The Experiment Report of Machine Learning



Grade:

Undergraduate

December 12, 2017

Student ID：

201530612378

Supervisor:

Mingkui Tan

Author:

Xiaotao Liu

**SUBJECT:**SOFTWARE ENGINEERING

**SCHOOL:** SCHOOL OF SOFTWARE ENGINEERING

Comparison of Various Stochastic Gradient Descent for solving Classification Problem

Abstract—The purpose of this experiment is to request us to independently accomplish logistic regression and linear classification (SVM) by using SGD. And then compare various optimization algorithm of stochastic Gradient on solving classification problem such as SVM and logistic regression , with respect to further understand the essence of the 4 algorithms and their differences.

# INTRODUCTION

Logistic regression and linear classification are the two of most fundamental machine learning models. Stochastic gradient descent(SGD) is an improved version of traditional GD.It accelerates the process that the model reaches its convergence point. From the overall effect, most of the time it can only approach the local optimal solution, so it is suitable for larger training set case. This experiment aims to compare Logistic regression and SVM using SGD to help understanding the differences and relations. What’s more, we compare 4 kinds of optimization algorithms of stochastic gradient descent on classification problem and understand their differences. Lastly, we practice SVM on larger data to have a better command of its principles.

# METHODS AND THEORY

## A.The selected loss function and its derivatives

### Logistic regression:

The loss function I select is cross-entropy cost function.



where 

Its derivatives:



While, 

So,

### Linear classification(SVM):

The loss function I select is cross-entropy cost function:

** **

Its derivatives:

1.

2.

if**,**





else



It can turn into another form:

****

if **** then**,**

else ****

*B*. Experiment steps:

*Logistic Regression and Stochastic Gradient Descent*

1. Load the training set and validation set.
2. Initialize logistic regression model parameters, you can consider initializing zeros, random numbers or normal distribution.
3. Select the loss function and calculate its derivation, find more detail in PPT.
4. Calculate gradient G 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, , and .
7. Repeat step 4 to 6 for several times, and drawing graph of ，， and with the number of iterations.

*Linear Classification and Stochastic Gradient Descent*

1. Load the training set and validation set.
2. Initialize SVM model parameters, you can consider initializing zeros, random numbers or normal distribution.
3. Select the loss function and calculate its derivation, find more detail in PPT.
4. Calculate gradient G 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 ，，and ..
7. Repeat step 4 to 6 for several times, and drawing graph of ，，and . with the number of iterations.

# Experiment

## Data Set using in the experiment and data analysis

Data set: We use a9a of LIBSVM Data, including 32561/16281(testing) samples and each sample has 123/123 (testing) features.

Data analysis:There are 123 features and 1 label in training set. However, it has 122 features and 1 label in testing set.After checking, I find that the last features in training set don’t be in testing set after compression.

## Initialization of model parameters

For both Logistic regression and SVM,I initialize their parameters into zeros

## Implementation

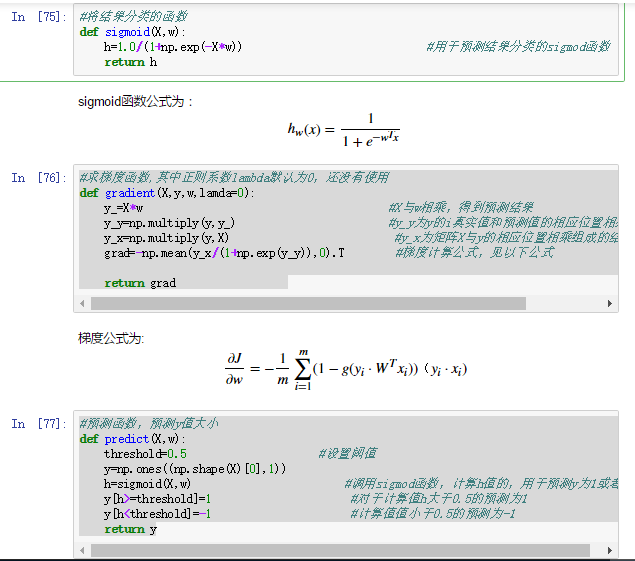
### Logistic regression:

