South China University of Technology

The Experiment Report of Machine Learning

SCHOOL: SCHOOL OF SOFTWARE ENGINEERING

SUBJECT: SOFTWARE ENGINEERING

Author:
Shaojia hong

Supervisor:
Qingyao Wu

Student ID：
201720144993

Grade:
Undergraduate

December 14, 2017

Logistic Regression, Linear Classification and Gradient Descent

Abstract—The experiment is further understand of the difference and connection between the gradient descent and the random gradient descent,and the difference and connection between logistic regression and linear classification

I. INTRODUCTION

A. Logistic regression

Logistic regression is a linear classification model. The difference between linear regression and linear regression is that in order to output large numbers of linear regression, for example, from negative infinity to positive infinity, it is compressed to 0 and 1.only need a logistic function is that

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

B. Linear Classification

Linear Classification is a Classification that given training data (xi, yi) for i = 1 . . . n, with $x_i \in R^m$ and $y_i \in \{-1, 1\}$, learn a classfier f(x) such that $f(x_i)\begin{cases}\geq 0 & y_i=+1 \\ <0 & y_i=-1\end{cases}$ and $y_i f(x_i) > 0$ for a correct classification.

In order to further understand of the difference and connection between the gradient descent and the random gradient descent,and the difference and connection between logistic regression and linear classification is compared.Finally further understand the principle of SVM and practice it on larger data.Finally we use the SGD,NAG,RMSProp, AdaDelta, and Adam of gradient methods to gradient descent,and compare the loss of five methods.

II. METHODS AND THEORY

A. Logistic regression

1) Experimental environment

Python3 and at least the following Python packages are included such as sklearn，numpy，jupyter，matplotlib.
It is recommended to install anaconda3 directly, which has built in the above Python packages.The experimental code and drawing are all done on jupyter.

2) Step

The step of Logistic regression and gradient Descent：

1.Read the experimental training set and the validation set.
2.init the parameter of logistic regression model ,the initialization can consider all zero initialization, random initialization or normal distribution initialization.
3.Select the Loss function and seek guidance for it. The process is detailed in the courseware ppt.
4.The gradient of a partial sample to the Loss function G is obtained.
5.Update the model parameters using different optimization methods (NAG, RMSProp, AdaDelta, and Adam).
6.Choosing the appropriate threshold, we will verify that the mark of the concentrated calculation is more than the threshold as a positive class, and otherwise as a negative class. Test and get the Loss function values $L_{NAG}, L_{RMSProp}, L_{AdaDelta}$和$L_{Adam}$ of different optimization methods on the validation set.
7.Repeat step 4-6 several times, draw the graph of $L_{NAG}, L_{RMSProp}, L_{AdaDelta}$和$L_{Adam}$, and change graphs with the number of iterations.

3) Formula

Logistic function is

$$h_w(x_i) = \frac{1}{1+e^{-wx}}$$

Target function is

$$w = w - \frac{1}{n}\sum_{i=1}^{n}\alpha(h_w(x_i) - y_i)x_i$$

Loss function is

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

we use these two formula to update w and find the best w to minimize the J,and we use five different method to update w,

SGD method use

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

NAG method use

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

RMSProp method use

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

$$G_t \leftarrow \gamma G_{t-1} + (1-\gamma) g_t \Theta g_t$$

$$\theta_t \leftarrow \theta_{t-1} - \frac{\eta}{\sqrt{G_t + \varepsilon}} \Theta g_t$$

AdaDelta method use

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

$$G_t \leftarrow \gamma G_t + (1-\gamma) g_t \Theta g_t$$

$$\Delta\theta_t \leftarrow -\frac{\sqrt{\Delta_{t-1} + \varepsilon}}{\sqrt{G_t + \varepsilon}} \Theta g_t$$

$$\theta_t \leftarrow \theta_{t-1} - \Delta\theta_t$$

$$\Delta_t \leftarrow \gamma \Delta_{t-1} + (1-\gamma) \Delta\theta_t \Theta \Delta\theta_t$$

Adam method use

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

$$m_t \leftarrow \beta_1 m_{t-1} + (1-\beta_1) g_t$$

$$G_t \leftarrow \gamma G_t + (1-\gamma) g_t \Theta g_t$$

$$\alpha \leftarrow \eta \frac{\sqrt{1-\gamma^t}}{\sqrt{1-\beta^t}}$$

$$\theta_t \leftarrow \theta_{t-1} - \alpha \frac{m_t}{\sqrt{G_t + \varepsilon}}$$

B. Linear Classification

1) Experimental environment

Python3 and at least the following Python packages are included such as sklearn，numpy，jupyter，matplotlib.
It is recommended to install anaconda3 directly, which has built in the above Python packages.The experimental code and drawing are all done on jupyter.

