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:
Fangyuan Ding

Supervisor:
Qingyao Wu

Student ID：201720144979

Grade:
Graduate

December 14, 2017

Logistic Regression, Linear Classification and Stochastic Gradient Descent

Abstract—
Compare and understand the difference between gradient descent and stochastic gradient descent.Compare and understand the differences and relationships between Logistic regression and linear classification.Further understand the principles of SVM and practice on larger data.

I. Introduction

In the experiment, we use the Logistic Regression and Linear Classification model to solve the classification problem. And use different optimized methods(NAG,RMSProp,AdaDelta and Adam) to train the model in order to Compare and understand the difference between gradient descent and stochastic gradient descent.

II. Methods and Theory

A. Logistic Regression and Stochastic Gradient Descent

1) The loss function of logistic regression

$$\text{Loss} = -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1-y^{(i)}) \log(1-h_\theta(x^{(i)}))]$$

2) Derivatives

$$\frac{\partial Loss}{\partial w} = -\frac{1}{m}\sum_{i=1}^{m} \alpha(h_\theta(x^{(i)}) - y^{(i)})x^{(i)}$$

B. Linear Classification and Stochastic Gradient Descent

1)The loss function of linear classification

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

2) Derivatives

$$g_t = \begin{cases} w + C\sum_{i=1}^{n} -x_i^T y_i & 1-y_i(x_i w+b) \geq 0 \\ w & 1-y_i(x_i w+b) < 0 \end{cases}$$

C. Gradient descent optimized methods

1)NAG

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

2)RMSProp

$$g_t = \nabla J(\theta_{t-1})$$
$$G_t = \gamma G_t + (1-\gamma) g_t \odot g_t$$
$$\theta_t = \theta_{t-1} - \frac{\eta}{\sqrt{G_t+\epsilon}} \odot g_t$$

3)AdaDelta

$$g_t = \nabla J(\theta_{t-1})$$
$$G_t = \gamma G_t + (1-\gamma) g_t \odot g_t$$
$$\Delta\theta_t = -\frac{\sqrt{\Delta_{t-1}+\epsilon}}{\sqrt{G_t+\epsilon}} \odot g_t$$
$$\theta_t = \theta_{t-1} + \Delta\theta_t$$
$$\Delta_t = \gamma\Delta_{t-1} + (1-\gamma)\Delta\theta_t \odot \Delta\theta_t$$

4)Adam

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

III. Experiment

A. Dataset

In this experiment, we use a9a of LIBSVM Data, including 32561/16281(testing) samples and each sample has 123/123 (testing) features.

B. Experimental steps

1) 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.

2) Linear Classification and Stochastic Gradient Descent

1.Load the training set and validation set.

2.Initalize SVM model parameters, you can consider initalizing zeros, random numbers or normal distribution.

3.Select the loss function and calculate its derivation, find more detail in PPT.

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.

C .initialization model parameters

1)Logistic Regression and Stochastic Gradient Descent

All parameters are set into zero in the linear model.

2)Linear Classification and Stochastic Gradient Descent

All parameters are set into zero in the SVM model.

D. Experimental results and curve

For all gradient descent optimized methods,hyper-parameter were selected as follows:

NAG: γ=0.9 η=0.5

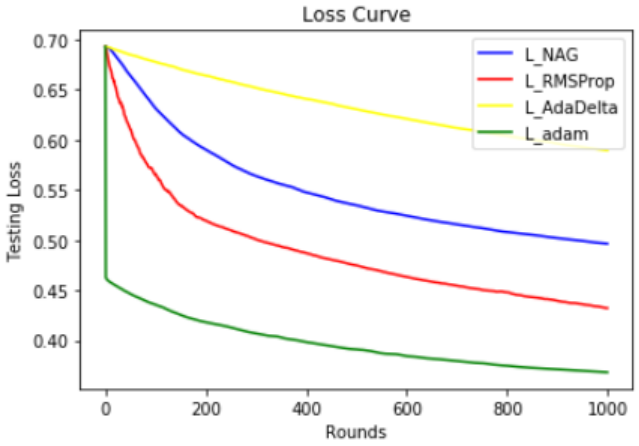
RMSProp: γ=0.9 η=0.5

AdaDelta: γ=0.95 Δt= 0

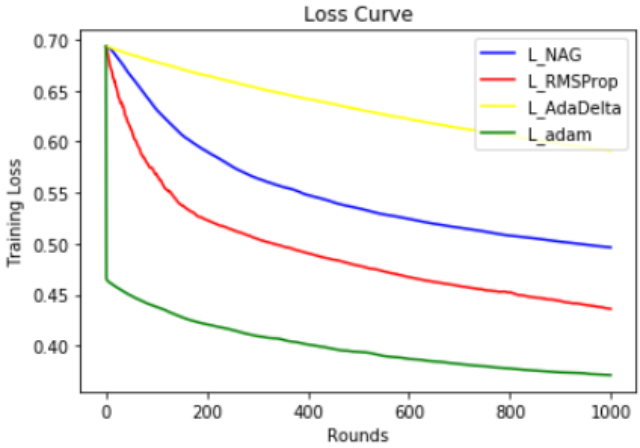
Adam: βt = 0.9 γ = 0.999 η = 0.05

1) Logistic Regression and Stochastic Gradient Descent

Testing Loss:



Training Loss:

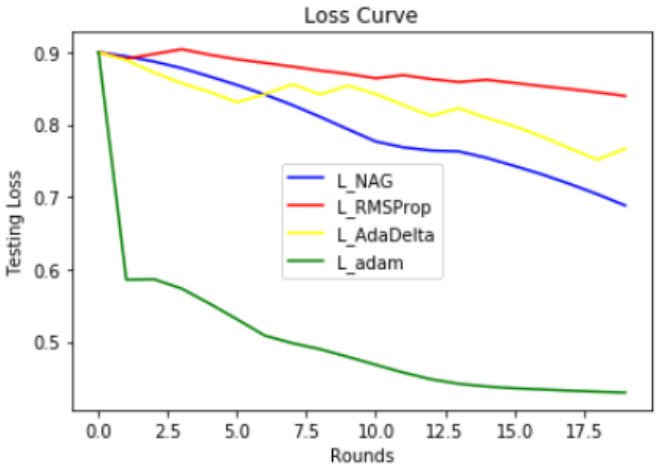


Accuracy:

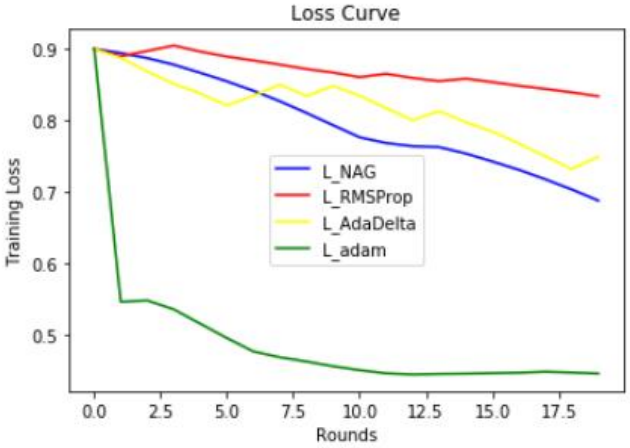
```
training accuracy_nag= 0.7591904425539756    testing accuracy_nag= 0.7637737239727289
training accuracy_rms= 0.7773102791683302    testing accuracy_rms= 0.7813402125176586
training accuracy_adad= 0.7591904425539756    testing accuracy_adad= 0.7637737239727289
training accuracy_adam= 0.8284450723257886    testing accuracy_adam= 0.8298016092377618
```

2) Linear Classification Loss:

Testing Loss:



Training Loss:



Accuracy:

```
training accuracy_nag= 3.071158748195694e-05    testing accuracy_nag= 6.142128861863522e-05
training accuracy_rms= 3.071158748195694e-05    testing accuracy_rms= 6.142128861863522e-05
training accuracy_adad= 3.071158748195694e-05    testing accuracy_adad= 6.142128861863522e-05
training accuracy_adam= 3.071158748195694e-05    testing accuracy_adam= 6.142128861863522e-05
```

IV. CONCLUSION

Finally,I understand the similarities and the differences between logistic regression and linear classification .

For the similarities,two models slove the classification problem.and both of them are linear model.

For the differences: the logistic regression can represent a probability throw mapping features on the sigmoid function. And the linear classification classify the datas by training a hyperplane to split the datas.