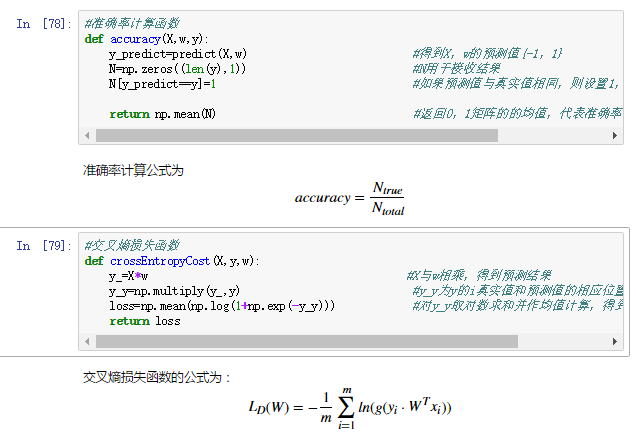
Having defined loss function(cross-entropy cost function) and its gradient, we can use SGD to realize logistic regression. We use five method respectively to reach the local optimal solution including SGD(without optimization), NAG, RMSProp, AdaDelta and Adam. The super parameters we select are as follows.

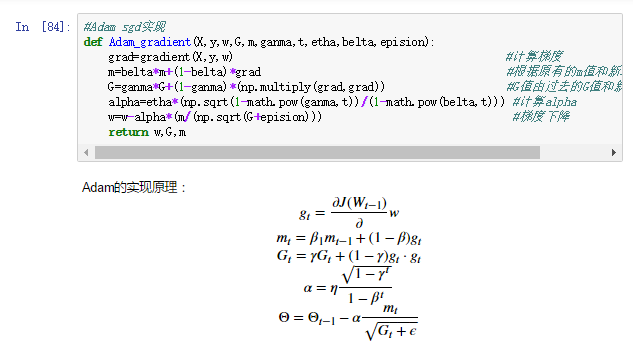
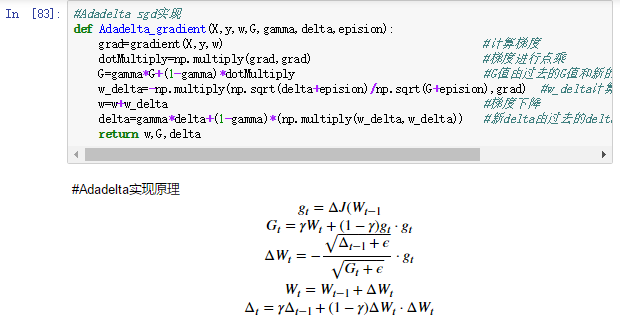
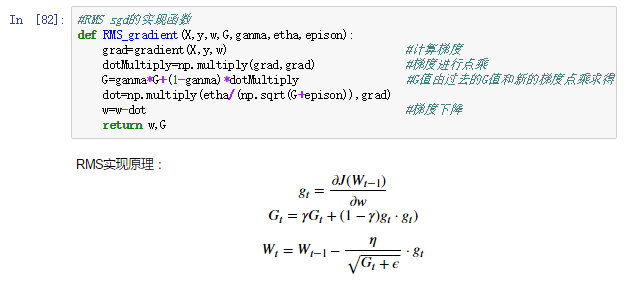
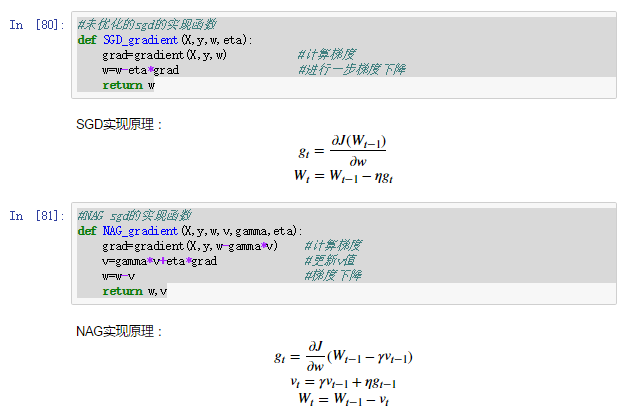
|  |  |  |
| --- | --- | --- |
| SGD | learning rate(eta) | 0.008 |
| epcohs | 4000 |
| NAG | learning rate(eta) | 0.001 |
| gamma | 0.9 |
| epcohs | 4000 |
| RMSProp | learning rate(eta) | 0.001 |
| gamma | 0.9 |
| epsilon | 0.0001 |
| epcohs | 4000 |
| AdaDelta | learning rate(eta) | 0.00001 |
| gamma | 0.9 |
| epsilon | 1e-6 |
| epcohs | 4000 |
| Adam | learning rate(eta) | 0.0008 |
| belta | 0.9 |
| gamma | 0.9 |
| epsilon | 0.0001 |
| epcohs | 4000 |

