华南理工大学

South China University of Technology

The Experiment Report of Deep Learning

SCHOOL: SCHOOL OF SOFTWARE ENGINEERING

SUBJECT: SOFTWARE ENGINEERING

Author:
Yiqun Wu

Supervisor:
Mingkui Tan

Student ID：
201720145082

Grade:
Graduate

December 14, 2017

Logistic Regression, Linear Classification and Stochastic Gradient Descent

Abstract—this experiment applies four optimized methods, NAG，RMSProp，AdaDelta and Adam in logistic regression and linear classification. The influence of learning rate in the four different methods is also discussed in logistic regression and linear classification. Finally graphs of loss with the number of iterations in validation dataset and a comparison of these four methods are presented. This experiment shows the apply of stochastic gradient descent using different optimized methods in logistic regression and linear classification.

I. INTRODUCTION

Logistic regression is a regression model where the dependent variable (DV) is categorical. This experiment covers the case of a binary dependent variable—that is, where the output can take only two values, "0" and "1", which represent outcomes such as pass/fail, win/lose, alive/dead or healthy/sick. Note that in this experiment dataset “0” is marked as “-1”. Logistic regression was developed by statistician David Cox in 1958. The binary logistic model is used to estimate the probability of a binary response based on one or more predictor (or independent) variables (features). It allows one to say that the presence of a risk factor increases the odds of a given outcome by a specific factor. Logistic regression is used in various fields, including machine learning, most medical fields, and social sciences. In the terminology of economics, logistic regression is an example of a qualitative response/discrete choice model. Conditional random fields, an extension of logistic regression to sequential data, are used in natural language processing.

Linear classification is a classification algorithm (Classifier) that makes its classification based on a linear predictor function combining a set of weights with the feature vector. The decision boundaries of linear classification is flat. In the field of machine learning, the goal of statistical classification is to use an object's characteristics to identify which class (or group) it belongs to. A linear classifier achieves this by making a classification decision based on the value of a linear combination of the characteristics. An object's characteristics are also known as feature values and are typically presented to the machine in a vector called a feature vector. Such classifiers work well for practical problems such as document classification, and more generally for problems with many variables (features), reaching accuracy levels comparable to non-linear classifiers while taking less time to train and use.

Stochastic gradient descent (often shortened to SGD), also known as incremental gradient descent, is a stochastic approximation of the gradient descent optimization and iterative method for minimizing an objective function that is written as a sum of differentiable functions. In other words, SGD tries to find minima or maxima by iteration. In other cases, evaluating the sum-gradient may require expensive evaluations of the gradients from all summand functions. When the training set is enormous and no simple formulas exist, evaluating the sums of gradients becomes very expensive, because evaluating the gradient requires evaluating all the summand functions' gradients. To economize on the computational cost at every iteration, stochastic gradient descent samples a subset of summand functions at every step. This is very effective in the case of large-scale machine learning problems.

The motivation of this experiment is as follows:

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

A. Logistic regression

Assume that the labels are binary:$y_i \in \{0,1\}$, the logistic model can be represented by the following formula:

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

The probability can be described as:

$$p = \begin{cases} h_w(x) & y_i = 1 \\ 1 - h_w(x) & y_i = 0 \end{cases}$$

Log-likelyhood loss function:

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

Then, we have:

$$\frac{\partial J(w)}{\partial \mathrm{w}} = \frac{1}{n}\sum_{i=1}^{n}(h_w(x_i) - y_i)x_i$$

So the gradient can be updated by the following formula:

$$\mathrm{w} := w - \frac{1}{m}\sum_{i=1}^{n}\alpha(h_w(x_i) - y_i)x_i$$

In which α is the learning rate, m is the number of samples in a batch.

B. *Linear classification*

For linear classification, the selected loss function and its derivatives can be described by the following formulas:

$$\min_{w,b} f: \frac{\|w\|^2}{2} + C\sum_{i=1}^{n} \max(0, 1 - y_i(w^T x_i + b))$$

$$g_w(x_i) = \begin{cases} -y_i x_i & 1 - y_i(w^T x_i + b) \geq 0 \\ 0 & 1 - y_i(w^T x_i + b) < 0 \end{cases}$$

$$g_b(x_i) = \begin{cases} -y_i & 1 - y_i(w^T x_i + b) \geq 0 \\ 0 & 1 - y_i(w^T x_i + b) < 0 \end{cases}$$

At last we have:

$$\frac{\partial J(w)}{\partial w} = w + C\sum_{i=1}^{m} g_w(x_i)$$

$$\frac{\partial J(w)}{\partial b} = C\sum_{i=1}^{m} g_b(x_i)$$

In which m is the number of samples in a batch.

