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: Zequan Zeng	Supervisor: Qingyao Wu
Student ID：201720144917	Grade: Graduate

December 14, 2017

Logistic Regression, Linear Classification and Stochastic Gradient Descent

Abstract—Regression and Classification are two basic techniques of machine learning.In this report,We will introduce Logistic Regression and Linear Classification .And we will update two models parameters using Stochastic Gradient Descent.Among our experiments, we update two models parameters using four optimized methods(NAG, RMSProp, AdaDelta and Adam) and compare the effect of them.

I. Introduction

In this report, we will introduce Logistic Regression and Linear Classification.We will use support vector machine(SVM) to solve Linear Classification in order to further understand the principles of SVM. And we will update two models parameters using Stochastic Gradient Descent.

Among our experiments, We will use four optimized methods(NAG, RMSProp, AdaDelta and Adam) to update the model parameters and compare the effect of them.

II. Methods and Theory

1. Logistic Regression

Logistic Regression model can be expressed as

$$\mathrm{P}(\mathrm{y_i}=1\,|\,\mathrm{X_i})=\frac{\mathrm{e}^{\mathrm{W}^{\mathrm{T}}X_i}}{1+\mathrm{e}^{\mathrm{W}^{\mathrm{T}}X_i}}$$

or

$$P(y_i=1\,|\,X_i)=\frac{1}{1+\mathrm{e}^{-\mathrm{W}^{\mathrm{T}}X_i}}$$

where $i=1,2,...,n$, $X_i\in R^m$ is a vector of the decision variables , $W\in R^m$ is a vector of the weights and $^{\mathrm{T}}$ denotes the transpose,so that $\mathrm{W}^{\mathrm{T}}X_i$ is the inner product between vectors $\mathrm{X_i}$ and W .

Assume that the data is binary as follow:

$$D=\{(x_1,y_1=\pm1),...,(x_n,y_n=\pm1)\}$$

then we can use the cross entropy error to measure the loss of the model as follow:

$$J(W)=\frac{1}{n}\sum_{i=1}^{n}\log(1+e^{-y_i w^T x_i})$$

We can also add regularization to avoid overfitting as follow:

$$J(W)=\frac{1}{n}\sum_{i=1}^{n}\log(1+e^{-y_i w^T x_i})+\frac{\lambda}{2}\|W\|_2^2$$

Assume that the data is binary as follow:

$$D=\{(x_1,y_1\in\{0,1\}),...,(x_n,y_n\in\{0,1\})\}$$

then the loss function of the model can be expressed as

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

$$h_w(x_i)=g(W^TX)=\frac{1}{1+e^{-W^TX}}$$

2. Linear Classification

Linear Classification model can be expressed as

$$\mathrm{y_i}=w_1x_{i1}+w_2x_{i2}+...+w_mx_{im}+b_i=X_i^TW+b_i$$

where $i=1,2,...,n$,and $^{\mathrm{T}}$ denotes the transpose,so that X_i^TW is the inner product between vectors $\mathrm{X_i}$ and W .

We can also stack these n equations together and get a vector form as

$$\mathrm{Y}=X^TW$$

where

$$\mathrm{Y}=\begin{pmatrix}y_1\\y_2\\...\\y_n\end{pmatrix},\mathrm{X}=\begin{pmatrix}x_{11},x_{21},...,x_{n1}\\x_{12},x_{22},...,x_{n2}\\....................\\x_{1m},x_{2m},...,x_{nm}\end{pmatrix},\mathrm{W}=\begin{pmatrix}w_1\\w_2\\...\\w_m\end{pmatrix}$$

Assume that $\mathrm{y}_i\in\{-1,1\}$, we will set a threshold that if the predictive value is higher than the threshold , then we set the classification result to 1,and if the predictive value is lower than the threshold , then we set the classification result to -1.

We use the Support Vector Machine (SVM) model to solve Linear Classification.Then the loss function of the model takes the form

$$\frac{\|\mathrm{W}\|^2}{2}+\frac{\mathrm{C}}{n}\sum_{i=1}^{n}\max(0,1-y_i(W^Tx_i+b))$$

where W is the weight vector of the decision variables, $\mathrm{y_i}$ is the real value of the sample i, $\mathrm{x}_i\in\mathrm{R}^{\mathrm{m}}$ is the decision variables, b is the bias, C is a hyper parameter.

3. Stochastic Gradient Descent

Gradient descent is a first-order iterative optimization algorithm for finding the minimum of a function.We use the expression $\frac{\partial \mathrm{J}(W_{\mathrm{x}_i})}{\partial W_{x_i}}$ to calculate the gradient each iteration.

Gradient descent update the gradient using the whole samples. Different from Gradient descent , Stochastic Gradient Descent only update the gradient using part of the samples.

