華南理工大學

South China University of Technology

The Experiment Report of *Machine Learning*

School: School of Software Engineering

Subject: Software Engineering

Author:
Dongming Sheng

Supervisor:
Mingkui Tan

Student ID:
201530612699

Grade:
Undergraduate

December 14, 2017

Logistic Regression, Linear Classification and Stochastic Gradient Descent

***Abstract*—For the purpose of comparing the effectiveness of NAG, RMSProp, AdaDelta and Adam, we design this experiment and implement logistic regression and linear SVM using these four methods to update parameters respectively.**

I. Introduction

STOCHASTIC gradient descent is one of the essential parts in deep learning. In this experiment, we managed to have an intuitive understanding of NAG, RMSProp, AdaDelta and Adam by comparing their performances.As a result, AdaDelta converges the fastest and has a better stability than NAG.

II. Methods and Theory

A. Logistic regression

In logistic regression, we select the cross entropy error measure as its loss function:

$$L(w) = \frac{\lambda}{2}\|w\|^2 + \frac{1}{n}\sum_{i=1}^{n} log(1 + e^{-y_i \cdot w^T x})$$

therefore, its gradient respect to w can be expressed as:

$$\nabla_w L(w) = \lambda w - \frac{1}{n}\sum_{i=1}^{n} \frac{e^{-y_i \cdot w^T x}}{1 + e^{-y_i \cdot w^T x}} y_i x_i$$

B. Linear SVM

In linear SVM, we select the hinge loss as its loss function:

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

so the gradients of L(w) respect to w and b are:

$$\nabla_w L(w, b) = w - C\sum_{i=1}^{n} \mathbb{1}\{1 - y_i(w^T x_i + b) > 0\} y_i x_i$$

$$\nabla_b L(w, b) = -C\sum_{i=1}^{n} \mathbb{1}\{1 - y_i(w^T x_i + b) > 0\} y_i$$

C. Optimization methods

Let w_i denotes the parameters in the ith iteration, and the main procedures of NAG, RMSProp, AdaDelta and Adam are shown below:

1) NAG:

$$v_{i+1} \leftarrow \rho v_i - \eta \nabla_w L(w_i)$$

$$w_{i+1} \leftarrow w_i - \rho v_i + (1 + \rho) v_{i+1}$$

2) RMSProp:

$$g_{i+1} \leftarrow \nabla_w L(w_i)$$

$$G_{i+1} \leftarrow \gamma G_i + (1 - \gamma) g_{i+1} \odot g_{i+1}$$

$$w_{i+1} \leftarrow w_i - \frac{\eta g_{i+1}}{\sqrt{G_{i+1}} + \epsilon}$$

3) AdaDelta:

$$g_{i+1} \leftarrow \nabla_w L(w_i)$$

$$G_{i+1} \leftarrow \gamma G_i + (1 - \gamma) g_{i+1} \odot g_{i+1}$$

$$\nabla w_{i+1} \leftarrow -\frac{\sqrt{\nabla_i + \epsilon}}{\sqrt{G_i + \epsilon}} \odot g_{i+1}$$

$$w_{i+1} \leftarrow w_i + \nabla w_{i+1}$$

$$\nabla_{i+1} \leftarrow \gamma \nabla_i + (1 - \gamma) \nabla w_i \odot \nabla w_i$$

4) Adam:

$$g_{i+1} \leftarrow \nabla_w L(w_i)$$

$$v_{i+1} \leftarrow \frac{\beta_1 v_i + (1 - \beta_1) g_{i+1}}{1 - \beta_1^{i+1}}$$

$$G_{i+1} \leftarrow \frac{\beta_2 G_i + (1 - \beta_2) g_{i+1} \odot g_{i+1}}{1 - \beta_2^{i+1}}$$

$$w_{i+1} \leftarrow w_i - \frac{\eta v_{i+1}}{\sqrt{G_{i+1} + \epsilon}}$$

III. Experiments

A. Dataset

In this experiment, we use a9a of LIBSVM Data, which includes 32561 training samples and 16281 testing samples. For each sample, there are 123 features. The negative class is labeled as -1.

B. Implementation

We implement the experiment in the following order:

1) Logistic Regression and Stochastic Gradient Descent:

- Load the training set and validation set.
- Initialize logistic regression model parameters. In this experiment, I use zeros to initialize the parameters.
- Select the loss function and calculate its derivation.
- Calculate gradient G toward loss function from partial samples.
- Update model parameters using different optimized methods(NAGRMSPropAdaDelta 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 $L_{NAG}L_{RMSProp}L_{AdaDelta}$ and L_{Adam}.
- Repeat 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.

2) Linear Classification and Stochastic Gradient Descent:

- Load the training set and validation set.
- Initialize SVM model parameters. In this experiment, I use zeros to initialize the parameters.
- Select the loss function and calculate its derivation.
- Calculate gradient G toward loss function from partial samples.
- Update model parameters using different optimized methods(NAGRMSPropAdaDelta 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 $L_{NAG}L_{RMSProp}L_{AdaDelta}$ and L_{Adam}.
- Repeat 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.

In these two experiment, we use hold-out method to evaluate the models and choose $threshold = 0$ to judge when predicting.Also, we use $batch\_size = 100$ in each iteration.

In logistic regression, we control the number of iterations to be 5000 so that all four optimization methods can converge. However, setting $\lambda = 0$ leads to a better result in our validation set.

In linear SVM, we find out that setting $C = 5$ results in a higher accuracy in validation set for all four methods when having $iter\_num = 10000$.