C. *Stochastic gradient descent*

SGD works similar as GD, but more quickly by estimating gradient from a few examples at a time

Algorithm1: SGD

1 Initialize parameter w and learning rate η

2 **while** an approximate minimum is not obtained do:

3 Randomly select an example i in the training set

4 $w = w - \eta \nabla L_i(w)$

5 end

MSGD works identically to SGD, except that we use more than one training example to make each estimate of the gradient.

Algorithm2: MSGD

1 Initialize parameter w and learning rate η

2 **while** an approximate minimum is not obtained do:

3 Randomly select $|S_k|$ examples in the training set

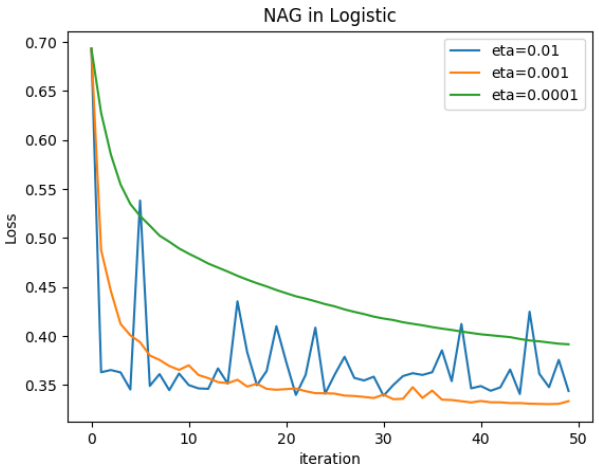
4 $w = w - \eta \frac{1}{|S_k|} \sum_{i \in S_k} \nabla_w L_i(w)$

