The Experiment Report of Machine Learning

SCHOOL: SCHOOL OF SOFTWARE ENGINEERING

SUBJECT: SOFTWARE ENGINEERING

Author:
Yihui Zhu

Supervisor:
Mingkui Tan

Student ID：
201530613955

Grade:
Undergraduate

December 12, 2017

Logistic Regression, Linear Classification and Stochastic Gradient Descent

Abstract— It is an experiment of machine learning about Logistic Regression, Linear Classification and Stochastic Gradient Descent(SGD). In this paper, by introducing a dataset which has the 0-1 control value to each input feature, we propose that both logistic regression and linear classification use different optimized methods of SGD to update the model parameters and build the models. Experimental results on the dataset show that the gap between different optimized methods and strengths and weakness of SGD.

I. INTRODUCTION

In the experiment, I try to use different optimized methods of SGD to realize logistic regression and linear classification separately and adjust parameters to get the right results.

The experiment is to compare and understand the difference between gradient descent and stochastic gradient descent, compare and understand the differences and relationships between Logistic regression and linear classification and further understand the principles of SVM and practice on larger data.

I use python3 as the environment of experiment, and I am going to accomplish the experiment by building the models, adjust the parameters, and using dataset to validate the effects of these models.

I wish to see that these methods have rapider convergence, smoother loss curve, and better learning and diagnostic properties to complex data, and the results have higher accuracies and lower losses finally.

II. METHODS AND THEORY

I am going to complete the models of logistic regression and linear classification, and use different optimized methods of SGD to update the parameters so that the loss curve can be smoother and the results could be more accurate. Then I will talk about those methods and theory.

Firstly, let's introduce the loss function and gradient function of the two models.

Logistic regression: loss function: $J(\mathbf{w})$ = -1/n $[\sum_{i=1}^{n} y_i \log h_{\mathbf{w}}(\mathbf{x}_i) + (1 - y_i) \log (1 - h_{\mathbf{w}}(\mathbf{x}_i))]$; gradient function: $\frac{\partial J(w)}{\partial (w)} = (h_{\mathbf{w}}(\mathbf{x}) - y)x$

Linear classification: loss function: $\min_{w,b} L : \frac{\|w\|_2}{2} + \frac{C}{n}\sum_{i=1}^{n} \max(0;\ 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b))$; gradient function: if $1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 0, g_{\mathbf{w}}(\mathbf{x}_i) = -y_i x_i$; if $1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b) < 0, g_{\mathbf{w}}(\mathbf{x}_i) = 0$.

Then, it is about the SGD: the SGD has some effects, but it also has problems, like learning rate has big influence on the convergence, the noise may bring the bad results, etc. There is the principle:

$$\mathbf{g}_t \leftarrow \nabla J_i(\boldsymbol{\theta}_{t-1})$$

$$\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \eta \mathbf{g}_t$$

Lastly, I will make use of the four different optimized methods. They are NAG, RMSprop, AdaDelta, Adam. all the methods are based on SGD, but they have more strengths such like rapider convergence or smoother loss curve than SGD.

NAG: it is a slightly different version of the momentum update that has recently been gaining popularity.

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

$$\mathbf{v}_t \leftarrow \gamma \mathbf{v}_{t-1} + \eta \mathbf{g}_t$$

$$\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \mathbf{v}_t$$

RMSprop: RMSprop is a very effective, but currently unpublished adaptive learning rate method.

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

$$G_t \leftarrow \gamma G_t + (1 - \gamma)\mathbf{g}_t \odot \mathbf{g}_t$$

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

AdaDelta: It doesn't have to initialize the learning rate at first, but its speed maybe slow sometimes.

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

$$G_t \leftarrow \gamma G_t + (1 - \gamma)\mathbf{g}_t \odot \mathbf{g}_t$$

$$\Delta\boldsymbol{\theta}_t \leftarrow -\frac{\sqrt{\Delta_{t-1} + \epsilon}}{\sqrt{G_t + \epsilon}} \odot \mathbf{g}_t$$

$$\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} + \Delta\boldsymbol{\theta}_t$$

$$\Delta_t \leftarrow \gamma\Delta_{t-1} + (1 - \gamma)\Delta\boldsymbol{\theta}_t \odot \Delta\boldsymbol{\theta}_t$$

Adam: Adam is a recently proposed update that looks a bit like RMSProp with momentum. It can do initialization bias correction and has very strong effects.

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

$$\mathbf{m}_t \leftarrow \beta_1 \mathbf{m}_{t-1} + (1 - \beta_1)\mathbf{g}_t$$

$$G_t \leftarrow \gamma G_t + (1 - \gamma)\mathbf{g}_t \odot \mathbf{g}_t$$

$$\alpha \leftarrow \eta \frac{\sqrt{1 - \gamma^t}}{1 - \beta^t}$$

$$\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \alpha \frac{\mathbf{m}_t}{\sqrt{G_t + \epsilon}}$$

III. EXPERIMENT

In this section, I analyze the performance of SGD for logistic regression and linear classification, and I also investigate how the different optimized methods improve my algorithm.

3.1 Data sets and data analysis:

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

3.2 Experimental steps:

The experimental code and drawing are completed on jupyter.

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 L_{NAG}， $L_{RMSprop}$， $L_{AdaDelta}$ and L_{Adam}.
7. Repeate step 4 to 6 for several times, and **drawing graph of L_{NAG}，$L_{RMSprop}$，$L_{AdaDelta}$ and L_{Adam} 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 L_{NAG}， $L_{RMSprop}$， $L_{AdaDelta}$ and L_{Adam}.
7. Repeate step 4 to 6 for several times, and **drawing graph of L_{NAG}，$L_{RMSprop}$，$L_{AdaDelta}$ and L_{Adam} with the number of iterations**.

