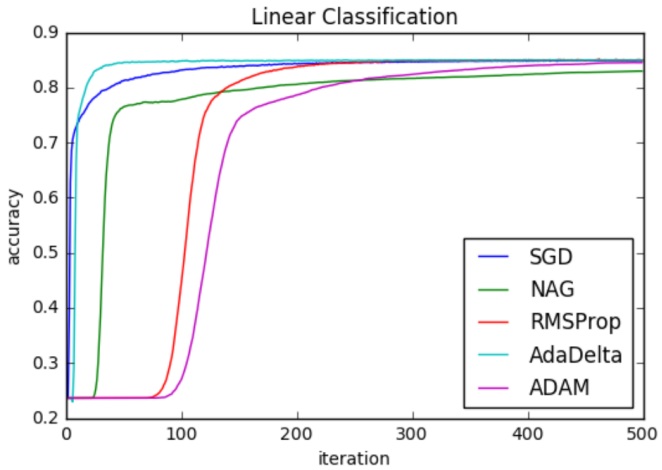


Fig. 4. The accuracy of SGD and four methods in linear classification.

It can be seen from figures that the either SGD and four other methods get good results, but the range of convergence rate is: SGD, AdaDelta, NAG, RMSProp, and Adam.

TABLE I

The final loss of these methods in logistic regression

| Method | Final loss |
|---|---|
| SGD | 0.326419 |
| NAG | 0.360658 |
| RMSProp | 0.324955 |
| AdaDelta | 0.327011 |
| ADAM | 0.328736 |

TABLE II

The final loss of these methods in logistic regression

| Method | Final loss |
|---|---|
| SGD | 5757.918488 |
| NAG | 6435.163494 |
| RMSProp | 5754.962126 |
| AdaDelta | 5749.142760 |
| ADAM | 5859.612263 |

Tables show each methods can get good results.

## IV. CONCLUSION

From the experimental results, it can be seen that in the case of the same learning rate (due to the large change of SGD gradient, the learning rate is lower than the other four), and the four gradually improved versions of SGD, in the simple experiment, regardless of the logic Regression, or linear classification, can optimize the good results, but due to the addition of attenuation coefficient, slow learning, but the benefits are more stable and can tend to the global optimal solution.