华南理工大学

South China University of Technology

The Experiment Report of Machine Learning

SCHOOL: SCHOOL OF SOFTWARE ENGINEERING

SUBJECT: SOFTWARE ENGINEERIN
G

Author:
Shoukai Xu and Yaofu Chen

Supervisor:
Mingkui Tan or Qingyao Wu

Student ID:
201530611111 and 20153060000

Grade:
Undergraduate or Graduate

December 9, 2017

Linear Regression, Linear Classification and Gradient Descent

Abstract—Use a variety of optimization methods of gradient descent to achieve logistic regression and linear classification

I. Introduction

This experiment is to achieve the logistic regression and linear regression, through several methods of stochastic gradient descent. At the same time compare and understand the differences and relationships between Logistic regression and linear classification. And further understand the principles of SVM and practice on larger data.

II. Methods and Theory

(1)Logistic Regression

In statistics, logistic regression, or logit regression, or logit model is a regression model where the dependent variable (DV) is categorical. This experiment covers the case of a binary dependent variable—that is, where the output can take only two values, "0" and "1", which represent outcomes such as pass/fail, win/lose, alive/dead or healthy/sick.

The logistic regression can be understood simply as finding the β parameters that best fit:

$$y = \begin{cases} 1 & \beta_0 + \beta_1 x + \varepsilon > 0 \\ 0 & \text{else} \end{cases}$$

where ε is an error distributed by the standard logistic distribution. (If the standard normal distribution is used instead, it is a probit model.)

The loss function is calculated using the sigmoid function

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

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}},$$

$$L_D(\mathbf{w}) = -\frac{1}{m}\sum_{i=1}^{m} \ln(g(y_i \cdot \mathbf{w}^T \mathbf{x}_i))$$

The loss is then reduced by gradient descent method.

$$\frac{\partial L_D(\mathbf{w})}{\partial \mathbf{w}} = -\frac{1}{m}\sum_{i=1}^{m}(1 - g(y_i \cdot \mathbf{w}^T \mathbf{x}_i))(y_i \cdot \mathbf{x}_i)$$

(2)Linear Classification

In the field of machine learning, the goal of statistical classification is to use an object's characteristics to identify which class (or group) it belongs to. A linear classifier achieves this by making a classification decision based on the value of a linear combination of the characteristics. An object's characteristics are also known as feature values and are typically presented to the machine in a vector called a feature vector. Such classifiers work well for practical problems such as document classification, and more generally for problems with many variables (features), reaching accuracy levels comparable to non-linear classifiers while taking less time to train and use.

This experiment use SVM model to achieve Linear Classification. And the Loss Function of SVM is:

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

And the gradient of loss function respecting to w is:

$$\frac{\partial L_D(\mathbf{w}, b)}{\partial \mathbf{w}} = \begin{cases} \mathbf{w}^T - C\mathbf{X}^T\mathbf{y}, & \text{if } 1 - Y_i(\mathbf{w}^T x_i + b) >= 0 \\ \mathbf{w}^T, & \text{if } 1 - Y_i(\mathbf{w}^T x_i + b) < 0 \end{cases}$$

(3)Various optimization methods of Stochastic Gradient Descent

a. NAG

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

b. RMSProp

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

c. AdaDelta

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

d. Adam

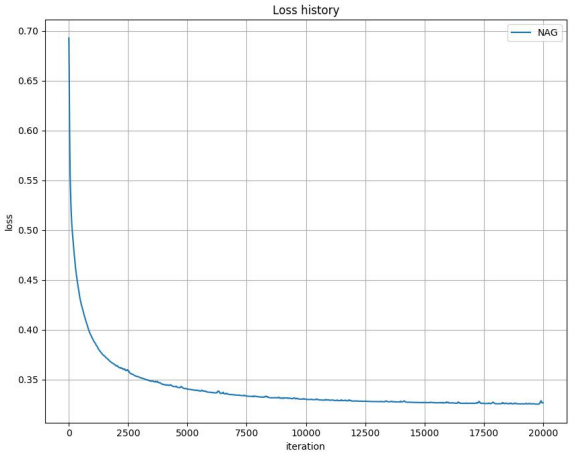
$$\mathbf{g}_t \leftarrow \nabla J(\boldsymbol{\theta}_{t-1})$$
$$\mathbf{m}_t \leftarrow \beta_1\mathbf{m}_{t-1} + (1-\beta_1)\mathbf{g}_t$$
$$G_t \leftarrow \gamma G_t + (1-\gamma)\mathbf{g}_t \odot \mathbf{g}_t$$
$$\alpha \leftarrow \eta\frac{\sqrt{1-\gamma^t}}{1-\beta^t}$$
$$\boldsymbol{\theta}_t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

