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:
Huakui Zhang

Supervisor:
Mingkui Tan

Student ID:
201530613610

Grade:
Undergraduate

December 15, 2017

Experimental Study on Stochastic Gradient Descent for Solving Classification Problems

***Abstract*—This experiment is set up to make us further understand the differences of gradient descent and stochastic gradient descent and the differences of logistic regression and linear classification. And from this experiment, we will further understand the principles of SVM and practice on larger data.**

I. Introduction

BATCH gradient descent and stochastic gradient descent are two important ways to apply on the algorithms that use gradient descent. Batch gradient descent computes the gradient using the whole dataset. However, when the dataset is large, the calculation will be expensive as well as other problems. Stochastic gradient descent computes the gradient using only the subset of the dataset, which often leads to faster convergence and gets to a reasonable solution quickly.

The logistic regression comes from the generalized linear regression using the linear equation $y = w^T X + b$ and regressing for the probability of a categorical outcome. And the linear classification is much more geometrically motivated trying to find a optimal separating hyperplane.

II. Methods and Theory

A. Logistic Regression

Logistic Regression use the binary logistic model to estimate the probability of a binary response. In this experiment, we use the cost function with the cross entropy error measure and the regularized term:

$$J(w) = \frac{1}{n}\sum_{i=1}^{n} log(1 + e^{-y_i \cdot w^T x_i}) + \frac{\lambda}{2}\|w\|_2^2,$$

and the gradient of the cost function:

$$\frac{\partial J(w)}{\partial w} = \lambda w - \frac{1}{n}\sum_{i=1}^{n} \frac{y_i x_i}{1+e^{-y_i \cdot w^T x_i}}.$$

B. Linear Classification(SVM)

Linear Classification(SVM) is to use an example's characteristics to identify which class it belongs to. In this experiment, we use the cost function with hinge loss:

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

The hinge loss is

$$\xi_i = max(0, 1 - y_i(w^T x_i + b)).$$

Let

$$g_w(x_i) = \frac{\partial \xi_i}{\partial w}.$$
$$g_b(x_i) = \frac{\partial \xi_i}{\partial b}.$$

So we have:

$$g_w(x_i) = \begin{cases} 0 & 1 - y_i(w^T x_i + b) \le 0 \\ -y_i x_i & 1 - y_i(w^T x_i + b) > 0 \end{cases}$$

$$g_w(x_i) = \begin{cases} 0 & 1 - y_i(w^T x_i + b) \le 0 \\ -y_i & 1 - y_i(w^T x_i + b) > 0 \end{cases}$$

And thus, the gradient is:

$$\frac{\partial J(w,b)}{\partial w} = w + \frac{C}{n}\sum_{i=1}^{n} g_w(x_i)$$
$$\frac{\partial J(w,b)}{\partial b} = \frac{C}{n}\sum_{i=1}^{n} g_b(x_i)$$

C. SGD and Its Different Optimized Methods

In this section, we will introduce the process of stochastic gradient descent(SGD) and its four optimized methods.

Taking linear classification model as an example, the process of SGD are shown in Fig. 1. There are four main variations of SGD - NAG, RMSProp, AdaDelta and Adam as its optimized methods. In the following figures, we will show you the detailed update procedures of the four optimized methods, respectively shown in Fig. 2, Fig. 3, Fig. 4 and Fig. 5.

Algorithm 2: SGD

Initialize parameter $\mathbf{w}$ and learning rate η
while stopping condition is not achieved **do**
  Randomly select an example i in the training set
  $\mathbf{w} = \mathbf{w} - \eta \nabla_{\mathbf{w}} L(\mathbf{w}, b)$
  $b = b - \eta \nabla_b L(\mathbf{w}, b)$
end

Fig. 1. The process of SGD.

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

Fig. 2. The update procedure of NAG.

III. Experiments

A. Dataset

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

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

Fig. 3. The update procedure of RMSProp.

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

Fig. 4. The update procedure of AdaDelta.

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

Fig. 5. The update procedure of Adam.

B. Implementation

In this part we will describe how we do our experiment in details.

