South China University of Technology

The Experiment Report of Machine Learning

SCHOOL: SCHOOL OF SOFTWARE ENGINEERING

SUBJECT: SOFTWARE ENGINEERING

Author:
DingKun Song

Supervisor:
Mingkui Tan or Qingyao Wu

Student ID：
201720145006

Grade:
Undergraduate

December 12, 2017

Logistic Regression, Linear Classification and Stochastic Gradient Descent

Abstract—The experiment is to use Logistic Regression to analysis a9a Data and use Linear Classification to analysis a9a Data.

I. Introduction

A. *Logistic Regression and Stochastic Gradient Descent*

This experiment is use Logistic Regression to find the best model fucntion to fix the dataset.

We need to use the a9a tranning set to train our model function and compute the loss function.

We use Stochastic Gradient Descent to update w.

We must Update model parameters using different optimized methods(NAG，RMSProp，AdaDelta and Adam).

Finally,use validation set to validate the model function and caculate the loss if these methods.Then show the loss as picture.

B. Linear Classification And *Stochastic Gradient Descent*

This experiment is use Linear Classification to find the best model fucntion to seperate the dataset.

We need to use the a9a tranning set to train our model function and compute the loss function.

We use Stochastic Gradient Descent to update w.

We must Update model parameters using different optimized methods(NAG，RMSProp，AdaDelta and Adam).

Finally,use validation set to validate the model function and caculate the loss if these methods.Then show the loss as picture.

II. Methods and Theory

A. *Logistic Regression and Stochastic Gradient Descent*

First,defined the model function as

$$h(x) = g\left(\sum_{i=1}^{m} w_i x_i\right) = g(\mathrm{w}^T X)$$

$$g(z) = \frac{1}{1+e^{-z}}$$

Second,find the loss function.

$$J(w) = -\frac{1}{n}\left[\sum_{i=1}^{n} y_i \log h_w(x_i) + (1-y_i)\log(1-h_w(x_i))\right]$$

Third,minimizing the loss function use Gradient Descent.

$$\frac{\partial L(w)}{\partial w} = \frac{1}{n}\sum_{i=1}^{n}(h_w(x_i) - y)x_i$$

We use different methods to update model parameters

SGD

$$g_t \leftarrow \nabla J_i(\theta_{t-1})$$
$$\theta_t \leftarrow \theta_{t-1} - \eta g_t$$

NAG

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

RMSProp

$$g_t \leftarrow \nabla J(\theta_{t-1})$$
$$G_t \leftarrow \gamma G_t + (1-\gamma) g_t * g_t$$
$$\theta_t \leftarrow \theta_{t-1} - \frac{\eta}{\sqrt{G_t + \varepsilon}} * g_t$$

AdaDelta

$$g_t \leftarrow \nabla J(\theta_{t-1})$$
$$G_t \leftarrow \gamma G_t + (1-\gamma) g_t * g_t$$
$$\Delta\theta_t \leftarrow -\frac{\sqrt{\Delta_{t-1} + \varepsilon}}{\sqrt{G_t + \varepsilon}} * g_t$$
$$\theta_t \leftarrow \theta_{t-1} + \Delta\theta_t$$
$$\Delta_t \leftarrow \gamma \Delta_{t-1} + (1-\gamma)\Delta\theta_t * \Delta\theta_t$$

Adam

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

We use these to update w and find the best w to minimize loss function.

B. Linear Classification And *Stochastic Gradient Descent*

First,defined the model function as

$$f(x) = w_0 + w_1x_1 + w_2x_2 + ... + w_mx_m = \mathrm{w}^T X$$

Second,find the loss function.

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

Third,minimizing the loss function use Gradient Descent.

$$g_{\mathbf{w}}(\mathbf{x}_i) = \begin{cases} -y_i\mathbf{x}_i & 1 - y_i(\mathbf{w}^\top\mathbf{x}_i + b) >= 0 \\ 0 & 1 - y_i(\mathbf{w}^\top\mathbf{x}_i + b) < 0 \end{cases}$$

$$\frac{\partial \mathrm{L}(w)}{\partial w} = \mathrm{w} + \frac{C}{n}\sum_{i=1}^{n} g_w(x_i)$$

We use different methods to update model parameters

SGD

$$g_t \leftarrow \nabla J_i(\theta_{t-1})$$
$$\theta_t \leftarrow \theta_{t-1} - \eta g_t$$

NAG

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

RMSProp

$$g_t \leftarrow \nabla J(\theta_{t-1})$$
$$\mathrm{G}_t \leftarrow \gamma \mathrm{G}_t + (1-\gamma) g_t * g_t$$
$$\theta_t \leftarrow \theta_{t-1} - \frac{\eta}{\sqrt{\mathrm{G}_t + \varepsilon}} * g_t$$

AdaDelta

$$g_t \leftarrow \nabla J(\theta_{t-1})$$
$$\mathrm{G}_t \leftarrow \gamma \mathrm{G}_t + (1-\gamma) g_t * g_t$$
$$\Delta\theta_t \leftarrow -\frac{\sqrt{\Delta_{t-1} + \varepsilon}}{\sqrt{\mathrm{G}_t + \varepsilon}} * g_t$$
$$\theta_t \leftarrow \theta_{t-1} + \Delta\theta_t$$
$$\Delta_t \leftarrow \gamma \Delta_{t-1} + (1-\gamma)\Delta\theta_t * \Delta\theta_t$$

Adam

$$g_t \leftarrow \nabla J(\theta_{t-1})$$
$$\mathrm{m}_t \leftarrow \beta_1 m_{t-1} + (1-\beta_1) g_t$$
$$\mathrm{G}_t \leftarrow \gamma \mathrm{G}_t + (1-\gamma) g_t * g_t$$
$$\alpha \leftarrow \eta \frac{\sqrt{1-\gamma^t}}{\sqrt{1-\beta^t}}$$
$$\theta_t \leftarrow \theta_{t-1} - \alpha \frac{\mathrm{m}_t}{\sqrt{\mathrm{G}_t + \varepsilon}}$$

We use these to update w and find the best w to minimize loss function.

