The Experiment Report of Machine Learning

SCHOOL: SCHOOL OF SOFTWARE ENGINEERING

SUBJECT: SOFTWARE ENGINEERING

Author:
Yu Guo

Supervisor:
Mingkui Tan

Student ID：
2015306115000

Grade:
Graduate

December 9, 2017

Logistic Regression, Linear Classification and Stochastic Gradient Descent

Abstract—In this experiment we run logistic regression and linear classification with SVM using different gradient descent ways including nag, RMSProp, Adadelta, Adam. They have different learning effect.

I. Introduction

We often use gradient descent to solve linear classification problems, and there are several different methods to optimize custom gradient descent, including nag, RMSProp, adaDelta and adam. We will implement these methods in logistic regression and SVM to test these algorithms.

II. Methods and Theory

We use four kinds of gradient descent methods to minimize the loss function of logistic regression and SVM.

Theory of nag:

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

Theory of RMSProp:

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

Theory of adaDelta:

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

Theory of adam:

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

The selected loss function and its derivatives:

Logistic regression:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \mathrm{Cost}(h_\theta(x^{(i)}), y^{(i)})$$
$$\mathrm{Cost}(h_\theta(x), y) = -\log(h_\theta(x)) \quad \text{if } y = 1$$
$$\mathrm{Cost}(h_\theta(x), y) = -\log(1 - h_\theta(x)) \quad \text{if } y = 0$$

Linear classification:

loss function:

$$\frac{\|\mathbf{w}\|^2}{2} + C\sum_{i=1}^{N} \max(0, 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b))$$

Gradient = ||w|| + 0 yi(Wtxi+b)>=1

||w|| - Cxiyi else

III. Experiment

A. Dataset

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

B. Implementation

Experimental steps of Logistic Regression

1. Load the training set and validation set.
2. Initialize logistic regression model parameters with all zero.
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.
7. Repeat step 4 to 6 for several times, and drawing graph of loss with the number of iterations.

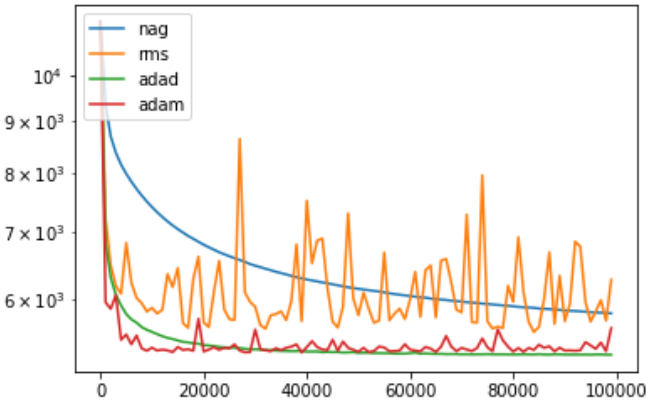
TABLE I
Hyper-parameter selection:

NAG	learing rate = 0.05
RMSProp	gamma = 0.99, epslion = 10e-7, learing rate = 0.05
Adadelta	gamma = 0.9, epslion = 10e-7, learing rate = 0.05
Adam	gamma1 = 0.9, gamma2 = 0.99, epslion = 10e-7, learing rate = 0.05

TABLE II
Predicted Error Rate for Logistic Regerssion (Best Results)

NAG	0.1595 at 100000 epoch
RMSProp	0.1592 at 18153 epoch
AdaDelta	0.1482 at 50056 epoch
Adam	0.1517 at 23156 epoch

GRAPH I
Loss Curve for Logistic Regression



Experimental steps of Linear Classification

1. Load the training set and validation set.
2. Initialize SVM model parameterswith all zeros.
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.
7. Repeat step 4 to 6 for several times, and drawing graph of loss with the number of iterations.

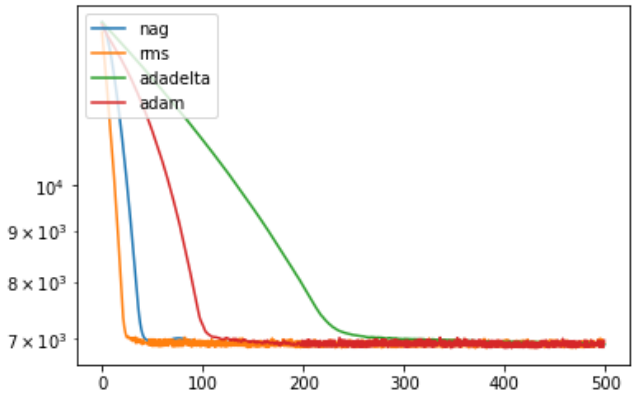
TABLE III
Hyper-parameter selection:

NAG	learing rate = 0.05
RMSProp	gamma = 0.99, epslion = 10e-7, learing rate = 0.05
Adadelta	gamma = 0.9, epslion = 10e-7 , learing rate = 0.05
Adam	gamma1 = 0.9, gamma2 = 0.99, epslion = 10e-7, learing rate = 0.05

TABLE VI
Predicted Error Rate for SVM(Best Results)

NAG	0.1970 at 402 epoch
RMSProp	0.1973 at 173 epoch
AdaDelta	0.1968 at 458 epoch
Adam	0.2035 at 318 epoch

GRAPH II
Loss Curve for Linear regression with SVM:



IV. CONCLUSION

In this experiment, logistic regression performs better than SVM, with lower error rate. However it seems that RMSProp cannot perform well on logistic regression.

As we see the four gradient descent methods, the curves show that Adadelta and NAG are more stable gradient descent ways since the curves have less fluctuate. I find that nag is a bit slower in convergence. RMSprop and Adam are fast in convergence but with more fluctuate. In this experiment Adadelta and Adam shows the best result.