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:
黄海南

Supervisor:
Mingkui Tan or Qingyao Wu

Student ID：
201720145150

Grade:
Graduate

December 14, 2017

Linear Regression, Linear Classification and Gradient Descent

Abstract—The experiment is mainly about gradient descent and stochastic gradient descent. Besides,we should learn more about the SVM, Logistic regression and linear classification.

I. INTRODUCTION

The motivation of experiment is to compare and understand the difference between gradient descent and stochastic gradient descent and to compare and understand the differences and relationships between Logistic regression and linear classification.What's more,we should further understand the principles of SVM and practice on larger data.

In order to do this,most important step of the experiment is that we should learn to update model parameters using different optimized methods(NAG，RMSProp，AdaDelta and Adam).

II. METHODS AND THEORY

The different optimized methods are as follows.

1) NAG

The core idea is to use Momentum to predict the next level of gradient rather than the current one.

$$\begin{aligned}\mathbf{g}_t &\leftarrow \nabla J(\boldsymbol{\theta}_{t-1} - \gamma \mathbf{v}_{t-1}) \\ \mathbf{v}_t &\leftarrow \gamma \mathbf{v}_{t-1} + \eta \mathbf{g}_t \\ \boldsymbol{\theta}_t &\leftarrow \boldsymbol{\theta}_{t-1} - \mathbf{v}_t\end{aligned} \tag{2-1}$$

2) AdaDelta

Even the initial learning rate is not set, AdaDelta is sometimes relatively slow.

$$\begin{aligned}\mathbf{g}_t &\leftarrow \nabla J(\boldsymbol{\theta}_{t-1}) \\ G_t &\leftarrow \gamma G_t + (1-\gamma)\mathbf{g}_t \odot \mathbf{g}_t \\ \Delta\boldsymbol{\theta}_t &\leftarrow -\frac{\sqrt{\Delta_{t-1}+\epsilon}}{\sqrt{G_t+\epsilon}} \odot \mathbf{g}_t \\ \boldsymbol{\theta}_t &\leftarrow \boldsymbol{\theta}_{t-1} + \Delta\boldsymbol{\theta}_t \\ \Delta_t &\leftarrow \gamma\Delta_{t-1} + (1-\gamma)\Delta\boldsymbol{\theta}_t \odot \Delta\boldsymbol{\theta}_t\end{aligned} \tag{2-2}$$

3) RMSProp

RMSProp is the solution to the problem of learning rate trending 0 in AdaGrad.

$$\begin{aligned}\mathbf{g}_t &\leftarrow \nabla J(\boldsymbol{\theta}_{t-1}) \\ G_t &\leftarrow \gamma G_t + (1-\gamma)\mathbf{g}_t \odot \mathbf{g}_t \\ \boldsymbol{\theta}_t &\leftarrow \boldsymbol{\theta}_{t-1} - \frac{\eta}{\sqrt{G_t+\epsilon}} \odot \mathbf{g}_t\end{aligned} \tag{2-3}$$

4) Adam

Adam exploited the advantages of AdaGrad and RMSProp in sparse data. The correction of the initialization bias also made Adam perform better.

$$\begin{aligned}\mathbf{g}_t &\leftarrow \nabla J(\boldsymbol{\theta}_{t-1}) \\ \mathbf{m}_t &\leftarrow \beta_1 \mathbf{m}_{t-1} + (1-\beta_1)\mathbf{g}_t \\ G_t &\leftarrow \gamma G_t + (1-\gamma)\mathbf{g}_t \odot \mathbf{g}_t \\ \alpha &\leftarrow \eta \frac{\sqrt{1-\gamma^t}}{1-\beta^t} \\ \boldsymbol{\theta}_t &\leftarrow \boldsymbol{\theta}_{t-1} - \alpha \frac{\mathbf{m}_t}{\sqrt{G_t+\epsilon}}\end{aligned} \tag{2-4}$$

A. Logistic Regression and Stochastic Gradient Descent

The Experiment Step are as follows.

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, mark the sample whose predict scores greater than the threshold as positive, on the contrary as negative. 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 loss function are as follows.

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x_i) - y_i) x_i^j \quad (2\text{-}5)$$

B. Linear Classification and Stochastic Gradient Descent

The Experiment Step are as follows.

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, mark the sample whose predict scores greater than the threshold as positive, on the contrary as negative. 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 loss function are as follows.

$$L_i = \sum_{j \neq y_i} [max(0, w_j^T x_i - w_{y_i}^T x_i + \Delta)]$$

$$L = \frac{1}{N} \sum_i \sum_{j \neq y_i} [max(0, f(x_i; W)_j - f(x_i; W)_{y_i} + \Delta)] + \lambda \sum_k \sum_l W_{k,l}^2$$

$$\nabla_{w_{y_i}} L_i = -\left(\sum_{j \neq y_i} \mathbb{1}(w_j^T x_i - w_{y_i}^T x_i + \Delta > 0) \right) x_i \quad (2\text{-}6)$$

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

The environment of experiment is on the basis of python3 including following python package: sklearn，numpy，jupyter，matplotlib. In order to achieve better experimental results. Different optimization methods may set different parameters. The parameters of different optimization methods are set as follows.

1).Logistic Regression and Stochastic Gradient Descent

TABLE I
SIMULATION PARAMETERS

	γ	v	η	∈	G_t	b1	b2	batch	iters
NAG	0.999	0	0.01	--	--	--	--	256	100
RMSProp	0.99	--	0.01	1e-8	0	--	--	256	100
AdaDelta	0.999	--	--	1e-5	0	--	--	256	100
Adam	0.99	0	0.1	1e-5	0	0.9	0.99	256	100

As for AdaDelta, the values of the initial parameter ∂ is 0. Draw graph of different optimized method NAG, RMSProp, AdaDelta and Adam loss with the number of iterations are as follows.