III. Experiment

A. Logistic Regression and Stochastic Gradient Descent

A. Dataset

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

B Implementation

The Initialization value is showned ad the table.

sgd_rate	0.001
nag_rate	0.001
rmsp_rate	0.001
adadel_rate	0.001
adam_rate	0.001
nga_yta	0.9
rmsp_yta	0.9
adadel_yta	0.95
adam_yta	0.999
ε	10**(-8)
β	0.9
Tranning time	1000

Then I use the formula above to caculate loss function and update the w.

I use array,loss_train loss_test, to save the loss of validation set.

Finally,I use matplotlib to show the loss_test.

B. Linear Classification And *Stochastic Gradient Descent*

A. Dataset

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

B. Implement

The Initialization value is showned ad the table.

sgd_rate	0.001
nag_rate	0.001
rmsp_rate	0.001
adadel_rate	0.001
adam_rate	0.001
nga_yta	0.9
rmsp_yta	0.9
adadel_yta	0.95
adam_yta	0.999
ε	10**(-8)

β	0.9
C	0.1

Then I use the formula above to caculate loss function and update the w.

I use array,loss_test, to save the loss of validation set.

Finally,I use matplotlib to show the loss_test.

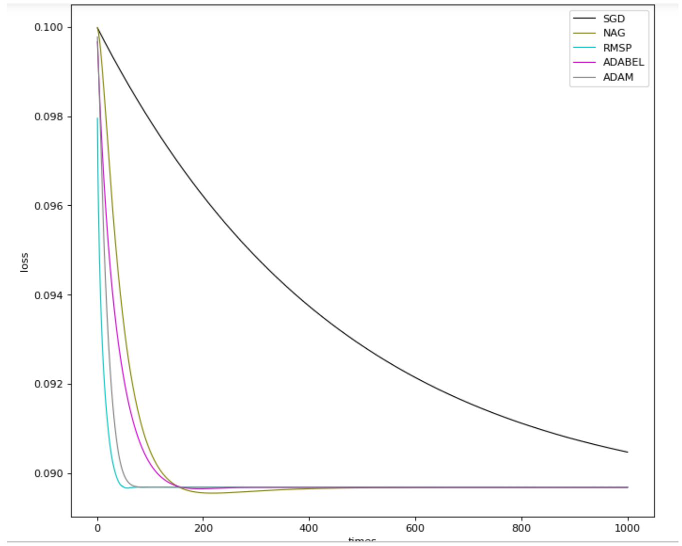
IV. CONCLUSION

A,Logistic Regression and Stochastic Gradient Descent

With the appeals method, as the number of training increases, the loss is getting smaller and smaller, finally tends to be smooth.

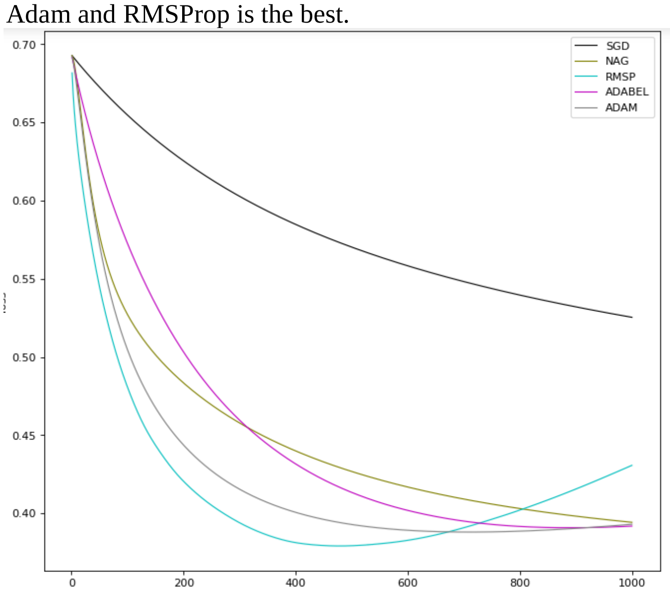
This is mean we find the best w.

The result is showned as follow.



As we can see, the loss of RMSProp first drop and then rise,I think is that the rate of it is too large so that it is descent to quickly.The loss of SGD,NAG,AdaDelta,Adam is firt drop and then smooth,so that it find the best w.

The loss of Adam and RMSProp is drop faster than other,and the SGD is the lowest,so that the convergence speed of Adam and RMSProp is the best.



B. Linear Classification And *Stochastic Gradient Descent*

With the appeals method, as the number of training increases, the loss is getting smaller and smaller, finally tends to be smooth.

This is mean we find the best w.

The result is showned as follow.

As we can see,the loss of Adam and RMSProp is drop faster than other,and the SGD is the lowest,so that the convergence speed of Adam and RMSProp is the best,and the SGD is the worst of these methods.