2) Step

The step of Linear Classification and gradient Descent：

1.Read the experimental training set and the validation set.
2.The support vector machine model parameter initialization can consider all zero initialization, random initialization or normal distribution initialization.
3.Select the Loss function and seek guidance for it. The process is detailed in the courseware ppt.
4.The gradient of a partial sample to the Loss function G is obtained.
5.Update the model parameters using different optimization methods (NAG, RMSProp, AdaDelta, and Adam).
6.Choosing the appropriate threshold, we will verify that the mark of the concentrated calculation is more than the threshold as a positive class, and otherwise as a negative class. Test and get the Loss function values $L_{NAG}, L_{\mathrm{RMSProp}}, L_{AdaDelta}$和$L_{Adam}$ of different optimization methods on the validation set.
7.Repeat step 4-6 several times, draw the graph of $L_{NAG}, L_{\mathrm{RMSProp}}, L_{AdaDelta}$和$L_{Adam}$, and change graphs with the number of iterations.

3) Formula

Target function is

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

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

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

Loss function is

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

we use these two formula to update w and find the best w to minimize the J,and we use five different method to update w,

SGD method use

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

$$\theta_t \leftarrow \theta_{t-1} - \eta g_t$$

NAG method use

$$g_t \leftarrow \nabla J_i(\theta_{t-1} - \gamma v_{t-1})$$

$$v_t \leftarrow \gamma v_{t-1} + \eta g_t$$

$$\theta_t \leftarrow \theta_{t-1} - v_t$$

RMSProp method use

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

$$G_t \leftarrow \gamma G_{t-1} + (1-\gamma) g_t \Theta g_t$$

$$\theta_t \leftarrow \theta_{t-1} - \frac{\eta}{\sqrt{G_t + \varepsilon}} \Theta g_t$$

AdaDelta method use

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

$$G_t \leftarrow \gamma G_t + (1-\gamma) g_t \Theta g_t$$

$$\Delta\theta_t \leftarrow -\frac{\sqrt{\Delta_{t-1}+\varepsilon}}{\sqrt{G_t+\varepsilon}} \Theta g_t$$

$$\theta_t \leftarrow \theta_{t-1} - \Delta\theta_t$$

$$\Delta_t \leftarrow \gamma\Delta_{t-1} + (1-\gamma)\Delta\theta_t \Theta \Delta\theta_t$$

Adam method use

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

$$m_t \leftarrow \beta_1 m_{t-1} + (1-\beta_1) g_t$$

$$G_t \leftarrow \gamma G_t + (1-\gamma) g_t \Theta g_t$$

$$\alpha \leftarrow \eta \frac{\sqrt{1-\gamma^t}}{\sqrt{1-\beta^t}}$$

$$\theta_t \leftarrow \theta_{t-1} - \alpha \frac{m_t}{\sqrt{G_t+\varepsilon}}$$

III. Experiment

A. Logistic regression

1) DataSet

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

2) Implementation

1. Logistic regression parameter is such
iteration=500
SGD method parameter:
η=0.005
NAG method parameter:
$\gamma = 0.9$
η=0.005
RMSProp method parameter:
$\gamma = 0.9$
$\varepsilon = 10^{-6}$
η=0.001
AdaDelta method parameter:
$\gamma = 0.95$
$\varepsilon = 10^{-6}$
Adam method parameter:
β=0.9
$\gamma = 0.999$
$\varepsilon = 10^{-6}$
η=0.001

I use the above formula to update w and use array to save loss value,then show the loss result using matplotlib.

