华南理工大学

South China University of Technology

The Experiment Report of Machine Learning

SCHOOL: SCHOOL OF SOFTWARE ENGINEERING

SUBJECT: SOFTWARE ENGINEERING

Author:
Shumin Yan

Supervisor:
Mingkui Tan

Student ID：
201530613337

Grade:
Undergraduate

December 14, 2017

Logistic Regression, Linear Classification and Stochastic Gradient Descent

Abstract—In this experiment, we implement logistic regression and SVM using four optimized methods which includes NAG，RMSProp，AdaDelta and Adam. And camparing SVM and logistic regression for solving classfication problems.

I. Introduction

The experiment's purpose is to compare and understand the difference between gradient descent and stochastic gradient descent, compare and understand the differences and relationships between Logistic regression and linear classification, further understand the principles of SVM and practice on larger data.

In the experiment we need implement logistic regression and SVM using four optimized methods which includes NAG，RMSProp，AdaDelta and Adam.

II. Methods and Theory

As for logistic regression,

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

Logistic regression's loss function is

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

For a sample, the gradient function is

$$\frac{\partial J(w}{\partial w} = -\frac{1}{\partial w}\cdot \partial[y\cdot \log h_w(x) + (1-y)\log(1-h_w(x))]$$
$$= -\mathrm{y}\cdot\frac{1}{h_w(x)}\cdot\frac{\partial h_w(x)}{\partial w} + (1-y)\cdot\frac{1}{1-h_w(x)}\frac{\partial h_w(x)}{\partial w}$$
$$= -\mathrm{y}\cdot\frac{1}{h_w(x)}\cdot\frac{\partial g(w^T x)}{\partial w} + (1-y)\cdot\frac{1}{1-h_w(x)}\frac{\partial g(w^T x)}{\partial w}$$
$$= \left(-\frac{xy}{h_w(x)} + \frac{x(1-y)}{1-h_w(x)}\right)\cdot \mathrm{g}(w^T\mathrm{x})\cdot[1-\mathrm{g}(w^T x)]$$
$$= (h_w(\mathrm{x}) - \mathrm{y})\mathrm{x}$$

As for SVM,
Its loss function is

$$\mathrm{L} = \max(0, 1 - y_i(w^T x^i + \mathrm{b}))$$

Optimization:

$$\mathrm{f} = \frac{||w||^2}{2} + C\sum_{i=1}^{N}\max(0, 1 - y_{i(}(w^T x_i + b))$$

Derivative:

$$g_w(x_i) = \begin{cases} -y_i x_i & 1 - y_i(w^T x_i + b) \ge 0 \\ 0 & 1 - y_i(w^T x_i + b) < 0 \end{cases}$$
$$g_b(x_i) = \begin{cases} -y_i & 1 - y_i(w^T x_i + b) \ge 0 \\ 0 & 1 - y_i(w^T x_i + b) < 0 \end{cases}$$
$$\frac{\partial f(w,b)}{w} = w + C\sum_{i=1}^{N} g_w(x_i)$$
$$\frac{\partial f(w,b)}{b} = C\sum_{i=1}^{N} g_b(x_i)$$

The four optimized methods includes NAG,RMSProp, AdaDelte, and Adam.

NAG

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

RMSProp

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

AdaDelte

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

Adam

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

III. EXPERIMENT

A.Dataset

Experiment uses a9a of LIBSVMData. The trainning dataset has 32561 samples and each sample has 123 features. The validation dataset has 16281 samples and each has 123 features.

B.Implementation

Logistic regression

Logistic regression's experimental steps:

1. Load the training set and validation set.
2. Initalize logistic regression model parameters, you can consider initalizing 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 L_{NAG} ， $L_{RMSProp}$， $L_{AdaDelta}$ and L_{Adam}.
7. Repeate 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.