4. NAG

NAG(Nesterov accelerated gradient) can be expressed as

$$g_t \leftarrow \nabla J(W_{t-1} - \gamma V_{t-1})$$
$$V_t \leftarrow \gamma V_{t-1} + \eta g_t$$
$$W_t \leftarrow W_{t-1} - V_t$$

where $\nabla J(W_{t-1} - \gamma V_{t-1}) = \frac{\partial J(W_{t-1} - \gamma V_{t-1})}{\partial (W_{t-1} - \gamma V_{t-1})}$ calculate the gradient using $W_{t-1} - \gamma V_{t-1}$ instead of W_{t-1}

5. RMSProp

RMSProp can be expressed as

$$g_t \leftarrow \nabla J(W_{t-1})$$
$$G_t \leftarrow \gamma G_t + (1-\gamma) g_t^{\ 2}$$
$$W_t \leftarrow W_{t-1} - \frac{\eta}{\sqrt{G_t + \varepsilon}} g_t$$

6. AdaDelta

AdaDelta is a method that don’t initialize the learning-rate,it can be expressed as

$$g_t \leftarrow \nabla J(W_{t-1})$$
$$G_t \leftarrow \gamma G_{t\text{-}1} + (1-\gamma) g_t^{\ 2}$$
$$\Delta W_t \leftarrow -\frac{\sqrt{\Delta_{t-1} + \varepsilon}}{\sqrt{G_t + \varepsilon}} g_t$$
$$W_t \leftarrow W_{t-1} + \Delta W_t$$
$$\Delta_t \leftarrow \gamma \Delta_{t-1} + (1-\gamma) \Delta W_t^2$$

7. Adam

Adam can be expressed as

$$g_t = \nabla J(W_{t-1})$$
$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$$
$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$$
$$\hat{m}_t = m_t / (1-\beta_1^t)$$
$$\hat{v}_t = v_t / (1-\beta_2^t)$$
$$W_t = W_{t-1} - \alpha \hat{m}_t / (\sqrt{\hat{v}_t} + \varepsilon)$$

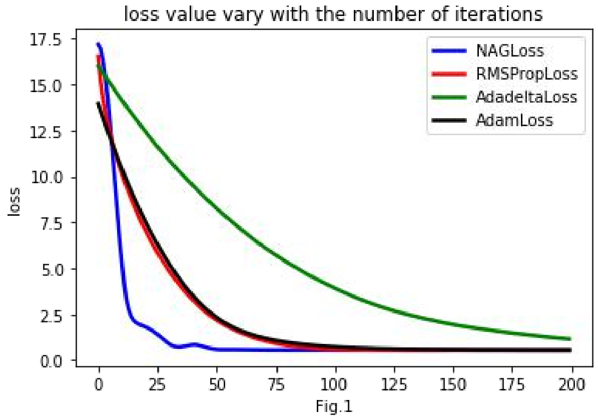
III. Experiment

1. Dataset

Both Logistic Regression and Linear Classification uses the training edition and testing edition of a9a in LIBSVM Data,including 32561 train samples and 16281 test samples and each sample has 124 features.

2. Experiment of Logistic Regression

We set the learning-rate of NAG and RMSProp to 0.01, the γ of NAG and RMSProp to 0.9, the ε of RMSProp to 10^{-8} ,the γ of AdaDelta to 0.95, the ε of AdaDelta and Adam to 10^{-6} , the β_1 of Adam to 0.9, the β_2 of Adam to 0.999, the α of Adam to 0.01, the iterations to 200,and initialize both model parameters to zeros.The Fig.1 shows that test-loss vary with the number of iterations using the four optimized method.



From the Fig.1,we can find that both four optimized method can minimize the model.It is shown that NAG has the fastest convergence and AdaDelta has the slowest convergence.

3. Experiment of Linear Classification

We set the learning-rate of NAG and RMSProp to 0.01, the γ of NAG and RMSProp to 0.9, the ε of RMSProp to 10^{-8} ,the γ of AdaDelta to 0.95, the ε of AdaDelta and Adam to 10^{-6} , the β_1 of Adam to 0.9, the β_2 of Adam to 0.999, the α of Adam to 0.01, the iterations to 200,and initialize both model parameters to zeros.The Fig.1 shows that test-loss vary with the number of iterations using the four optimized method.

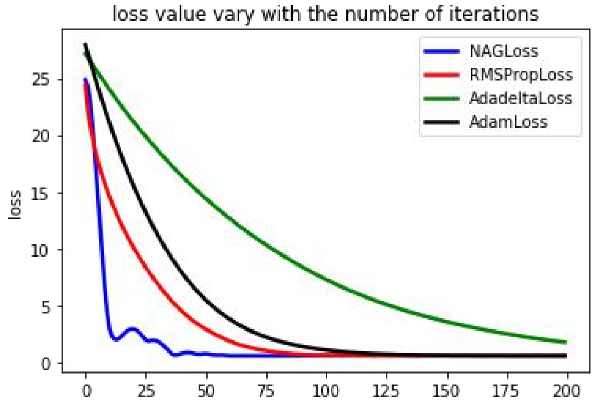


Fig.2

From the Fig.2, we can find that both four optimized method can minimize the model.It is shown that NAG has the fastest convergence and AdaDelta has the slowest convergence.

IV. CONCLUSION

Among our experiments, we can find that four optimized methods are the good algorithms to minimize the loss function, when we use the cross entropy error as the loss function of Logistic Regression and use SVM to solve the Linear Classification.

It is also shown that NAG has the fastest convergence and AdaDelta has the slowest convergence both in Logistic Regression and Linear Classification.