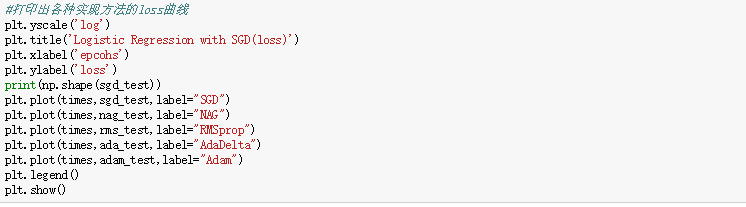
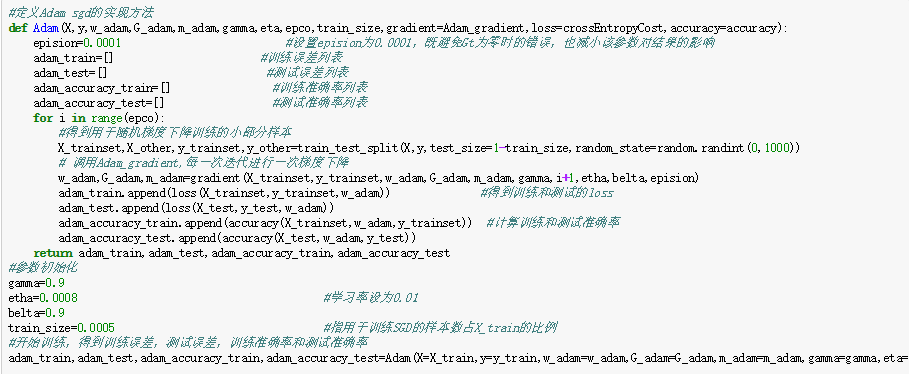
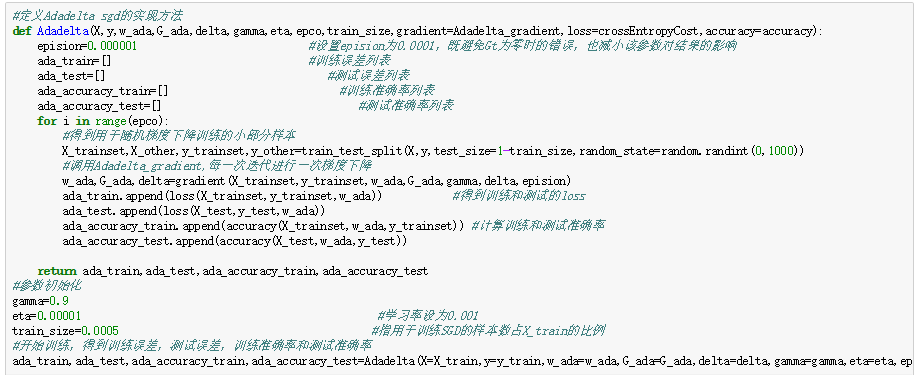
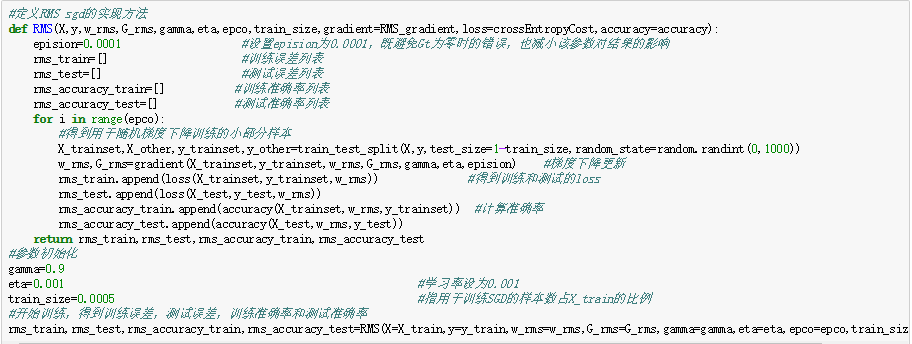
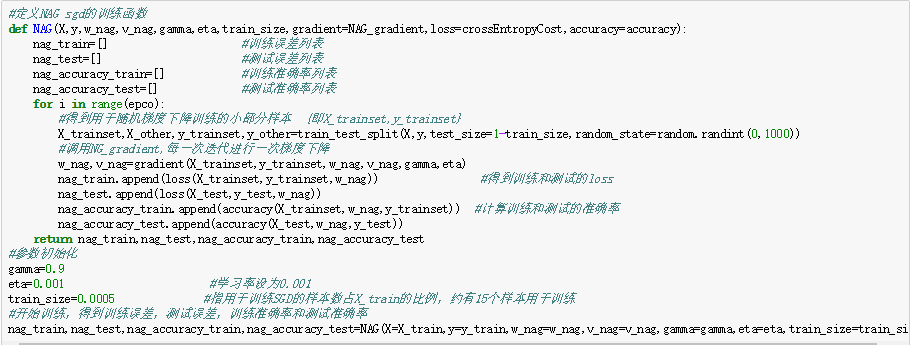
