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:
Shuai Zou

Supervisor:
Qingyao Wu

Student ID：
201720145167

Grade:
graduate

December 14, 2017

Logic Regression, Linear Classification and Stochastic Gradient Descent

Abstract—The experiments contains Logic Regression and Linear Classification using Stochastic Gradient Descent to adjust parameters. In the process of SGD, we use different optimized methods to update the parameters. Through the experiments, we should understand logic regression and stochastic gradient descent deeper and then realize the different process of optimization and adjusting parameters.

I. Introduction

The experiments are based on python3 and use the following packages which include sklearn,numpy,jupyter and matplotlib.

We use the function train_test_split to split the datasets into training set and validation set randomly. We use the train set to adjust the parameters through SGD and then use the model we get to validate on our validate dataset. Finally we get the loss on the validate datasets according to the iterations by using different optimized methods and show them on the figure1 and 2.

II. Methods and Theory

In the experiment of Logic Regression, we first initial a linear model by setting all parameters into zero. The linear model is shown as $\mathrm{f(x)} = \theta * x + b$.

And we define a sigmod function as

$$\mathrm{sigmoid(x)} = 1/(1 + e^{-f(x)})$$

Then we choose the loss function

$\mathrm{Loss} = (-1)*(\mathrm{y} * log^{sigmoid(x)} + (1 - \mathrm{y})*log^{(1-sigmoid(x))})$

to calculate gradient G from all samples. Choose the opposite direction of gradient G as D to update model

$$\theta_t = \theta_{t-1} - \alpha * \mathrm{D}$$

where α stands for the learning rate. The D can be calculated as

$$\frac{\partial L}{\partial \theta} = (f(x) - y) * x$$

$$\frac{\partial L}{\partial b} = (f(x) - y)$$

We get the loss L_{train} under the training set and $L_{validate}$ by validating under validation set.

Finally we repeat the updating progress of θ and draw the corresponding loss L_{train} , $L_{validate}$ according to the iterations.

In the experiment of Linear Classification which is also called SVM, the loss function is defined as

$$Loss = \frac{\|\theta\|^2}{2} + C * max(0, 1 - y(\theta * x + b)).$$

The D can be calculated as

$$\frac{\partial L}{\partial \theta} = \begin{cases} 0 & 1 - y * f(x) \le 0 \\ -c * x * y & 1 - y * f(x) > 0 \end{cases}$$

$$\frac{\partial L}{\partial b} = \begin{cases} 0 & 1 - y * f(x) \le 0 \\ -c * y & 1 - y * f(x) > 0 \end{cases}$$

We use four different optimized methods NAG, RMSProp, AdaDelta and Adam. They all have different methods to update the learning rate.

NAG:

$$g_t = \nabla J\ (\theta_{\mathrm{t\text{-}1}} - \gamma v_{t-1})$$

$$v_t = \gamma v_{t-1} + \alpha g_t$$

$$\theta_t = \theta_{t-1} - v_t$$

γ=0.95 $\varepsilon = 1\mathrm{e} - 6$

RMSProp:

$$g_t = \nabla J\ (\theta_{\mathrm{t-1}})$$

$$G_t = \gamma G_t + (1 - \gamma) g_t \odot g_t$$

$$\theta_t = \theta_{t-1} - \frac{\alpha}{\sqrt{G_t + \varepsilon}} * g_t$$

γ=0.95 α=0.001 $\varepsilon = 1\mathrm{e} - 6$

AdaDelta:

$$g_t = \nabla J\ (\theta_{\mathrm{t-1}})$$

$$G_t = \gamma G_t + (1 - \gamma) g_t \odot g_t$$

$$\Delta\theta_t = -\frac{\sqrt{\Delta_{t-1} + \varepsilon}}{\sqrt{G_t + \varepsilon}} \odot \mathrm{g}_t$$

$$\theta_t = \theta_{t-1} + \Delta\theta_t$$

$$\Delta_t = \gamma \Delta_{t-1} + (1 - \gamma)\ \Delta\theta_t \odot \Delta\theta_t$$

γ=0.95 $\varepsilon = 1\mathrm{e} - 6$

Adam:

$$g_t = \nabla J\ (\theta_{\mathrm{t-1}})$$

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$$

$$G_t = \gamma G_t + (1 - \gamma) g_t \odot g_t$$

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

$$\theta_t = \theta_{t-1} - \alpha\ m_t / \sqrt{G_t + \varepsilon}$$

β_1=0.9 γ=0.999 $\varepsilon = 1\mathrm{e} - 8$

III. Experiment

A. Dataset

The a9a in LIBSVM Data which contains 32561train samples with 123 features is used in the experiment of Logic Regression.

The a9a.t in LIBSVM Data which contains 16281 samples with 123 features is used in the experiment of Linear classification.

B. Implementation

- Logic Regression

  We have a 200 iteration and the learning rate is defined as 0.001 and the bias is set to 0.1. All θ are set to 0 in the beginning.

  The Loss on train and validate is shown below:

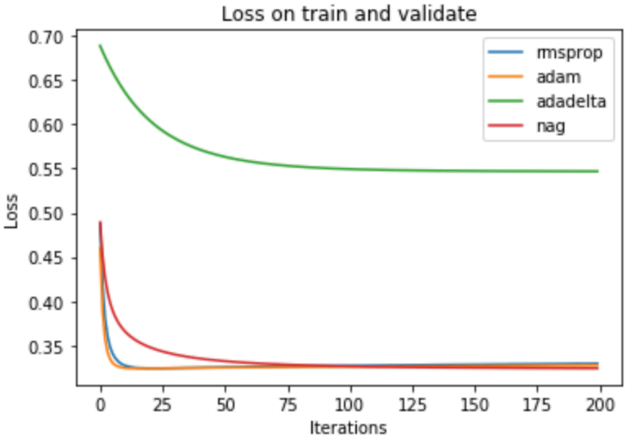

Figure 1

- Linear Classification

  We have a 100 iteration. The learning rate is defined as 0.001 for RMSProp and 0.0001 for Adam and 0.1 for AdaDelta and 0.05 for Nag and the bias is set to 0.01.c is set to 0.01. All θ are set to 0 in the beginning.

  The Loss on train and validate is shown below:

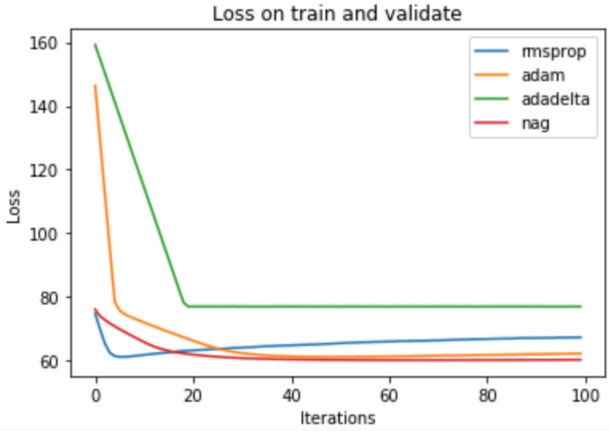

Figure 2

IV. Conclusion

In the experiments, there are some traps which will lead to some unexpected result. First of all, the function that we use to load dataset load_svmlight_file must specify the dimension. For the a9a dataset, it has 123 features but if you use this function to read a9a.t, you will only get a test dataset with 122 features. Secondly, the label in logic regression should be binary and they must be 0 and 1. So we should set all the -1 in y to 0 otherwise it will be wrong. In the experiments, the importance of hyper parameters is obviously to see and we should realize this and change the parameters in order to get a good look of the figure of the loss.