Hyper-parameter selection:

SGD:

$$\eta = 0.1$$

NAG:

$$\eta = 0.01$$
$$\gamma = 0.9$$

RMSProp:

$$\eta = 0.01$$
$$\gamma = 0.9$$
$$\epsilon = 1\mathrm{e}-8$$

AdaDelte:

$$\gamma = 0.9$$
$$\epsilon = 1\mathrm{e}-8$$

Adam:

$$\eta = 0.001$$
$$\gamma = 0.999$$
$$\epsilon = 1\mathrm{e}-8$$
$$\beta_1 = 0.9$$

Predicted Results (Best Results):

Accuracy rate:

SGD: 0.832

NAG: 0.812

RMSProp: 0.807

AdaDelte: 0.764

Adam: 0.764

Loss value:

SGD:0.378

NAG:0.391

RMSProp:0.398

AdaDelte:0.523

Adam:0.621

Loss curve:

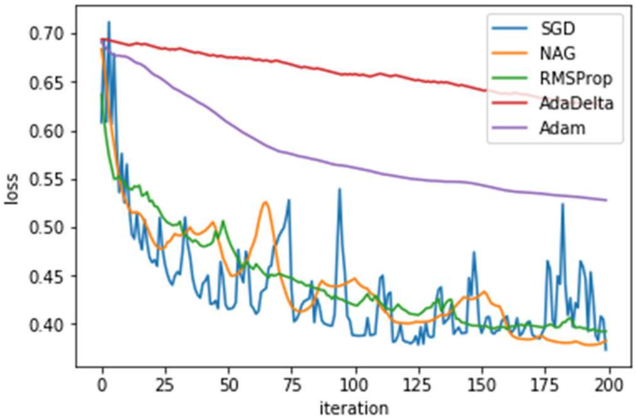

Fig.1 Logistic Regression's loss curve

SVM

SVM's experimental steps:

1. Load the training set and validation set.
2. Initalize SVM model parameters, you can consider initalizing 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 L_{NAG} ， $L_{RMSProp}$， $L_{AdaDelta}$ and L_{Adam}..
7. Repeate 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.

Hyper-parameter selection:

SGD:

$$\eta = 0.003$$
$$\mathrm{C} = 0.1$$

NAG:

$$\mathrm{C} = 0.1$$
$$\eta = 0.01$$
$$\gamma = 0.9$$

RMSProp:

$$C = 0.1$$
$$\eta = 0.001$$
$$\gamma = 0.9$$
$$\epsilon = 1e - 8$$

AdaDelte:

$$C = 0.1$$
$$\gamma = 0.95$$
$$\epsilon = 1e - 8$$

Adam:

$$C = 0.1$$
$$\eta = 0.001$$
$$\gamma = 0.999$$
$$\epsilon = 1e - 8$$
$$\beta_1 = 0.9$$

Predicted Results (Best Results):
Loss value:
SGD: 0.048
NAG:0.058
RMSProp:0.051
AdaDelte: 0.052
Adam: 0.047

Loss curve:

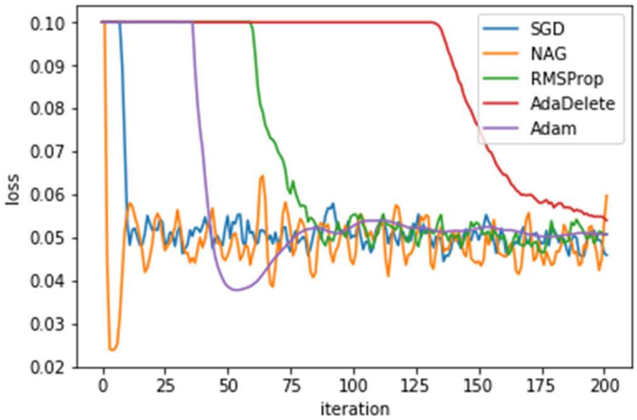

Fig.2 SVM's loss curve

IV. CONCLUSION

After the experiment, I know the differences and relations of the gradient descent, stochastic gradient descent, and batchGradient descent.I implement logic regression and linear classification, and understand thier principle and their relationships. In addition, through the optimization methods, I understand what is NAG, RMSProp, AdaDelta and Adam and understand how to use them.

In conclusion, logistic regression is different from SVM. First, the mothed of finding the optimal hyperplane is different. Logistic regression find the hyperplane, to let all point away from it.The hyperplane which SVM looks for, is the most close to the middle line. Second, the SVM can handle the nonlinear case.Third, their loss function is different.The loss of the Logistic regression function is cross entropy loss,and the SVM is the hinge loss.