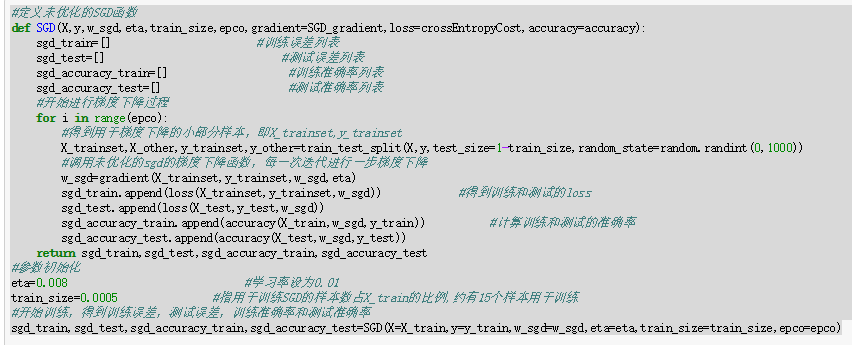
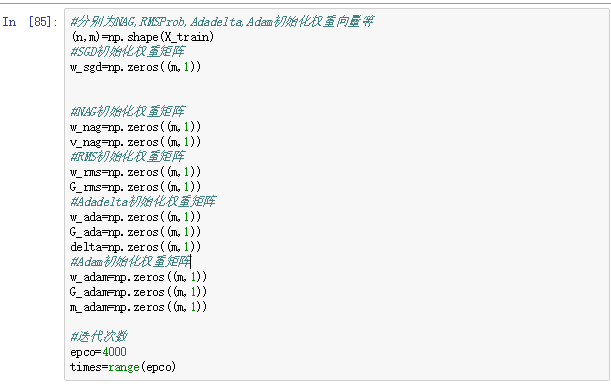
Next, we program to implement the above methods using python. The following are the screenshots of source code.



