South China University of Technology

The Experiment Report of Machine Learning

SCHOOL: SCHOOL OF SOFTWARE ENGINEERING

SUBJECT: SOFTWARE ENGINEERING

Author:
Zi longLi

Supervisor:
Qingyao Wu

Student ID：
201721045398

Grade:
Graduate

December 14, 2017

Logistic Regression, Linear Classification and Stochastic Gradient Descent

Abstract

We implemented a Logistic regression and a Linear classification algorithm ,which get very low loss in the LIBSVM a9a data set. More than that ,We used four type of optimization methods, which can help us update the parameters.

I. INTRODUCTION

In statistics, logistic regression, or logit regression, or logit In statistics, logistic regression, or logit regression, or logit model is a regression model where the dependent variable (DV) is categorical. This article 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. Cases where the dependent variable has more than two outcome categories may be analysed in multinomial logistic regression, or, if the multiple categories are ordered, in ordinal logistic regression. In the terminology of economics, logistic regression is an example of a qualitative response/discrete choice model.

Logistic regression was developed by statistician David Cox in 1958. The binary logistic model is used to estimate the probability of a binary response based on one or more predictor (or independent) variables (features). It allows one to say that the presence of a risk factor increases the odds of a given outcome by a specific factor.

II. METHODS AND THEORY

2.1 logistic function

The sigmoid function aka logistic function, is often used as the thresholding function, which is continuous and differentiable.

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

Given a data set $D\,\{y_i = \pm 1, \boldsymbol{x_i}\}_{i=1}^{n}$of n n statistical units, we assume that the model's parameter as $\boldsymbol{W}$.So, it can also be written as :

$$g(\boldsymbol{W}^T x) = \frac{1}{1+e^{-\boldsymbol{W^T X}}}$$

So,the probability function can be written as :

$$P(y|x) = \begin{cases} g(\boldsymbol{W}^T x), y = 1 \\ 1 - g(\boldsymbol{W}^T x), y = -1 \end{cases}$$

That is same to:

$$P(y|x) = g(y * \boldsymbol{W}^T x)$$

We used the Likelihood function to evaluate the parameters $\boldsymbol{W}^T$ in data set D:

$$P(y_1, \dots, y_n | x_1, \dots, x_n) = \prod_{i=1}^{n} P(y_i|x_i)$$

$$max \prod_{i=1}^{n} P(y_i|x_i) \Leftrightarrow \max \log_e \prod_{i=1}^{n} P(y_i|x_i)$$

Then, $J(\boldsymbol{w}) = \frac{1}{n} \log_e \prod_{i=1}^{n} P(y_i|x_i)$

$$J(\boldsymbol{w}) = \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + e^{-y_i * W^T * x_i}\right) + \frac{\mu}{2} \|W\|^2$$

The $\frac{\mu}{2} \|W\|^2$ is a penalty for model parameters to prevent overfitting.

The Gradient of J(W) is as follow:

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

So, we can update the parameter as follow:

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

2.2 svm (support vector machine):

Not like the single perceptron machine, support vector machine thought that the hyperplane is based on the support vector. It use hinge loss to calculate the gradient:

The loss function is:

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

And the update equation of parameter is:

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

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

2.3 NAG

As it's well known to us all, Momentum strategy is on the basis of SGD strategy. It consider that the update of parameter is not only depends on the gradient of this moment, but also depends on the gradient of last moment

It's update equation is as follows:

$d_i = \beta d_{i-1} + g(\theta_{i-1})$

$\theta_i = \theta_{i-1} - \alpha d_i$

NAG(Nesterov Accelerated Gradient) make some progress on the momentum algorithm.