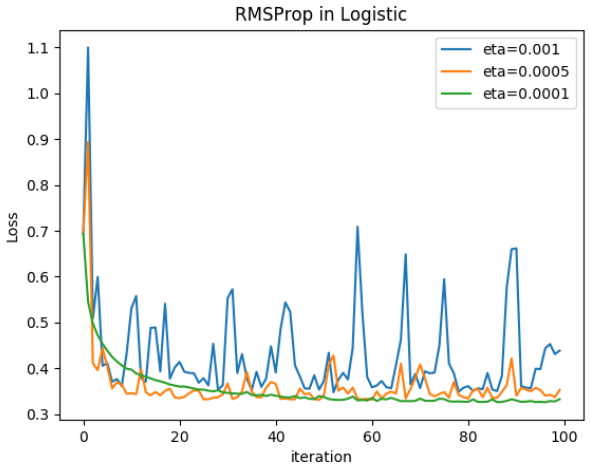
5 end

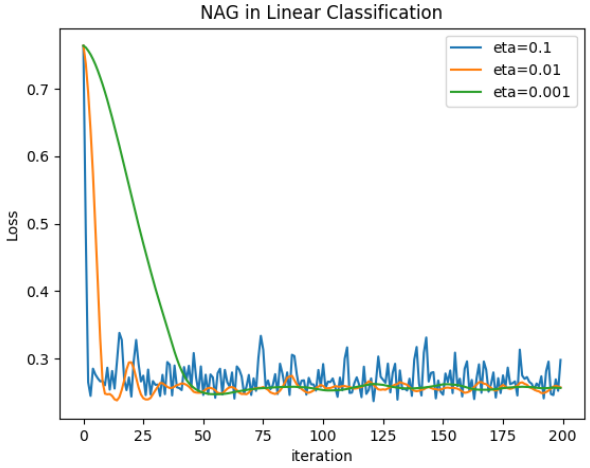
III. EXPERIMENT

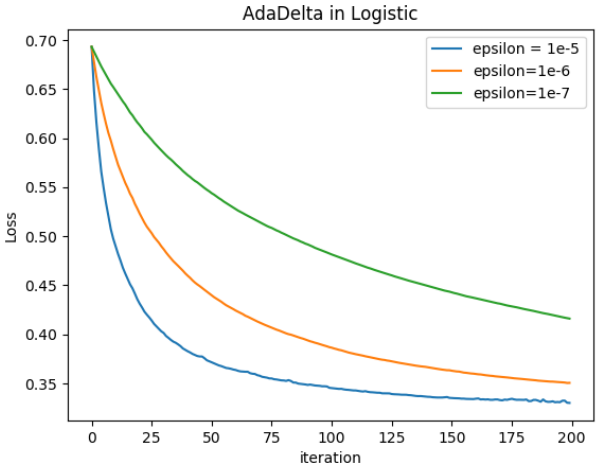
Experiment uses "a9a" dataset of LIBSVM Data, including 32561/16281(testing) samples and each sample has 123/123 (testing) features. Firstly load the training set and validation set. And initialize the logistic regression model parameters and SVM model parameters by setting all zeros. Then define the loss function and calculate its derivation as the formulas in part II. After calculating the gradient toward loss function from partial samples, update model parameters using different optimized methods(NAG，RMSProp，AdaDelta and Adam). 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}, $L_{RMSProp}$, $L_{AdaDelta}$, L_{Adam}. Repeat the above steps for several times, and drawing graph of L_{NAG}, $L_{RMSProp}$, $L_{AdaDelta}$, L_{Adam} with the number of iterations.

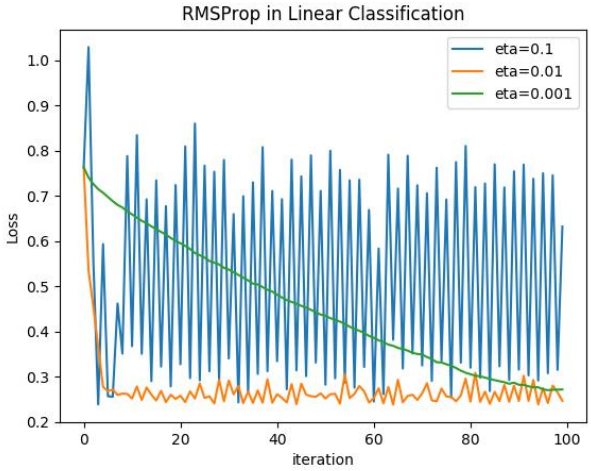
This experiment also discuss the influence of learning rate in the four optimized methods. To realize the influence of learning rate in experiment the other parameters remain the same. As for AdaDelta, which has an adaptive learning rate, this experiment discuss the parameter epsilon from 1e-5 to 1e-7.