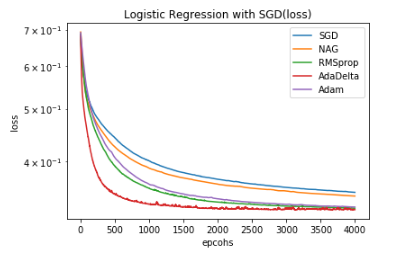








Then, we get the following loss graphs as results after running the program.



Result Analysis:

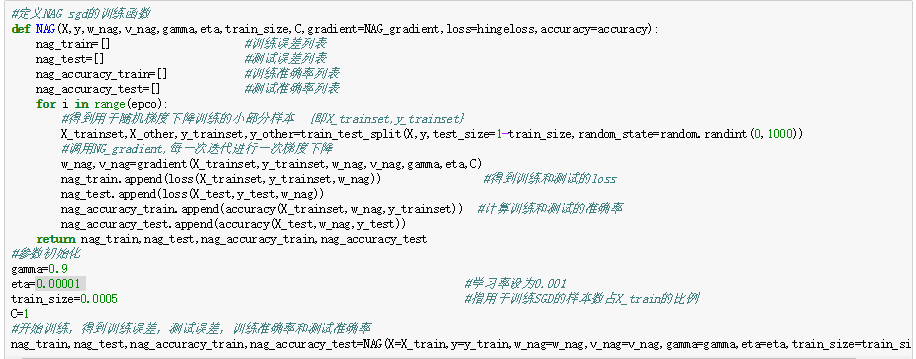
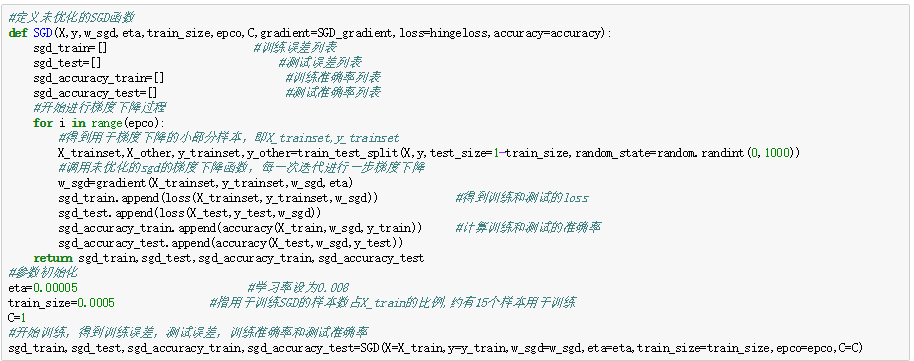
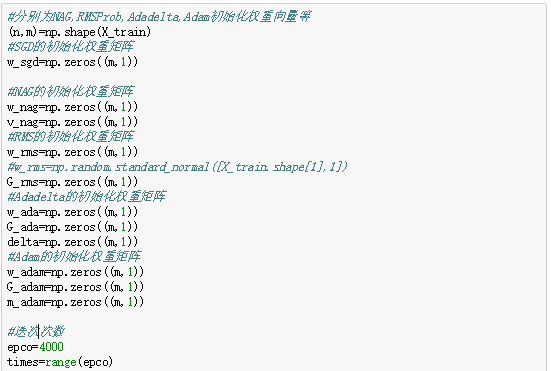
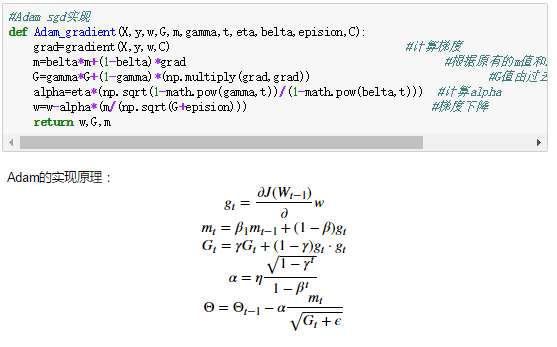
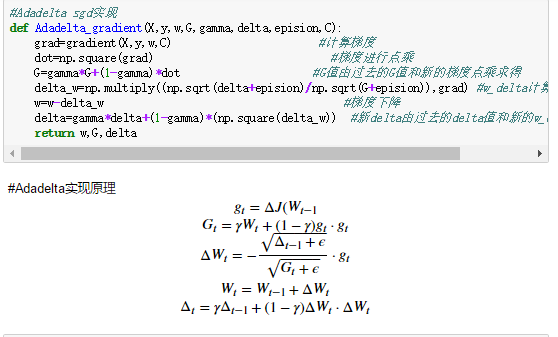
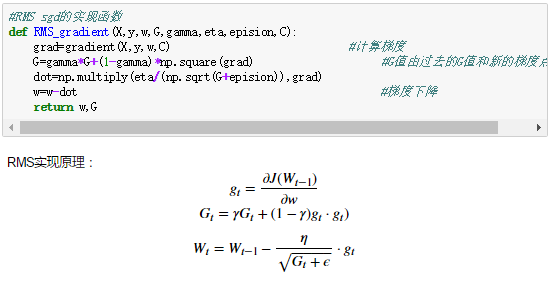
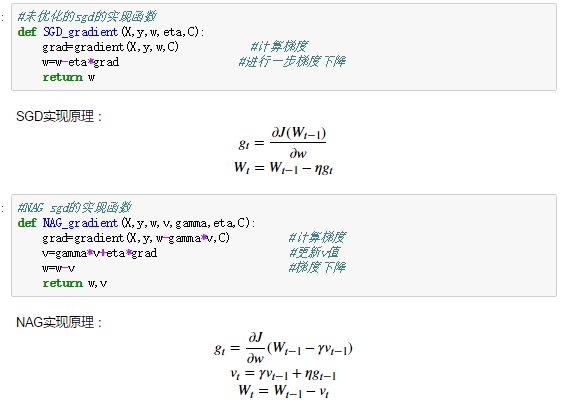
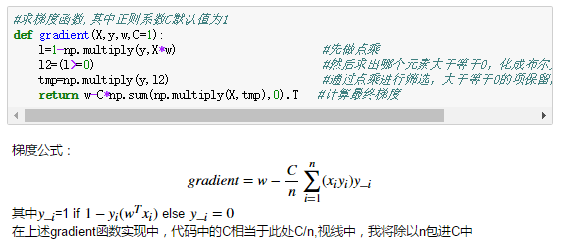
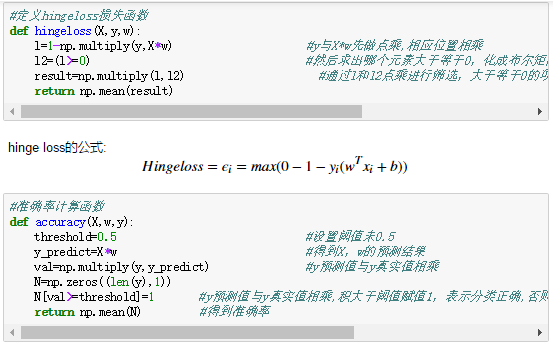
1. From the graph we find that AdaDelta reaches the local optimal solution fastest in solving logistic regression, but it also has obvious shaking at end than others. Its biggest characteristic is self-adaptive and high-speeding. That is because it only use first-order information and has small calculation costs.
2. RMSprop is a variant of AdaDelta, slightly less effective than AdaDelta. From the graph, we find that it have a more stable but slower curve. It is suitable for handling non-stationary targets
3. Adam is essentially RMSprop with momentum terms. It has a closely same result as RMSprop at end but slower than it at first. From its formula,it needs less memory requirement and calculate different adaptive learning rates for different parameters
4. By contrast, SGD is slowest one for which it use typically learning rate and gradient to update w. Which has a slow convergence speed. NAG plays relatively well than SGD.
5. Four methods reaches optimal solution far faster than traditional GD because they only randomly use a set of training set to train model.

