華南理工大學

South China University of Technology

The Experiment Report of *Machine Learning*

College Software College

Subject Software Engineering

Members 梁婧

Student ID 201530741368

E-mail qwers97@126.com

Tutor MingKui Tan

Date submitted 2017.12.2

1. Topic: Logistic Regression, Linear Classification and Stochastic Gradient Descent

2. Time: 2017/12/2

3. Reporter: 梁婧

4. Purposes:

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.

5. Data sets and data analysis:

Experiment uses a9a of LIBSVM Data, including 32561/16281(testing) samples and each sample has 123/122 (testing) features.

For the reason that the feature in the validation is no equal to that in training set, we use load_svmlight_file: n_feature to define the number of features in validation set.

6. Experimental steps:

Logistic Regression and Stochastic Gradient Descent

1. **Load the training set and validation set.**
2. **Initalize logistic regression model parameters, you can consider**

initalizing zeros, random numbers or normal distribution.

3. **Select the loss function and calculate its derivation, find more detail in PPT.**
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.**
7. **Repeate step 4 to 6 for several times, and drawing graph of losses with the number of iterations.**

Linear Classification and Stochastic Gradient Descent

1. **Load the training set and validation set.**
2. **Initalize SVM model parameters, you can consider initalizing zeros, random numbers or normal distribution.**
3. **Select the loss function and calculate its derivation, find more detail in PPT.**
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.

7. **Repeate step 4 to 6 for several times, and drawing graph of losses with the number of iterations.**

7. Code:

SEE the .ipynb files, with all the code, comments and curve graphs.

(Fill in the contents of 8-11 respectively for logistic regression and linear classification)

8. **The initialization method of model parameters:**

   All the parameters of logistic regression and linear classification are initialized to zero.

9. The selected loss function and its derivatives:

Logistic Regression:

The logic loss function:

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

$$cost(h_\theta(x), y) = -y_i log(h_\theta(x)) - (1 - y_i)log(1 - h_\theta(x))$$

and the derivatives is:

gradient=$(h_\theta(x_i)-y_i)x_i$

Linear Classification:

I select the hinge loss as the loss function:

$L(y)=C^* \sum\max(0,1-y^*wx)$

the derivatives is:

$-y_ix_i$ $1-y_i{*}w^Tx_i>=0$

0 $1-y_i{*}w^Tx_i<0$

10. Experimental results and curve:(Fill in this content for various methods of gradient descent respectively)

Hyper-parameter selection:

Regression:
epsilon is 1000, each time use 30 samples to update the gradient
NAG:
gramma=0.99
η =0.003
RMSProp:
gramma=0.9
ε =10^{-8}
η=0.003
AdaDelta:
gramma=0.95
ε=10-8
Adam:
beta1=0.9
beta2=0.99
ε =10-8
η=0.002

Classification:

epsilon is 250, each time use 30 samples to update the gradient, and the C is set to 0.9

NAG:
gramma=0.8

η=0.001

RMSProp:

gramma=0.94

ε=10-8

η=0.001

AdaDelta:

gramma=0.97

ε=10-6

Adam:

beta1=0.9

beta2=0.999

ε=10-6

η=0.0015

Predicted Results (Best Results):

Regresssion:

NAG:

loss less than 0.6 after 600 iterations

RMSProp:

loss less than 0.6 after 190 iterations

AdaDelta:

loss less than 0.6 after 900 iterations

Adam:

loss less than 1.2 after 800 iterations

Classification:

NAG:

loss less than 4700 after 225 iterations

RMSProp:

loss less than 4600 after 200 iterations

AdaDelta:

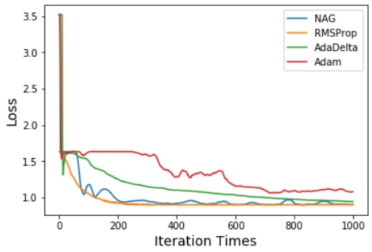
loss less than 4500 after 175 iterations

Adam:

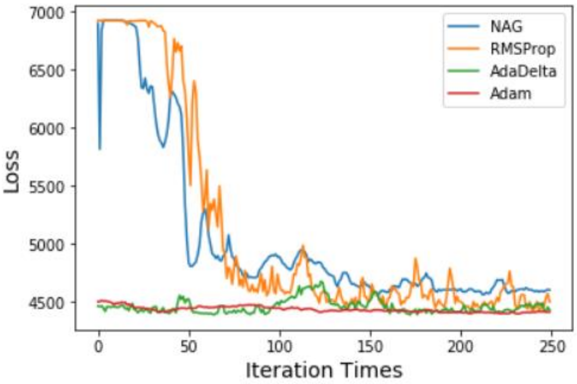
loss less than 4500 after 125 iterations

Loss curve:

Regression:



Classification:



11. Results analysis:

Regression:

All the ways will finally converge, no matter fast or slow.

NAG:

slower than RMSProp, and has more waves than other algorithms.

RMSProp:

smoothly descent and quickly converges.

AdaDelta:

converges slower than other method, it is because it do not have a default learning rate.

Adam:

loss fall quickly, but then has a strange platform, no matter how I check my code or change the parameters, this platform cannot be erased.

Classification:

NAG:

loss fall quickly first, then gradually converges, there are so many waves, but gradually become smaller as the iteration increases.

RMSProp:

acts like the NAG, but converges quicker and has less waves than NAG.

AdaDelta:

converge speed just as fast as RMSProp, but its curve seems more smooth.

Adam:

no matter how I adjust the parameters and learning rate, it just has a little fall in loss curve.

12. Similarities and differences between logistic regression and linear classification：

They both can solve the classification problem.

They both can output continuous values

However:

The linear classification uses a linear model, but the logistic model is a non-linear model.

The linear classification use 0 as threshold, by deciding the points below the line or not. The logistic model just output the

possibility, so we usually select 0.5 as threshold.

13. Summary:

Stochastic Gradient Descent is a good way to optimize the parameters in our model for big data sets, each time, it use only one sample to update the gradient and weights. However, it cannot gain a good result as the batch gradient descent. To optimize this, we apply four algorithm to SGD. They usually use the history gradient information or make a prediction to the gradient or the learning rate to optimize the gradient and learning rate.

Logistic regression is a classical classification model, it classifies samples by output the possibility with its non-linear sigmoid function.