South China University of Technology

The Experiment Report of Machine Learning

SCHOOL: SCHOOL OF SOFTWARE ENGINEERING

SUBJECT: SOFTWARE ENGINEERING

Author:
AoXiang Zhang

Supervisor:
Mingkui Tan

Student ID：
201530613580

Grade:
Undergraduate

December 9, 2017

Experimental Study on Stochastic Gradient Descent for Solving Classification Problems

Abstract—Logistic regression is a basic method of machine learning. In the previous experiment, a linear classification experiment was carried out. In this experiment, I will compare linear classification and logistic regression classification.

I. INTRODUCTION

Now there are many different algorithms for classifying logical feature datasets and there are many ways to update model parameters. This experiment aims to achieve both logistic regression and linear classification, using four of the model parameter update methods , Learn the core ideas, and compare different methods.

II. METHODS AND THEORY

In this experiment, we will use logistic regression and linear classification respectively to predict the data, and we will use four different methods to update the model parameters.

1. Linear classification

For linear classification, we use hinge loss to calculate the loss, and derive it to calculate the gradient.

$$\text{Hinge loss} = \varepsilon_i = \max(0, 1 - y_i(w^T x_i + b))$$

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

2. Logistic regression

For logistic regression, we will use the maximum likelihood estimation method to calculate the loss, and derive it to calculate the gradient.

$$\mathrm{J(w)} = \frac{1}{n}\sum_{i=1}^{n} \log(1 + e^{-y_i W^T X_i}) + \frac{\lambda}{2}||w||_2^2$$

$$\frac{\partial J(w)}{\partial w} = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i X_i}{1 + e^{y_i W^T X_i}}$$

3. NAG

Nesterov is a variant of Momentum, the only difference between Nesterov and Momentum is that to calculate the gradient difference, Nesterov first updates the parameter with the current velocity v, and calculates the gradient with the updated temporary parameter. Under GD, Nesterov converges the error from $O(1/k)$ to $O\left(1/k^2\right)$, but there is no improvement

4. RMSProp

RMSProp is improved from AdaGrad improvement. Since both neural networks are non-convex, RMSProp performs better under nonconvex conditions and alters the moving average of gradient accumulation to exponential decay to discard distant past history. Compared to the historical gradient of AdaGrad $r \leftarrow r + g \odot g$. RMSProp adds an attenuation factor to control how much history information gets $r \leftarrow \rho r + (1 - \rho) g \odot g$.

5. AdaDelta

The first-order method is used to estimate the Hessian matrix, but it is derived from the Ada-Grad method. Eliminating the need to manually set the learning rate of the process by this method. As the learning rate in the network training process should be gradually reduced, this is the learning rate of annealing.

6. Adam

Adam algorithm can be regarded as a modified Momentum and RMSProp.

III. EXPERIMENT

Datasets

In this experiment, we download the data from LIBSVM Data, the dataset has been devided two datasets as train set and test set. These datasets are binary classification with 123 features. The train set contains 32561 groups of data and test set contains 16281 groups of data

Experimental Setup

In this experiment, we use linear classification and logistic regression to classify the dataset, use SGD to update model parameters and we compare differences between NAG, RMSProp, AdaDelta and Adam, four kinds of methods to optimize SGD.

Experiment with logistic regression

In this experiment, we used logistic regression to classify the dataset, and used four kinds of optimization methods to update the model parameters.

Figure 1show the loss of logistic regression with four kinds of optimization methods. We tried different threshold to improve accuracy.

Experiment with linear classification

In this experiment, we used linear classification to classify the dataset, and used four kinds of optimization methods to update the model parameters.

Figure 2show the loss of linear classification with four kinds of optimization methods. We tried different threshold to improve accuracy.

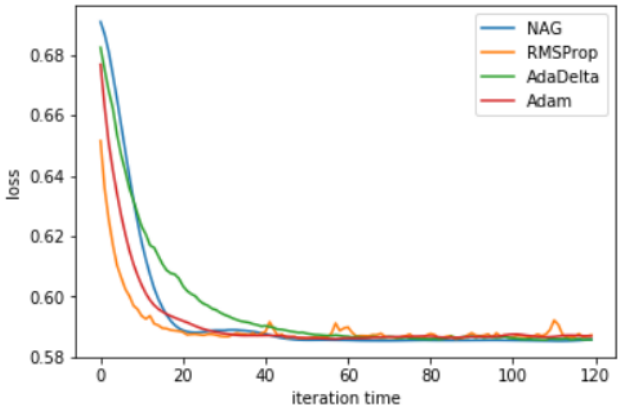

Figure 1

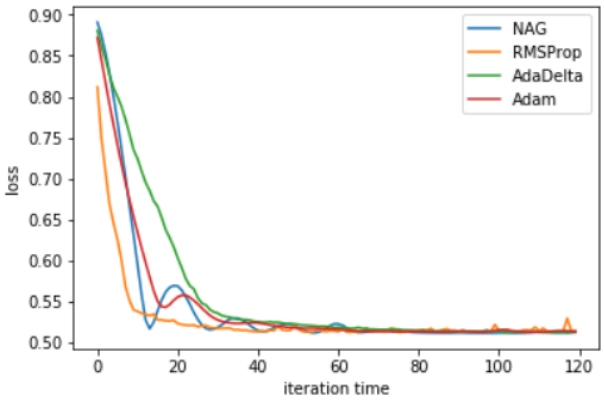

Figure 2

IV. Conclusion

Through this experiment, we classify a9a dataset by experimental logistic regression and linear classification, and compare the four optimization methods to update the model parameters.

Comparing the two algorithms, we find that the initial loss of logistic regression is smaller, the loss of final stability is smaller, the loss of change is smaller, and the linear classification is the opposite when we use exactly the same parameters.

For the four optimized model parameter update methods, RMSProp has the smallest initial loss and stabilizes as quickly as possible, but fluctuates when stabilized. Adam and NAG loss rate of decline in the middle, there are some fluctuations in the decline. AdaDelta lost the slowest loss. The four optimization methods eventually get the same loss, classification accuracy is also similar.

References

BVL10101111 Deep Learning 之 最优化方法 http://blog.csdn.net/BVL10101111/article/details/72614711

Csuhoward Caffe Solver 理解篇（2） SGD, AdaDelta, Ada-Grad, Adam, NAG, RMSprop 六种梯度下降方法横向对比 http://blog.csdn.net/csuhoward/article/details/53255918