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:
Jiachen Wang

Supervisor:
Qingyao Wu

Student ID：
201530612859

Grade:
Undergraduate

December 2, 2017

# Linear Regression, Linear Classification and Gradient Descent

**Abstract—**

## I. INTRODUCTION

1、Compare and understand the difference between gradient descent and stochastic gradient descent.
2、Compare and understand the differences and relationships between Logistic regression and linear classification.
3、Further understand the principles of SVM and practice on larger data.

## II. METHODS AND THEORY

In this experiment，I use SGD as method to implement Logistic Regression and SVM models。Then, I use four different optimization methods to optimize SGD，and compare their differences. In this experiment ， the loss function of Logistics Regression is $\sum_{i=1}^{m}(-y_i\beta^T x_i + \ln(1 + e^{\beta^T x_i}))$, and it's derivative is $-\sum_{i=1}^{m} x_i(y_i - \frac{e^{\beta^T}}{1+e^{\beta^T}})$. The loss function of SVM is $\frac{1}{2}||\omega||^2 + C\sum_i^n \max(0,1 - y_i(\omega^T x_i))$, it's derivative is $\omega - C\sum_i^n I[(1 - y_i(\omega^T x_i)) > 0]\, y_i x_i$

The four optimization methods mentioned above are NAG, RMSProp, AdaDelta, Adam.

The NAG's Mathematical formula is

$$\mathbf{g}_t \leftarrow \nabla J(\boldsymbol{\theta}_{t-1} - \gamma \mathbf{v}_{t-1})$$
$$\mathbf{v}_t \leftarrow \gamma \mathbf{v}_{t-1} + \eta \mathbf{g}_t$$
$$\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \mathbf{v}_t$$

The RMSProp's Mathematical formula is

$$\mathbf{g}_t \leftarrow \nabla J(\boldsymbol{\theta}_{t-1})$$
$$G_t \leftarrow \gamma G_t + (1-\gamma)\mathbf{g}_t \odot \mathbf{g}_t$$
$$\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \frac{\eta}{\sqrt{G_t + \epsilon}} \odot \mathbf{g}_t$$

The AdaDelta's Mathematical formula is

$$\mathbf{g}_t \leftarrow \nabla J(\boldsymbol{\theta}_{t-1})$$
$$G_t \leftarrow \gamma G_t + (1-\gamma)\mathbf{g}_t \odot \mathbf{g}_t$$
$$\Delta\boldsymbol{\theta}_t \leftarrow -\frac{\sqrt{\Delta_{t-1} + \epsilon}}{\sqrt{G_t + \epsilon}} \odot \mathbf{g}_t$$
$$\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} + \Delta\boldsymbol{\theta}_t$$
$$\Delta_t \leftarrow \gamma\Delta_{t-1} + (1-\gamma)\Delta\boldsymbol{\theta}_t \odot \Delta\boldsymbol{\theta}_t$$

The Adam Mathematical formula is

$$\mathbf{g}_t \leftarrow \nabla J(\boldsymbol{\theta}_{t-1})$$
$$\mathbf{m}_t \leftarrow \beta_1 \mathbf{m}_{t-1} + (1-\beta_1)\mathbf{g}_t$$
$$G_t \leftarrow \gamma G_t + (1-\gamma)\mathbf{g}_t \odot \mathbf{g}_t$$
$$\alpha \leftarrow \eta\frac{\sqrt{1-\gamma^t}}{1-\beta^t}$$
$$\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \alpha\frac{\mathbf{m}_t}{\sqrt{G_t + \epsilon}}$$

## III. EXPERIMENT

### A. Dataset

Experiment uses a9a of LIBSVM Data, including 32561/16281(testing) samples and each sample has 123/123 (testing) features. Data sets are represented using sparse matrices. It needs to be converted to a dense matrix, in which the 123th special test of the test set all 0, when the conversion will save skip this feature, I need to add manually.