III. Experiment

(1)Logistic Regression and Stochastic Gradient Descent

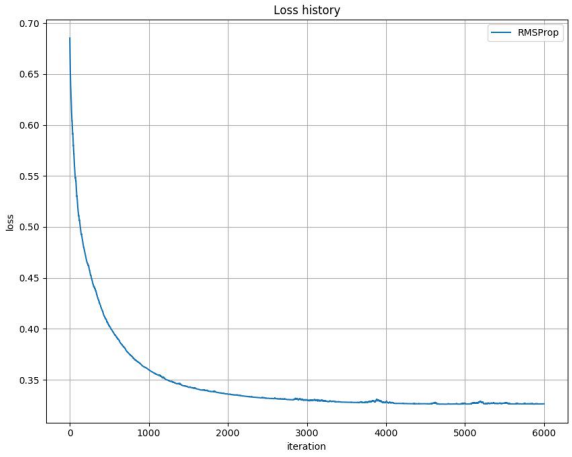
a. NAG

The number of iterations: 20000
The best accuracy is 0.852282
Loss Function figure:



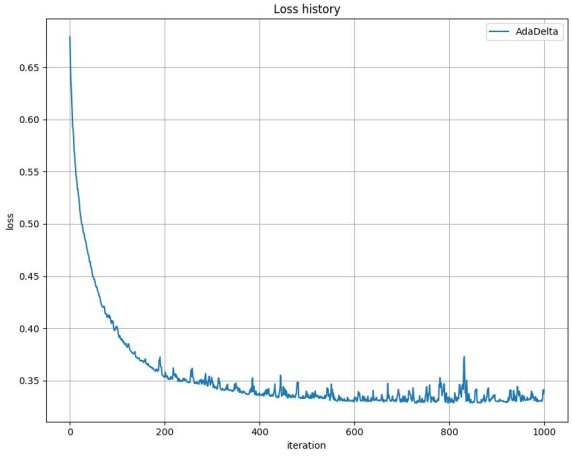
b. RMSProp

The number of iterations: 5000
The best accuracy is 0.850439
Loss Function figure:



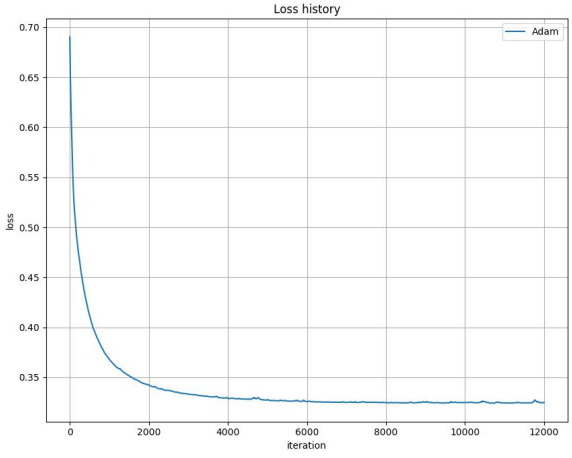
c. AdaDelta

The number of iterations: 1000
The best accuracy is 0.850071
Loss Function figure:

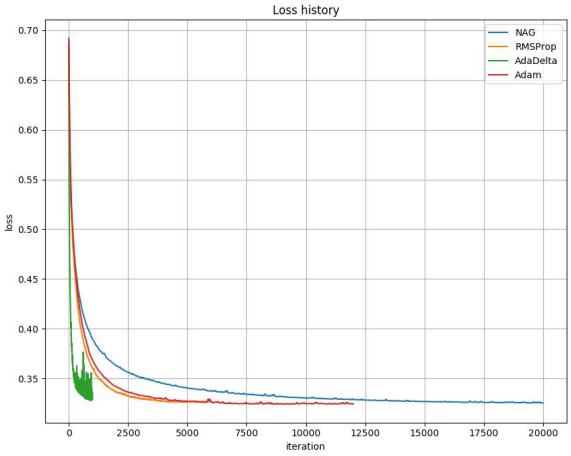


d. Adam

The number of iterations: 10000
The best accuracy is 0.851606
Loss Function figure:



