The Experiment Report of Machine Learning

SCHOOL: SCHOOL OF SOFTWARE ENGINEERING

SUBJECT: SOFTWARE ENGINEERING

Author:
肖纾予

Supervisor:
Mingkui Tan

Student ID：
201530613160

Grade:
Undergraduate

December 9, 2017

Comparison of Various Stochastic Gradient Descent Methods for Solving Classification Problems

***Abstract*—This report will show the readers the Comparison of Various Stochastic Gradient Descent Methods for Solving Classification Problems and how I solve the problems.**

I. Introduction

In machine learning, we usually face the Empirical risk minimization problem. We have to set a cost to every sample and figure out the average. When we build the model to solve the problem, we have to use iterator and gradient descent is the most popular iterator. However, gradient descent is not always the best tool to solve ERM problems. So we have to learn about more arithmetic to improve gradient descent.

:

II. Methods and Theory

The first: SGD

Related theories: Take one or several data randomly to make a gradient descent

Related equations:

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

Derivation process:

NAG

Related theories: The core idea is to use Momentum to predict the next step, rather than using the current ⊙

Related equations:

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

Derivation process:

RMSProp

Related theories: RMSProp is to solve the problem of learning rate of 0 in AdaGrad. To see how simple it is

Related equations:

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

Derivation process :

Adadelta

Related theories: Though 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.

Related equations:

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

Derivation process:

Adam

Related theories: 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. Why is it called Adam, because it is adaptive estimates of lower-order moments The 1 order moments (mean) and the 2 order moments (variance) are adaptively adjusted.

Related equations:

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

Derivation process:

III. Experiment

A. Datasets

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

B. Implementation

First: Read the experimental data and read the data using the load_svmlight_file function of the sklearn Library

Second: I use all zero initialization to make the parameter initialization.

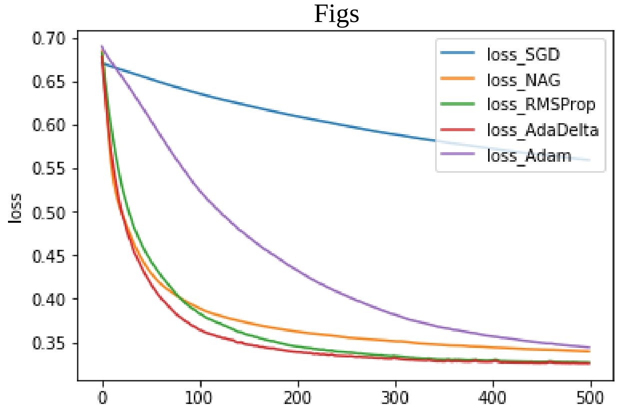
Third: In the model building, at first, I initialized the gradient. Then I figure out the gradient of the training set. Then I choose the average. Then I made the batch gradient descent. Then I figure out the accuracy.

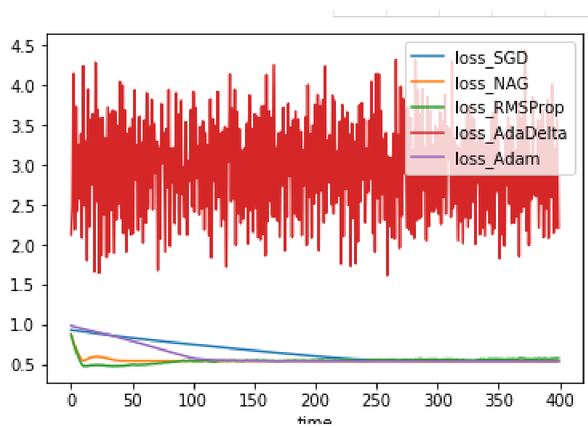
Parameters

NGA parameter	0.9
Learning rate	0.02

Threshold	-0.8
Iteration times	400
RMSProp parameters	0.9 and 1e-8
RMSProp learning rate	0.001
Adadelta parameters	0.95 and 1e-6
Adam parameters	0.999,0.9 and 1e-8

Figs





IV. CONCLUSION

The complexity of the experiment is far more than the last time. I've met a lot of difficulties. However, it let me have a deeper understanding of the linear classification problem and gradient descent.