The Experiment Report of Machine Learning



December 14, 2017

Grade:

Undergraduate

Student ID：

201530611937

Supervisor:

Mingkui Tan

Author:

Canguang Li

**SUBJECT:**SOFTWARE ENGINEERING

**SCHOOL:** SCHOOL OF SOFTWARE ENGINEERING

[[1]](#footnote-1)Logistic Regression, Linear Classification and Stochastic Gradient Descent

Abstract—In this experiment, logistic regression and linear classification algorithms were implemented using mini-batch, stochastic gradient descent and the model parameters were updated with four different optimization methods-NAG, RMSProp, AdaDelta and Adam.

# INTRODUCTION

Through implementing logistic regression and linear classification with stochastic gradient descent, this experiment explored the difference between logistic regression and linear classification as well as the difference between gradient descent and stochastic gradient descent. What’s more, this experiment explored the effect of four difference optimization methods-NAG, RMSProp, AdaDelta and Adam.

# METHODS AND THEORY

**A. Logistic regression**

The logistic regression model function used in the experiment is

The loss function used is

Its gradient is

**B. Linear classification**

The linear classification model function used in the

experiment is

The loss function used is

Denote

then the gradient is

**C. Optimization methods**

Four different optimization methods used in this experiment is NAG, RMSProp, AdeDelta and Adam.

The update step of NAG is

The update step of RMSProp is

The update step of AdaDelta is

The update step of Adam is

# Experiment

**A. Dataset**

Experiment uses a9a of LIBSVM Data, including

32561/16281(testing) samples and each sample has

123/123(testing) features.

**B. Implementation**

The parameters(of logistic regression and linear

classification) were initialized as follow:

TABLE1

Initialized parameters

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | NAG | RMSProp | AdaDelta | Adam |
|  | - | 1e-8 | 1e-8 | 1e-8 |
|  | 0 | - | - | - |
|  | - | 0 | 0 | 0 |
|  | - | - | - | 0 |
|  | - | - | 0 | - |
|  | 0.1 | 0.1 | - | 0.1 |
|  | 0.9 | 0.9 | 0.9 | 0.9 |
|  | - | - | - | 0.9 |
|  | 1 | 1 | 1 | 1 |
|  | 0 | 0 | 0 | 0 |

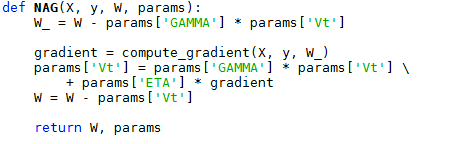
After exploring, some super parameters(of logistic regression and linear classification) were as follow:

TABLE2

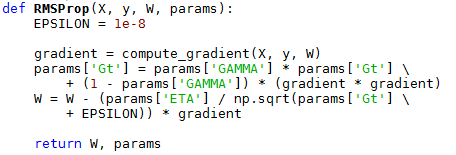
Adjusted super parameters

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | NAG | RMSProp | AdaDelta | Adam |
|  | 0.01 | 0.01 | - | 0.01 |
|  | 0.9 | 0.9 | 0.9999 | 0.9 |
|  | - | - | - | 0.9 |

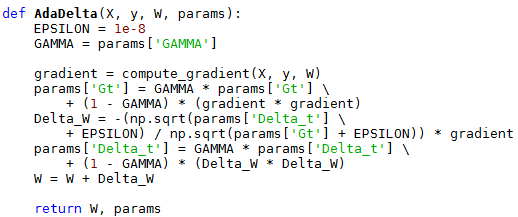
The implementation of four optimization method were as follow:



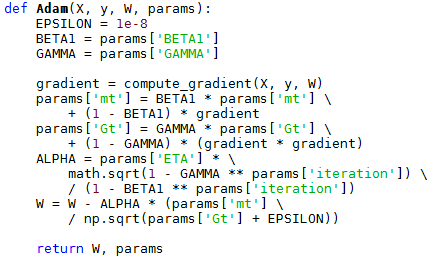
**Figure 1 Implementation of NAG**

****

**Figure 2 Implementation of RMSProp**

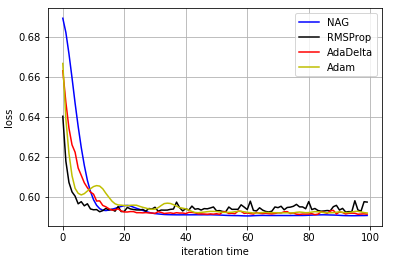
****

**Figure 3 Implementation of AdaDelta**

****

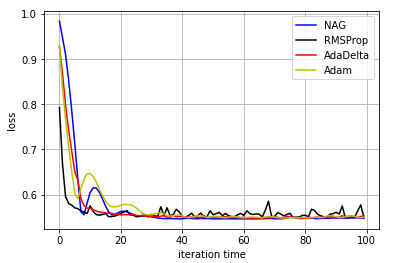
**Figure 4 Implementation of Adam**

As for logistic regression, I assigned 100 to batch number, 0.5 to threshold and 100 to iteration time.And the loss curve was as figure 5:



**Figure 5 Loss curve of four optimization method**

As for linear regression, I assigned 100 to batch number, 0 to threshold and 100 to iteration time.And the loss curve was as figure 4:



**Figure 6 Loss curve of four optimization method**

# conclusion

Through implementing logistic regression and linear classification with stochastic gradient descent, I learned that logistic regression is not a model to do regression but a model to classify. Logistic regression and linear classification all can solve two-classification problem. Besides, I learned that the difference between gradient descent and stochastic gradient descent. Gradient descent uses all the samples to compute the gradient each iteration but stochastic gradient descent uses just one samples to compute the gradient each iteration. Gradient descent performs well but slowly, while stochastic gradient descent performs not very well but fast. To balance, this experiment used some samples to compute gradient each iteration(mini-batch stochastic gradient descent), this method performs well and fast. As for four optimization methods, I learned that AdaDelta converge slower than other three methods under the same conditions. And RMSProp as well as Adam often oscillates slightly when they trend to converge. But all the optimization methods used in this experiment all can make the loss curve converge to a similar value.

1. [↑](#footnote-ref-1)