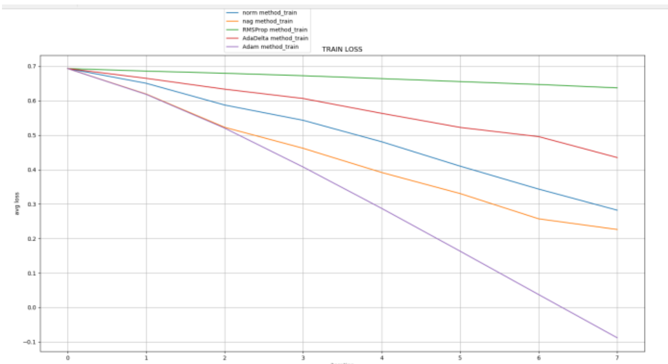
$$d_i = \beta d_{i-1} + g(\theta_{i-1} - \alpha\beta d_{i-1})$$

$$\theta_i = \theta_{i-1} - \alpha d_i$$

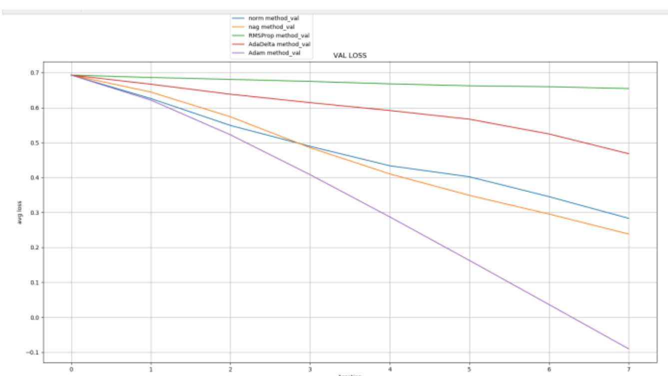
It looks a little longer than momentum strategy. So, this strategy can decides the update step ,and convergence faster.

III. EXPERIMENT

We conduct the regression experiment in the LIBSVM set and separate the traning data set and the vilification as 32561 :16281.In Regression experiment, we set the iteration size to 8,Batch size=16,learning rate=0.01 and we set the initial value of θ_i as zero.

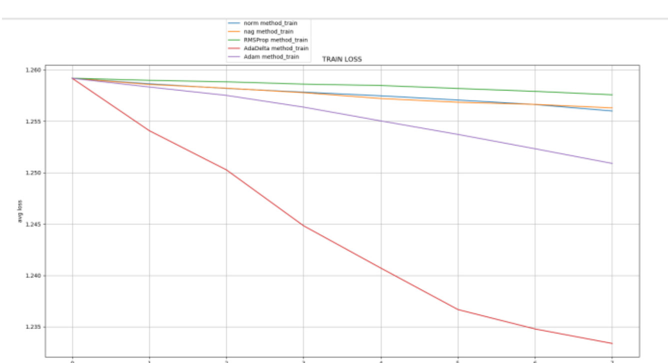


Regression Train Loss vs iteration

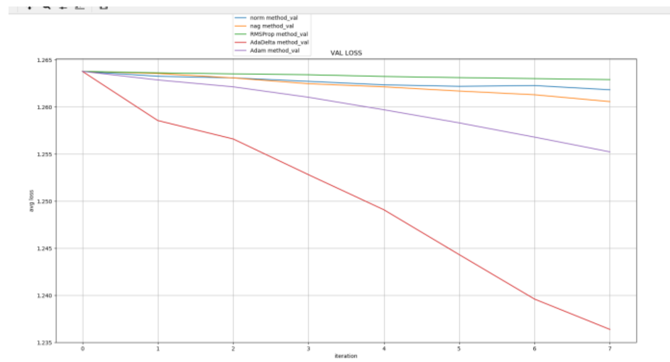


Regression Validation Loss vs iteration

We conduct the Classification experiment in the LIBSVM set and separate the traning data set and the vilification as 32561 :16281.In Regression experiment, we set the iteration size to 8,Batch size=16,learning rate=0.01 and we set the initial value of θ_i as zero.



Classification Train Loss vs iteration



Classification Validation Loss vs iteration

IV. CONCLUSION

We performed logistic regression and linear classification algorithms and got good results. Linear model operation is very simple and convenient, has a wide range of applications in the CS and financial fields.