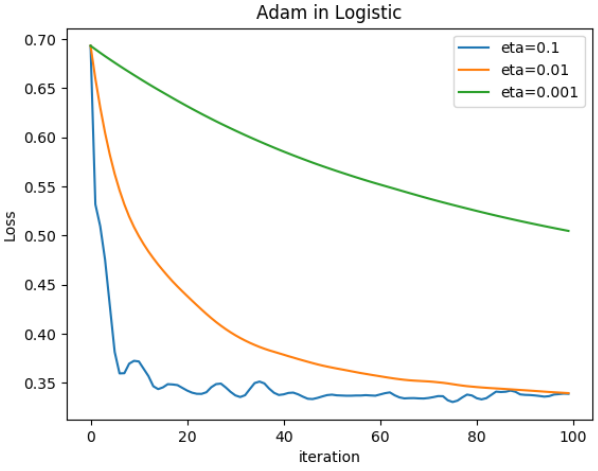


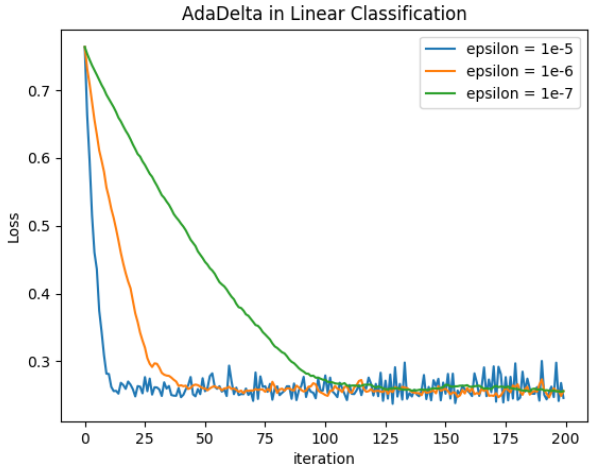

RMSProp in Logistic
eta=0.001
eta=0.0005
eta=0.0001
Loss
iteration

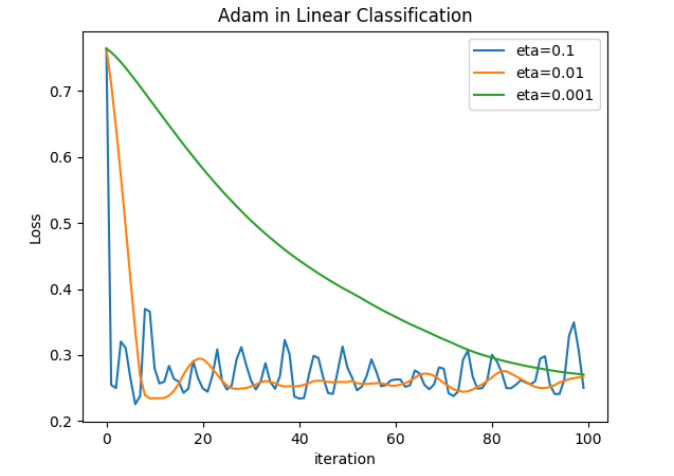

NAG in Linear Classification
eta=0.1
eta=0.01
eta=0.001
Loss
iteration

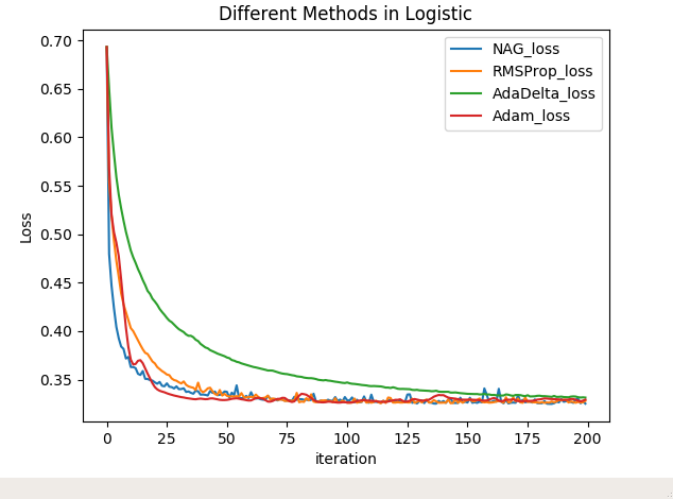

AdaDelta in Logistic
epsilon = 1e-5
epsilon=1e-6
epsilon=1e-7
Loss
iteration

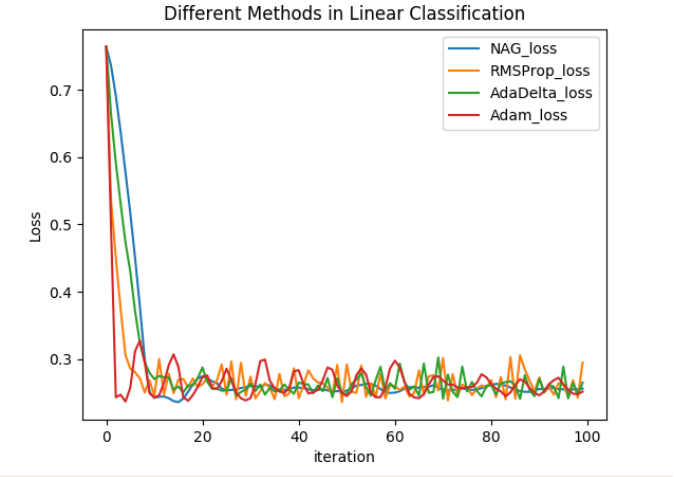

RMSProp in Linear Classification
eta=0.1
eta=0.01
eta=0.001
Loss
iteration


Adam in Logistic
eta=0.1
eta=0.01
eta=0.001
Loss
iteration


AdaDelta in Linear Classification
epsilon = 1e-5
epsilon = 1e-6
epsilon = 1e-7
Loss
iteration







As shown in the above graphs, generally, loss of different optimized methods decade with the number of iterations and finally converge to a value. As the effect of randomness, it will fluctuate on a value. The experiment result also shows that a suitable learning rate is helpful to make the loss converge quickly and reduce volatility. However, excessive learning rate will lead to the case in which the convergence cannot be found. And if the learning rate is too small, loss converge slowly.

IV. CONCLUSION

From this experiment I have better a better understand of the principle of stochastic gradient descent, logistic regression and linear classification. Also, by implementing four different optimized methods, I realize the mechanism of NAG， RMSProp， AdaDelta and Adam, which is powerful in optimization. Through this apply of stochastic gradient descent using different optimized methods in logistic regression and linear classification, I also realize that some parameter have a great influence in the experiment result, and it's necessary to master the technique of tuning parameters.