The hyper-parameters corresponding to different optimization methods are listed in Table I and Table II. The loss curves and the accuracy curves of logistic regression model and linear SVM model using different optimization methods are shown in Fig 1 Fig 4.As a result, AdaDelta converges the fastest and has a better stability than NAG.

TABLE I
HYPER-PARAMETER SELECTION IN LOGISTIC REGRESSION

NAG	$\eta = 1e-3$, $\rho = 0.9$
RMSProp	$\eta = 1e-3$, $\gamma = 0.9$, $\epsilon = 1e-7$
AdaDelta	$\gamma = 0.95$, $\epsilon = 1e-6$
Adam	$\eta = 1e-3$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e-7$

TABLE II
HYPER-PARAMETER SELECTION IN LINEAR SVM

NAG	$\eta = 1e-4$, $\rho = 0.99$
RMSProp	$\eta = 1e-4$, $\gamma = 0.99$, $\epsilon = 1e-7$
AdaDelta	$\gamma = 0.99$, $\epsilon = 1e-6$
Adam	$\eta = 1e-4$, $\beta_1 = 0.99$, $\beta_2 = 0.999$, $\epsilon = 1e-7$

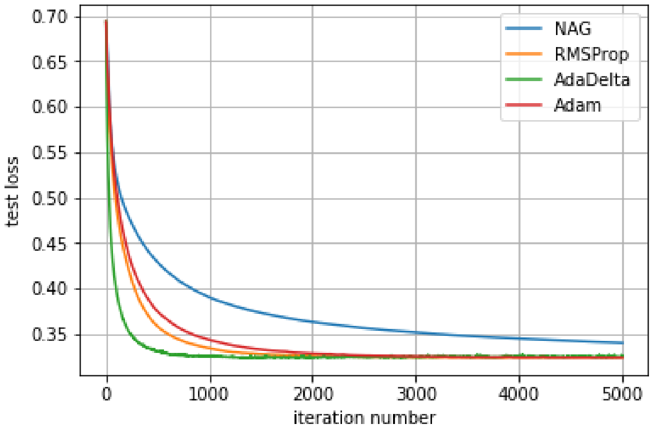

Fig. 1. The loss curves of logistic regression models on a9a.t dataset using different optimization methods.

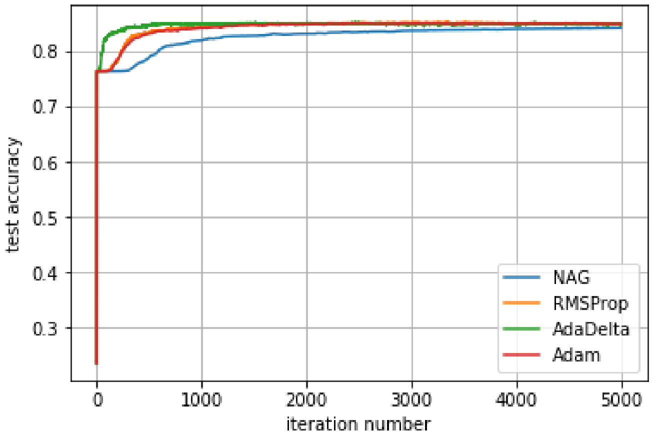

Fig. 2. The accuracy curves of logistic regression models on a9a.t dataset using different optimization methods.

IV. Conclusion

In the experiment, we implement logistic regression model and linear SVM model using four different optimization methods. With fixed number of iterations and learning rate, we have an intuitive understanding of NAG, RMSProp, AdaDelta and Adam by comparing their performances. As a result, AdaDelta converges the fastest and has a better stability than NAG.

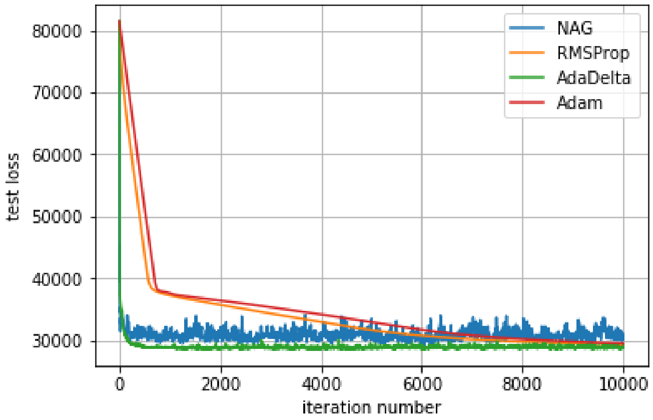

Fig. 3. The loss curves of SVM models on a9a.t dataset using different optimization methods.

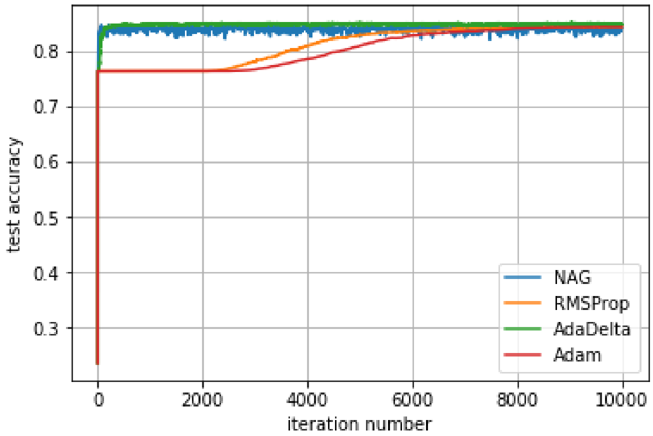

Fig. 4. The accuracy curves of SVM models on a9a.t dataset using different optimization methods.