### Linear classification(SVM):

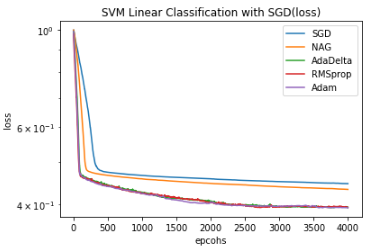
Having defined loss function(hinge loss )and its gradient, we can use SGD to get realize the SVM. We use five method respectively to reach the local optimal solution including SGD(without optimization), NAG, RMSProp, AdaDelta and Adam. The super parameters we select are as follows.

|  |  |  |
| --- | --- | --- |
| SGD | learning rate(eta) | 0.0005 |
| C | 1 |
| threshold | 0.5 |
| epcohs | 4000 |
| NAG | learning rate(eta) | 0.00001 |
| gamma | 0.9 |
| C | 1 |
| threshold | 0.5 |
| epcohs | 4000 |
| RMSProp | learning rate(eta) | 0.001 |
| gamma | 0.9 |
| epsilon | 0.00001 |
| C | 1 |
| threshold | 0.5 |
| epcohs | 4000 |
| AdaDelta | learning rate(eta) | 0.000001 |
| gamma | 0.9 |
| epsilon | 1e-7 |
| C | 1 |
| threshold | 0.5 |
| epcohs | 4000 |
| Adam | learning rate(eta) | 0.001 |
| belta | 0.9 |
| epsolon | 0.0001 |
| C | 1 |
| threshold | 0.5 |
| gamma | 0.9 |

Next, we program to implement the above methods. The following are the screenshots of source code.



We get the following loss graphs as results after running the program.



Result analysis:

1. From the graph we find that AdaDelta, RMSprop, Adam perform almost when solving SVM, but AdaDelta also has obvious shaking at end than others that is because its biggest characteristic is self-adaptive and high-speeding.
2. RMSprop is a variant of AdaDelta, slightly less effective than AdaDelta. From the graph, we find that it have a more stable curve than AdaDelta. That means it is more suitable for handling non-stationary targets
3. Adam is essentially RMSprop with momentum terms. It has a closely same result as RMSprop. From its formula,it needs less memory requirement and calculate different adaptive learning rates for different parameters
4. By contrast, SGD is slowest one for which it use typically learning rate and gradient to update w.It plays a more stable curve then three algorithms above while it has a slow convergence speed. NAG plays relatively well than SGD. And It’s curve is also more stable more first three algorithm.
5. Four methods reaches optimal solution far faster than traditional GD because they only randomly use a set of training set to train model.

## D.Similarities and differences between linear regression and linear classification:

1. Similarities:The fundamental purpose of both is the same. Logistic regression is a classifier, not really regression. The purpose of these two loss functions is to increase the weight of the data points that have a greater impact on the classification and reduce the weight of the data points that have a smaller relationship with the classification.SVM processing method only considers support vectors.And the logistic regression through nonlinear Mapping significantly reduces the weight of points farther from the classification plane and relatively increases the weight of the data points most relevant to the classification.
2. Differences:
   1. From the objective function point of view, the difference is that logistic loss is used in logistic regression, svm uses hinge loss.
   2. SVM more belongs to the non-parametric model, and logistic regression is a parameter model so that they are essentially different.
   3. The major difference between them is the way they evaluate final super plane.

# conclusion

1. SGD improves traditional GD with a faster converging speed and considerable result. SGD has more performance while deal with a large data set.
2. Four specific methods are different improvement of plain SGD with respective advantages. From this experiment, I can summaries some features of these four algorithms whild solving the classification.
   1. SGD generally takes longer to train, but with good initialization and learning rate scheduling schemes, the results are more reliable
   2. Adadelta, RMSprop, Adam are relatively similar algorithms that perform similarly under similar conditions. As an adaptive algorithm, AdaDelta tends to be more convergent and prone to frequent jitter
   3. RMSprop and Adam are driven by momentum, often with relatively fast and stable performance
3. The four kinds of optimization algorithms of SGD have different performance in solving the classification problem, but all of them can speed up the convergence speed and reach a lower error
4. Logistic regression and linear classification both solves classification problem to predict new samples. The major difference between them is the way they evaluate final super plane.SVM more belongs to the non-parametric model, and logistic regression is a parameter model so that they are essentially different.