華南理工大學

South China University of Technology

The Experiment Report of Machine Learning

SCHOOL: SCHOOL OF SOFTWARE ENGINEERING

SUBJECT: SOFTWARE ENGINEERING

Author: Shoubin Li	Supervisor: Mingkui Tan
Student ID： 201530612040	Grade: Undergraduate

December 12, 2017

Linear Regression, Linear Classification and Gradient Descent

Abstract—In order to compare and understand the difference between gradient descent and stochastic gradient descent. We did this experiment, using different optimization methods (NAG, RMSProp, AdaDelta and Adam) and compare the effectiveness.

I. Introduction

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.

II. METHODS AND THEORY

. **Logistic Regression:**

$$\sigma(t) = \frac{e^t}{e^t+1} = \frac{1}{1+e^{-t}}$$

$$h_w(x) = g(w^T x) = \frac{1}{1+e^{-w^T x}}$$

Loss fuction

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

Derivation

$$\frac{\partial J(w)}{\partial w} = -y\frac{1}{h_w(x)}\frac{\partial h_w(x)}{\partial w} + (1-y)\frac{1}{1-h_w(x)}\frac{\partial h_w(x)}{\partial w}$$

Linear Classification(SVM):

Loss fuction；

$$J(w) = \frac{1}{2}\|w\|^2 + C\sum_i \max(0, 1-y_i(w_i^x+b))$$

Derivation

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

Four optimized methods：

1. NAG

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

2. RMSProp:

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

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

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

NAG	η = 0.01, γ =0.9
RMSProp	η = 0.01, γ = 0.9
AdaDelta	η =0.01, γ = 0.99
Adam	η =0.01, β 1 = 0.9, β 2 = 0.99

NAG	η = 1*10^-6, ρ = 0.9
RMSProp	η = 0.01, γ = 0.9
AdaDelta	η =0.01, γ = 0.9
Adam	η =0.01, β 1 = 0.99, β 2 = 0.999

III. Experiment

Logistic Regression and Stochastic Gradient Descent

1.Load the training set and validation set.

2.Initalize logistic regression 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 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.

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 toward loss function from partial samples.

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

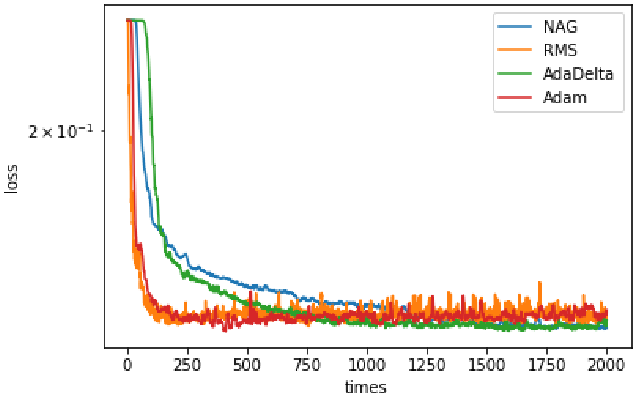
6.Select the appropriate threshold, mark the sample whose predict scores greater than the threshold as positive, on the contrary as negative. Predict under validation set and get the different optimized method loss L_NAG, L_RMSProp, L_AdaDelta and L_Adam.

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.

IV. CONCLUSION

We can compare four methods’ performance through the graph

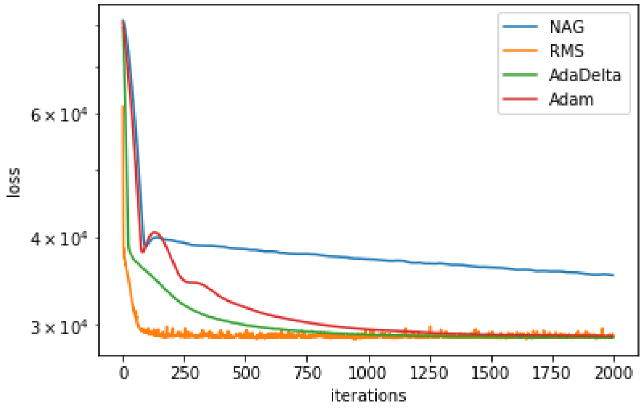
Logistic Regression



RMS decline fastest with η = 0.01, γ = 0.9, but constantly shaking.

NAG decline slower than other method with η = 0.01, γ =0.9. As the iterations increase. AdaDelta performs best

Linear Classification:



We can see that the RMS method is also the quickest, and the NGA still the lowest. The curve of AdaDelta is more stable then other three curves.