### B. Implement

Logistic Regression and Linear Classification and Stochastic Gradient Descent

1. Load the training set and validation set.
2. Initalize logistic regression model parameters,
3. Select the loss function and calculate its derivation.
4. Calculate gradient toward loss function from partial samples.
5.Update model parameters using different optimized methods(NAG，RMSProp，AdaDelta and Adam).
6. Select the appropriate threshold, mark the sample whose predict scores greater than the threshold as positive, on the contrary as negative. Predict under

validation set and get the different optimized method loss Loss-NAG, Loss-RMSProp, Loss-AdaDelta and Loss-Adam.

7. Repeate step 4 to 6 for several times, and drawing graph of Loss-NAG, Loss-RMSProp, Loss-AdaDelta and Loss-Adam. with the number of iterations.

Table of parameters(Logistic Regression):

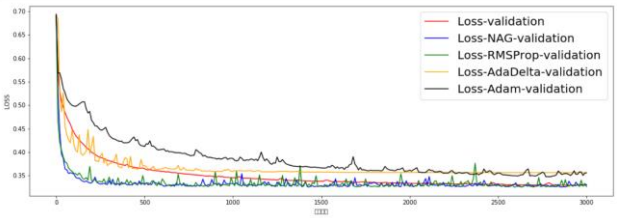
| Optimization | |
|---|---|
| NAG | learning rate=0.002,epoch=3000, $\gamma$ =0.9 |
| RMSProp | learning rate=0.1,epoch=3000, $\gamma$ =0.9 |
| AdaDelta | initial delta=0.003， epoch=3000, $\gamma$ =0.95 |
| Adam | learning rate=0.2,epoch=3000, $\gamma$ =0.9， $\beta$ =0.999 |

Table of parameters(SVM):

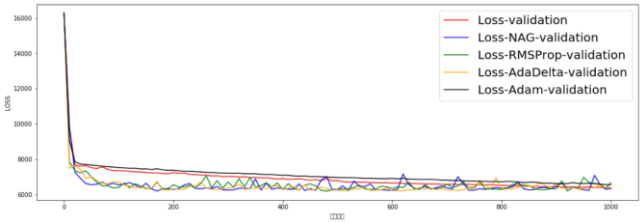
| Optimization | |
|---|---|
| NAG | learning rate=0.001,epoch=1000, $\gamma$ =0.9 |
| RMSProp | learning rate=0.1,epoch=1000, $\gamma$ =0.9 |
| AdaDelta | initial delta=0.003， epoch=1000, $\gamma$ =0.95 |
| Adam | learning rate=0.1,epoch=3000, $\gamma$ =0.9， $\beta$ =0.999 |

## IV. CONCLUSION

Logistic Regression LOSS with the number of iterations.




SVM LOSS with the number of iterations.




Conclusion:

Through the four different optimization methods and the stochastic gradient descent method, we can see that the method of stochastic gradient descent has a great fluctuation in LOSS of the validation set with the number of iterations. Among the four optimization methods, NAG and The optimization of RMSProp is more volatile than the other two methods. Adam's convergence is the smoothest among the four optimization methods, but Adam did not achieve the optimum in the logistic regression. By adjusting the parameters it is difficult to achieve the best, I did not find the reason. However, it can be seen from the linear classification that the Adam algorithm curve is the smoothest and the convergence speed is fast at the same time.

In this experiment I implemented logistic regression and linear classification using stochastic gradient descent, and used four different optimization methods to optimize the stochastic gradient descent algorithm. Through this experiment I deepened my understanding of stochastic gradient descent, logistic regression and linear classification, and further enhanced my use of the python language. In the realization of the logistic regression, the direct use of the loss function of the book, but did not notice, the book y labels are 1 and 0, and the experiment y labels are 1 and -1 so the LOSS function does not converge, do not bother It was a long time since I discovered this problem, which also deepened my understanding of the logistic regression loss function.