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:
Shoukai Xu and Yaofu Chen

Supervisor:
Mingkui Tan or Qingyao Wu

Student ID：
201530611111 and 20153060000

Grade:
Undergraduate or Graduate

December 9, 2017

Logistic Regression, Linear Classification and Stochastic Gradient Descent

Abstract—In this experiment, we are going to realize regression and linear classification using stochastic gradient descent, and use different methods to optimize them.

I. Introduction

Sometimes given a dataset which has elements with different features, you need to classify them into two or more different group. In this experiment we focus on the binary classification problem, and try to solve them by two methods.

II. Methods and Theory

In this experiment we are going to use two methods: logistic regression and linear classification.

Logistic regression is based on linear regression. Suppose that we have already completed linear regression, how do we use it to classify? The predicted value of the testing data is a real number, but its tag is binary: either 0 or 1. So we need a function to map the real value to 0/1. Then we use gradient descent to decide the parameters of the function, and then use it to classify. The loss function is:

$$J(\mathbf{w}) = -\frac{1}{n}\left[\sum_{i=1}^{n} y_i \log h_{\mathbf{w}}(\mathbf{x}_i) + (1-y_i)\log\left(1-h_{\mathbf{w}}(\mathbf{x}_i)\right)\right]$$

Where

$$h_{\mathbf{w}}(\mathbf{x}) = g(\mathbf{w}^\top \mathbf{x}) = \frac{1}{1+e^{-\mathbf{w}^\top \mathbf{x}}}$$

And we can calculate its gradient:

$$\frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} = \left(h_{\mathbf{w}}(\mathbf{x}) - y\right)\mathbf{x}$$

Now we can use gradient descent to calculate the parameters.

Linear classification is to use a hyperplane to divide the whole space into two different parts, and one is class 0, the other is class 1. To choose the hyperplane, we hope that all data in class 0 is on one side, and all data in class 1 is in the other one. What's more, we hope the smallest distance between any element and the hyperplane to be the greatest. So, we use the algorithm svm to decide the hyperplane, and use it to classify data. We use hinge loss to measure the loss. The loss function is:

$$Hinge\ loss = \xi_i = \max(0, 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b))$$

The optimization problem becomes:

$$\min_{\mathbf{w},b} \quad f: \frac{\|\mathbf{w}\|^2}{2} + C\sum_{i=1}^{N} \max(0, 1 - y_i(\mathbf{w}^\top x_i + b))$$

We can use gradient descent to calculate the parameters of all the functions we described above, but as it calculates the gradient of all samples in each iteration, it is too time consuming. So, in stochastic descent, we randomly choose a sample, calculate its gradient and update parameters. To avoid instability, we can choose more than one example to make estimate of the gradient, which is called mini-batch. We can also use some methods to optimize the gradient descent process.

III. Experiment

A. Dataset

Experiment uses a9a of LIBSVM Data, including 32561/16281(testing) samples and each sample has 123/123 (testing) features. Please download the training set and validation set.

B. Implementation

The parameters chosen are optimized by multiple times of experiments. Parameters of the two methods are shown in TABLE 1 and 2.

TABLE 1
Parameters of logistic regression:

NAG	Gama = 0.9 Nita = 0.01
RMSProp	Gama = 0.9 Nita = 0.01
AdaDelta	Gama = 0.99
Adam	Gama = 0.999 Nita = 0.01 Beta = 0.9

TABLE 2
Parameters of linear classification:

NAG	Gama = 0.9 Nita = 0.001
RMSProp	Gama = 0.9 Nita = 0.001

AdaDelta	Gama = 0.9
Adam	Gama = 0.999 Nita = 0.001 Beta = 0.9

From the experiment of logistic regression, we get the result of accuracy and loss, which are separately shown in Fig 1 and 2. From the figures we can see that in all of the four optimization method, the accuracy converge to around 0.85, and the loss converge to about 0.32. RMSProp and Adam are the fastest method to converge, while AdaDelta is slower than other methods.

