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:
Liangyu Xiao

Supervisor:
Qingyao Wu

Student ID：
201530613146

Grade:
Undergraduate

December 9, 2017

Logistic Regression, Linear Classification and Stochastic Gradient Descent

Abstract—

This article is written for an experiment of Logistic Regression, Linear Classification and Stochastic Gradient Descent, which include the realization of linear regression and linear classification and stochastic gradient descent with NAG, AdaDelta, RMSProp, Adam.

I. INTRODUCTION

In this experiment, we are going to compare and understand the difference between gradient descent and stochastic gradient descent. We will also compare and understand the differences and relationships between Logistic regression and linear classification. And get a further understand the principles of SVM and practice on larger data.

II. METHODS AND THEORY

A. *Logistic Regression:*

Logistic Model:

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

Loss Function:

$$\boldsymbol{J(w)} = -\frac{1}{n}\left[\sum_{i=1}^{n} y_i \log h_w(x_i) + (1 - y_i)\log(1 - h_w(x_i))\right]$$

Gradient:

$$\frac{\boldsymbol{\partial J(w)}}{\boldsymbol{\partial w}} = (\boldsymbol{h_w(X) - y})\boldsymbol{X}$$

B. *Linear Classification*

Linear Model:

$$\boldsymbol{y = w^T X}$$

Loss Function:

$$\boldsymbol{L(w)} = \frac{\|\boldsymbol{w}\|^2}{\boldsymbol{2}} + \boldsymbol{C}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{g(w)}$$

$$\boldsymbol{g(w)} = \boldsymbol{\max(0, 1 - y_i w^T x_i)}$$

Gradient:

$$\boldsymbol{g_w(x)} = \begin{cases} -\boldsymbol{y_i x_i} & 1 - \boldsymbol{y_i w^T x_i} >= 0 \\ \boldsymbol{0} & 1 - \boldsymbol{y_i w^T x_i} < 0 \end{cases}$$

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

C. *Stochastic Gradient Descent*

In this experiment, we use SGD to compute the gradient to reduce the loss, and use NAG, AdaDelta, RMSProp, Adam to improve SGD.

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*

B.1 Logistic Regression:

Initialization:

Batch size:128

Feature size: 123

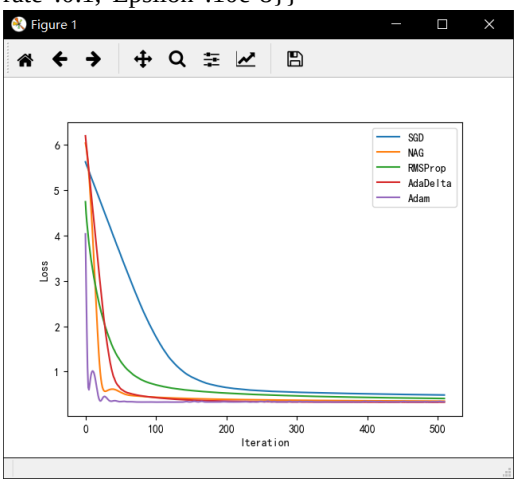
w: random

Process:

1. Load the training set and validation set. Shuffle the data.
2. Initalize logistic regression model.
3. Select the loss function and calculate its derivation.
4. compute the gradient by the selected method.
5. Update model parameters using different optimized methods(NAG，RMSProp，AdaDelta and Adam).
6. Predict under validation set and get the different optimized method
7. Repeate step 4 to 6 for several times, and drawing graph of NAG，RMSProp，AdaDelta and Adam

Result:

"SGD":{"learning rate":0.01},
"NAG":{"learning rate":0.01,"Gamma":0.9},
"RMSProp":{"learning rate":0.01,"Gamma":0.9,"Epsilon":10e-8},
"AdaDelta":{"Gamma":0.95,"Epsilon":10e-6},
"Adam":{"Beta":0.9,"Gamma":0.999,"learning rate":0.1,"Epsilon":10e-8}}



B. 2 Linear Classification

Initialization:

Batch size:256

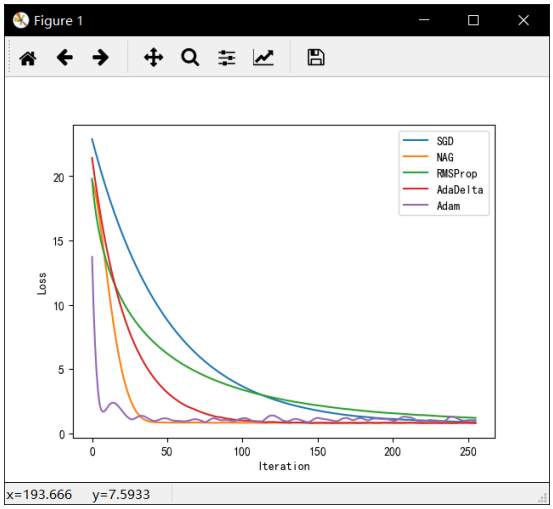
Feature size: 123

w: random

Process:

1. Load the training set and validation set. Shuffle the data.
2. Initalize SVM model parameters
3. Select the loss function and calculate its derivation.
4. compute the gradient by the selected method.
5. Update model parameters using different optimized methods(NAG，RMSProp，AdaDelta and Adam).
6. Predict under validation set and get the different optimized method
7. Repeate step 4 to 6 for several times, and drawing graph of NAG，RMSProp，AdaDelta and Adam

Result:



IV. CONCLUSION

From those experiment data and graph above, we can find that SGD is a bit slow than others, in which proves that when facing a great number of data, the accuracy of SGD may decrease. Otherwise, the other four way to SGD is significant efficient to increase the speed of loss decreasing.