B. Linear Classification

1) DataSet

Linear Classification use a9a of LIBSVM Data,including 32561/16281(testing) samples and each sample has 123/123(testing) features

2) Implementation

Linear Classification parameter is such
SGD method parameter:
η=0.001
NAG method parameter:
$\gamma = 0.9$
η=0.001
RMSProp method parameter:
$\gamma = 0.9$
$\varepsilon = 10^{-8}$
η=0.001
AdaDelta method parameter:
$\gamma = 0.9$
$\varepsilon = 10^{-6}$
Adam method parameter:
β=0.9
$\gamma = 0.999$
$\varepsilon = 10^{-6}$
η=0.001
I use the above formula to update w and use array to save loss value,then show the loss result using matplotlib.

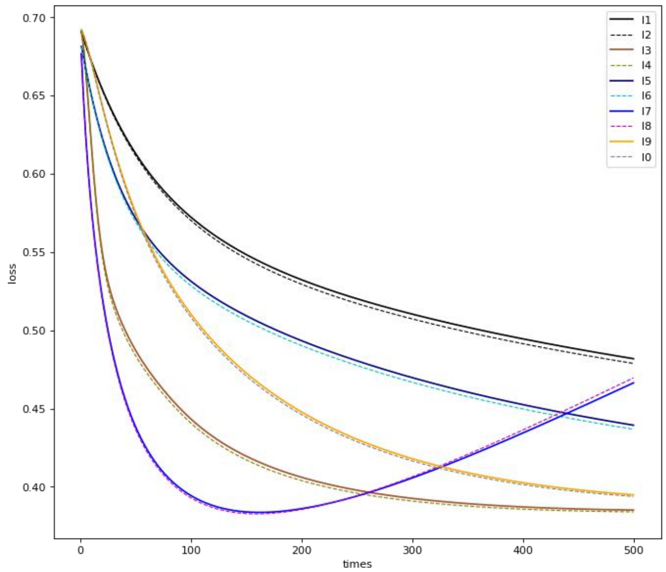
IV. Conclusion

A. Logistic regression

we can see that select different descent method, the loss descent rate are different.The SGD method drop the slowest ,then is Adam,RMSProp ,AdaDelta ,NAG,AdaDelta.

l1,l2 is SGD method loss and test loss line
l3,l4 is NAG method loss and test loss line
l5,l6 is RMSProp method loss and test loss line

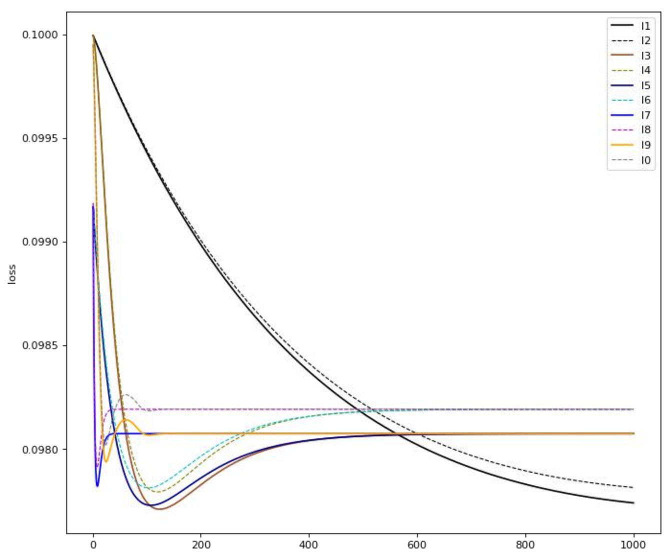
l7,l8 is AdaDelta method loss and test loss line
l9,l0 is Adam method loss and test loss line



B. Linear Classification

we can see that select different descent method, the loss descent rate are different.The SGD method drop the slowest ,then is NAG ,RMSProp,Adam,AdaDelta.

l1,l2 is SGD method loss and test loss line
l3,l4 is NAG method loss and test loss line
l5,l6 is RMSProp method loss and test loss line
l7,l8 is AdaDelta method loss and test loss line
l9,l0 is Adam method loss and test loss line



C. Summary

Through this experiment,I further understood the principle of logistic regression and linear classification and the SGD,NAG,RMSProp, AdaDelta, and Adam gradient descent. By learning and compare the five gradient descent method, we can further understand the important content of the gradient learning, and realize the process of optimizing and adjusting the parameters.