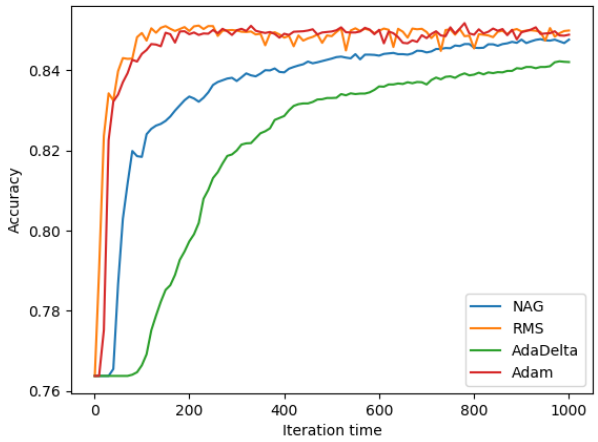

Fig 1. Accuracy of logistic regression

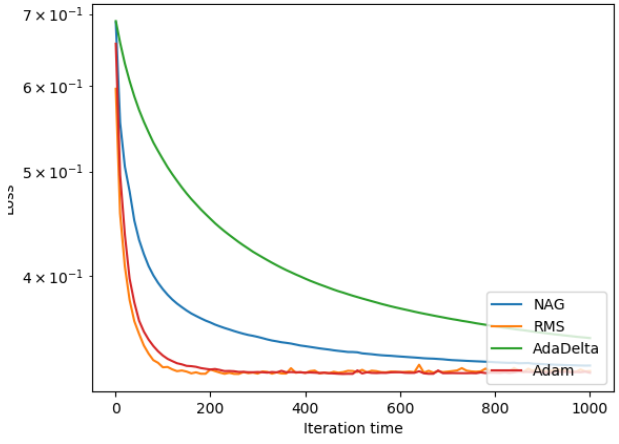

Fig 2. Loss of logistic regression

The accuracy and loss of linear classification are shown in Fig 3 and 4. From the figures we can see that in all of the methods accuracy converge to about 0.75, and loss converge to about 0.47. In this experiment NAG is the fastest one, while AdaDelta is still the slowest.

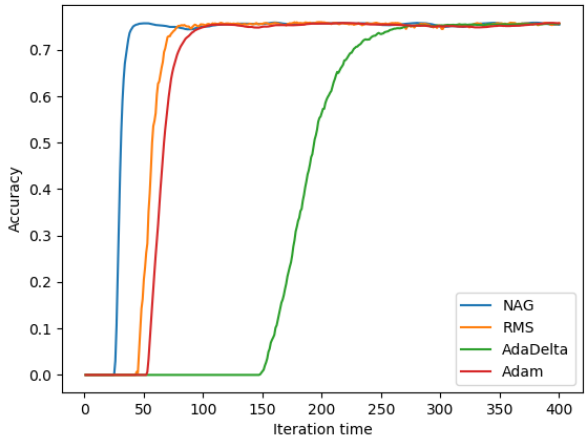

Fig 3. Accuracy of linear classification

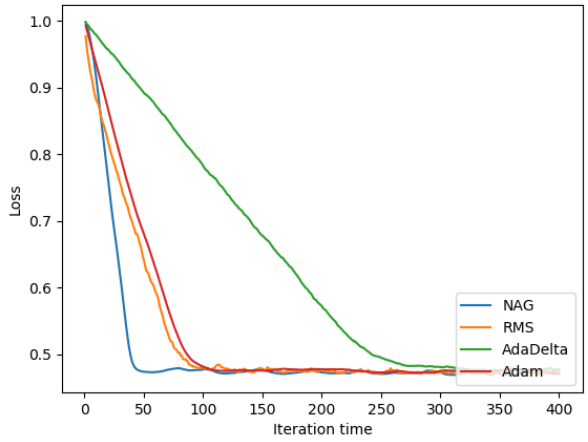

Fig 4. Loss of linear classification

IV. CONCLUSION

In this experiment, we realized logistic regression and linear classification using stochastic gradient decent and tried four different methods to optimize it. We compared the performance of the four different methods: the accuracy and the loss and how many iterations they need to converge. We found that both logistic regression and linear regression can efficiently make binary classification, as their accuracy are rather high. For the four optimizing methods, AdaDelta is always slower than the other three, but the accuracy of the four methods are similar.