Finishing experiment report according to result: The template of report can be found in example repository.

3.3 Experimental results and curve:

The below are my experimental results and the loss curves of two models.

Logistic Regression and Stochastic Gradient Descent

All zero initialization;

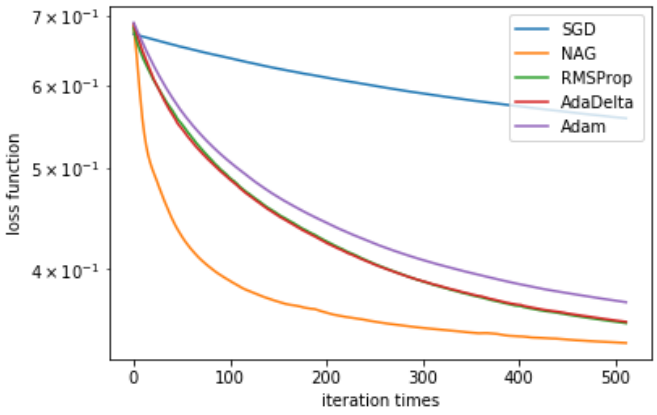
Logistic regression: loss function: $J(\mathbf{w}) = -1/n\,[\sum_{i=1}^{n} y_i \log h_{\mathbf{w}}(\mathbf{x}_i) + (1 - y_i)\log(1 - h_{\mathbf{w}}(\mathbf{x}_i))]$; gradient function: $\frac{\partial J(w)}{\partial (w)} = (h_{\mathbf{w}}(\mathbf{x}) - y)x$

Iteration time: 512; Number of samples each time: 128;

The hyper-parameter of various methods:

Methods	Hyper-parameter
SGD	η=0.05;
NAG	η=0.01; γ=0.9;
RMSprop	η=0.001; γ=0.9;
AdaDelta	γ=0.95;
Adam	η=0.001; γ=0.999; β=0.9;

The loss curves:



Linear Classification and Stochastic Gradient Descent

All zero initialization;

Linear classification: loss function: $\min_{w;b} L : \frac{\|w\|_2}{2} + \frac{C}{n}\sum_{i=1}^{n} \max(0;\ 1 - y_i(\mathbf{w}^{\top}\mathbf{x}_i + b))$; gradient function: if $1 - y_i(\mathbf{w}^{\top}\mathbf{x}_i + b) \geq 0, g_w(\mathbf{x}_i) = -y_i x_i$; if $1 - y_i(\mathbf{w}^{\top}\mathbf{x}_i + b) < 0, g_w(\mathbf{x}_i) = 0$.

Iteration time: 400; Number of samples each time: 128;

The hyper-parameter of various methods:

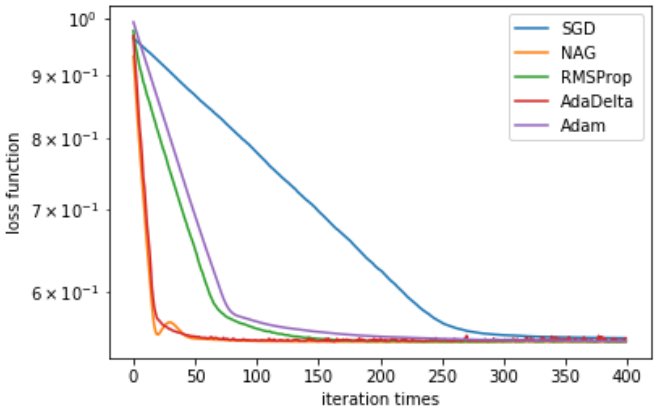
Methods	Hyper-parameter
SGD	η=0.02;
NAG	η=0.02; γ=0.9;
RMSprop	η=0.001; γ=0.9;
AdaDelta	γ=0.95; ε=1e-6;
Adam	η=0.001; γ=0.999; β=0.9;

Threshold: -0.825

The accuracy of various methods:

Methods	Accuracy
SGD	0.7788219396842946
NAG	0.8099625330139426
RMSprop	0.8100239543025612
AdaDelta	0.8095325839936122
Adam	0.8079356304895277

The loss curves:



IV. CONCLUSION

In this experiment, I use logistic regression and linear classification which are updated by SGD to learn a sparse feature subset for classification, and improve the performance of two models by introducing four optimized methods. There are some discoveries that come from the process. First of all, I learn that the difference of particular gradient descent and SGD is number of samples selected to calculate the gradient. Next, though both logistic regression and linear classification can handle classification problems, logistic regression uses logistic loss and linear classification adopt hinge loss, the two losses try to increase the weight of data points that have greater impact on the classification. What's more, the SVM which linear classification used considers the part(support vector), but logistic regression considers the global situation. At last, we can apply the optimized methods to solve SGD problems, which gave the loss function faster convergence and higher accuracy. However, there were some problems with my experiment. In particular, the effect of Adam method may be the best, but two loss curves didn't show the feature. Moreover, the selection of parameters could not be great, so the curves were not beautiful.

V. REFERENCES

The website:

https://blog.slinuxer.com/2016/09/sgd-comparison;
https://www.zybuluo.com/chenyaofo/note/961083;

The slides:

2.Linear Classfication;
3.Logistic Regression and Softmax Regression.