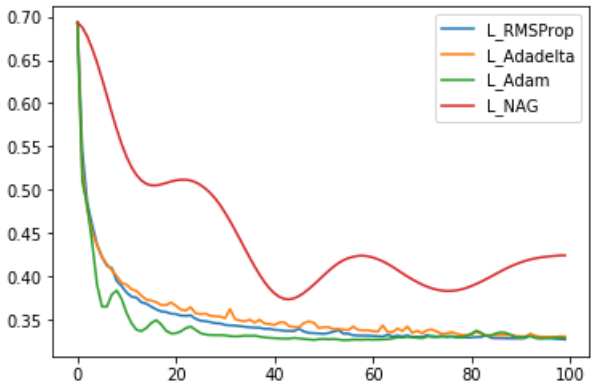

Fig. 1. different optimized method NAG, RMSProp, AdaDelta and Adam loss with the number of iterations(batch=256)

From the Fig.1,we can see that the optimization method of Amda has the best performance,while the NAG is the worst. Because the Amda method converges fastest and the NAG method is volatile.

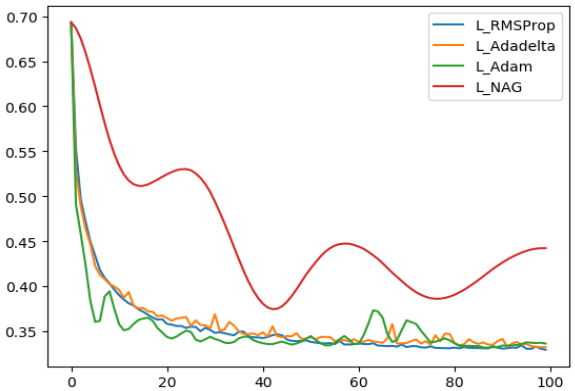

Fig. 2. different optimized method NAG, RMSProp, AdaDelta and Adam loss with the number of iterations(batch=128)

Compare Fig.1 and Fig.2,we can see that the convergence effect of each method is not affected by parameter batch.

2). Linear Classification and Stochastic Gradient Descent

TABLE II
SIMULATION PARAMETERS

	γ	v	η	∈	G_t	b1	b2	batch	iters
NAG	0.999	0	0.01	--	--	--	--	256	100
RMSProp	0.99	--	0.01	1e-8	0	--	--	256	100
AdaDelta	0.999	--	--	1e-8	0	--	--	256	100
Adam	0.99	0	0.1	1e-5	0	0.9	0.99	256	100

As for AdaDelta, the values of the initial parameter ∂ is 0. Draw graph of different optimized method NAG, RMSProp, AdaDelta and LAdam loss with the number of iterations are as follows.

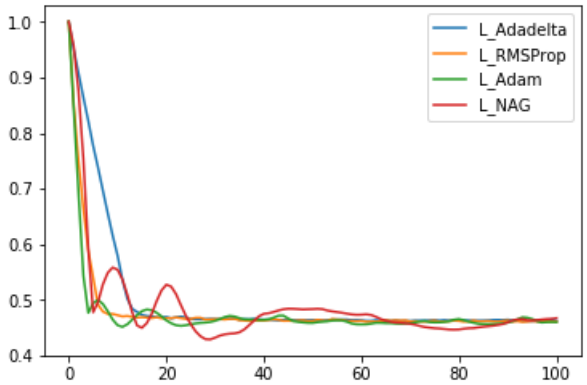

Fig. 3. different optimized method NAG, RMSProp, AdaDelta and Adam loss with the number of iterations(batch=256)

From the picture,we can see that these methods can converge to the nearest minimum.However, the method of AdaDelta converges at the slowest rate. Though he NAG method is volatile,it can converge to the minimum.

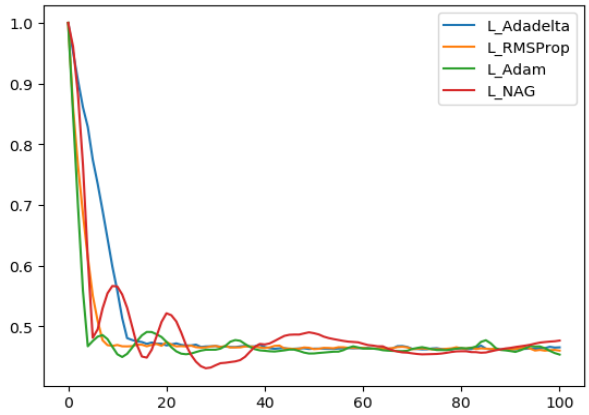

Fig. 4. different optimized method NAG, RMSProp, AdaDelta and Adam loss with the number of iterations(batch=128)

Compare Fig.3 and Fig.4,we can see that the convergence effect of each method is not affected by parameter batch.

IV. CONCLUSION

By comparing these kinds of optimization methods, We can draw the following conclusions.

1)Different optimized method NAG, RMSProp, AdaDelta and LAdam can converge to the nearest minimum.

2)For different problems, each method may have different effects.

3)The effect of each method is also related to the number of iterations. But after a certain number of iterations, the algorithm converges.

4)Various algorithms may not be sensitive to certain parameters. For example, we change the parameter batch, the effect of various optimization methods is basically unaffected.

In general, all kinds of optimization methods NAG, RMSProp, AdaDelta and Adam have some effect, but for different problems and different parameters, the performance of the method can have different effects.

Moreover, the above results show that the optimization method of Amda has the best performance