Logistic Regression and Stochastic Gradient Descent

1. Load the training set and validation set.
2. Initialize 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. Repeat 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. We initialize the gradient with zero. The rest of the paramenters are shown in TABLE I, TABLE II, TABLE III, TABLE IV, TABLE V and TABLE VI, the loss figure is shown in Fig. 6.

TABLE I
LOGISTIC REGRESSION SIMULATION PARAMETERS

Batch Size	1024
Number of Epoch	4

TABLE II
LOGISTIC REGRESSION SGD PARAMETERS

λ	0.0001
Learning Rate	0.5

TABLE III
LOGISTIC REGRESSION NAG PARAMETERS

λ	0.0001
Learning Rate	0.2
γ	0.9

TABLE IV
LOGISTIC REGRESSION RMSPROP PARAMETERS

λ	0.0001
Learning Rate	0.015
γ	0.9
ϵ	1e-4

TABLE V
LOGISTIC REGRESSION ADADELTA PARAMETERS

λ	0.0001
γ	0.95
ϵ	1e-4

TABLE VI
LOGISTIC REGRESSION ADAM PARAMETERS

λ	0.0001
Learning Rate	0.05
γ	0.999
ϵ	1e-4
δ	0.9

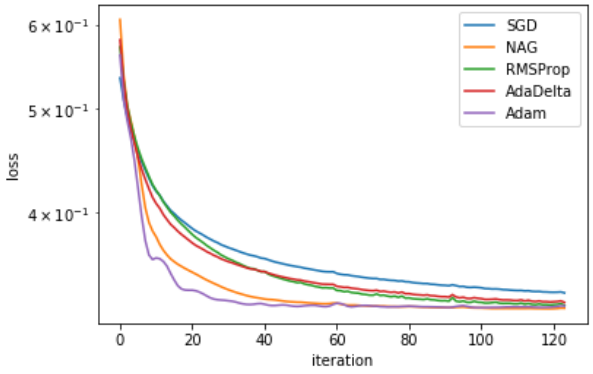

Fig. 6. Logistic Regression Loss Figure.

Linear Classification and Stochastic Gradient Descent

1. Load the training set and validation set.
2. Initailize 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. Repeat 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.

We initialize the gradient with zero. The rest of the paramenters are shown in TABLE VII, TABLE VIII, TABLE IX, TABLE X, TABLE XI and TABLE XII, the loss figure is shown in Fig. 7.

TABLE VII
SVM SIMULATION PARAMETERS

Batch Size	1024
Number of Epoch	4

TABLE VIII
SVM SGD PARAMETERS

C	21
Learning Rate	0.01

TABLE IX
SVM NAG PARAMETERS

C	21
Learning Rate	0.01
γ	0.9

TABLE X
SVM RMSPROP PARAMETERS

C	21
Learning Rate	0.01
γ	0.9
ϵ	1e-4

TABLE XI
SVM ADADELTA PARAMETERS

C	21
γ	0.95
ϵ	1e-4

TABLE XII
SVM ADAM PARAMETERS

C	21
Learning Rate	0.012
γ	0.999
ϵ	1e-4
δ	0.9

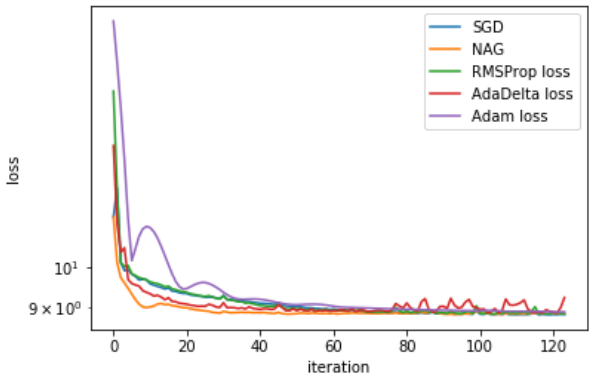

Fig. 7. SVM Loss Figure.

IV. CONCLUSION

In this experiment, we implement two models and four different SGD optimized algorithm, and compared their performance.We can draw a conclusion that compared to batch gradient descent, SGD performs better after parameter tuning.