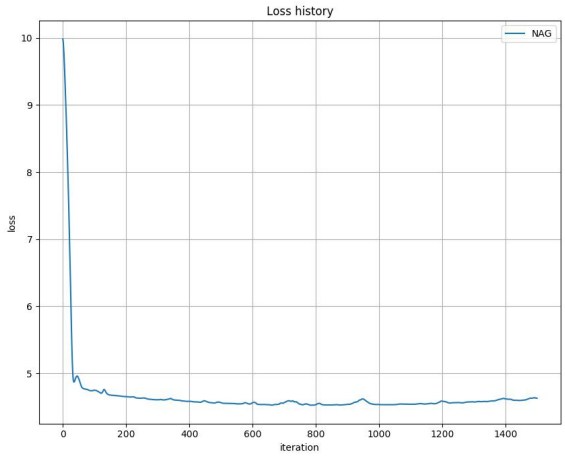
e.Composite figure



(2)Linear Classification and Stochastic Gradient Descent

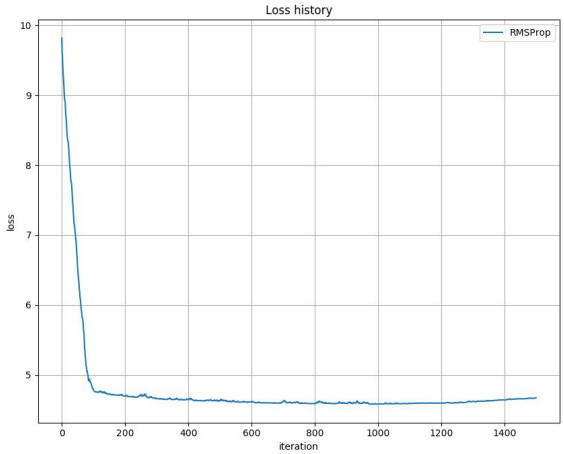
a. NAG

The number of iterations: 1500
The best accuracy is 0.838339
Loss Function figure:



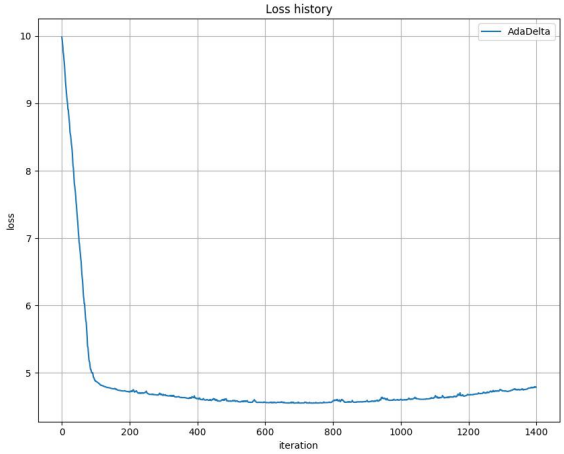
b. RMSProp

The number of iterations: 1500
The best accuracy is 0.829187
Loss Function figure:



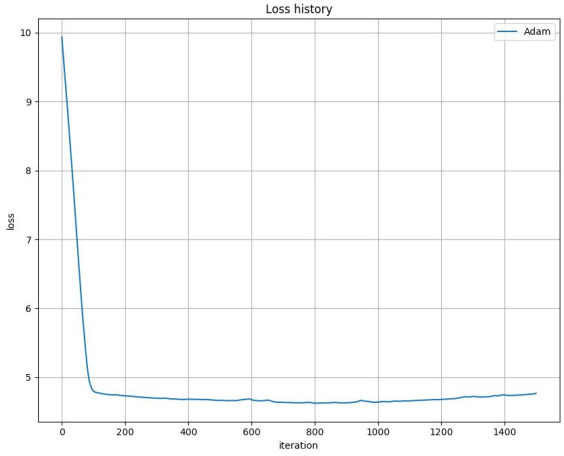
c. AdaDelta

The number of iterations: 1500
The best accuracy is 0.839322
Loss Function figure:

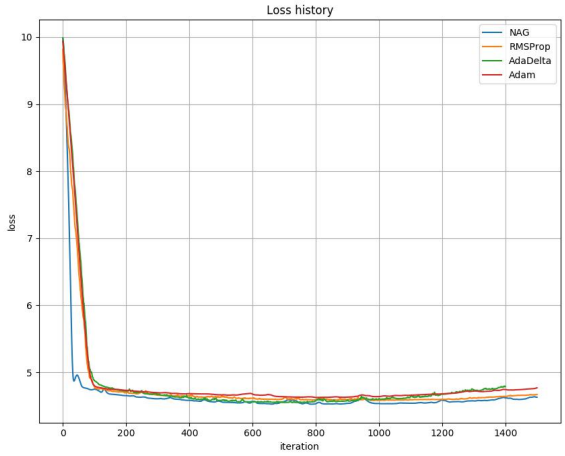


d. Adam

The number of iterations: 1500
The best accuracy is 0.826116
Loss Function figure:



e.Composite figure



IV. CONCLUSION

In this experiment, not only can we get a deeper understanding of the realization of logistic regression and linear classification by using various optimization methods of stochastic gradient descent, but also realize the realization of various optimization methods of stochastic gradient descent and the application of machine learning in practice In the fitting effect.

From the descent graph of the loss function, except that AdaDelta fluctuated greatly during the later iteration, the other three optimization methods showed very stable convergence, and RMSProp and Adam converged faster. And from the point of view of accuracy, Adam looks better.