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:
Riyang Hu

Supervisor:
Mingkui Tan

Student ID：
201530611616

Grade:
Undergraduate

December 14, 2017

Linear Regression, Linear Classification and Stochastic Gradient Descent

Abstract—In this experiment, we would compare the difference between NAG, RMSProp, AdaDelta and Adam algorithm, these four different optimization methods, under the implementations of logistic regression and SVM.

I. Introduction

Gradient descent method is an optimization algorithm, usually called the steepest descent method. The steepest descent method is one of the simplest and oldest methods for solving unconstrained optimization problems. Although it is not practical now, many effective algorithms are based on it and are improved and corrected. The steepest descent method is to use the negative gradient direction for the search direction, steepest descent method closer to the target value, the smaller the step, the slower the progress. We aim to compare NAG, RMSProp, AdaDelta and Adam and use these four methods to update the models of logistic regression and SVM.

II. Methods and Theory

A. Logistic regression

In statistics, logistic regression is a regression model where the dependent variable (DV) is categorical. This article covers the case of a binary dependent variable—that is, where the output can take only two values, "0" and "1", which represent outcomes such as pass/fail, win/lose, alive/dead or healthy/sick. Cases where the dependent variable has more than two outcome categories may be analyzed in multinomial logistic regression, or, if the multiple categories are ordered, in ordinal logistic regression. In the terminology of economics, logistic regression is an example of a qualitative response/discrete choice model.

In this experiment, the labels are binary. And for all sample, the log-likelihood loss function is

$$\mathrm{L}(\beta)=\sum_{\mathrm{i}=1}^{\mathrm{m}}\left(-\mathrm{y_i}\beta^{\mathrm{T}}\mathbf{x}_{\mathrm{i}}+\ln\quad\left(1+\mathrm{e}^{\beta^{\mathrm{T}}\mathbf{x}_{\mathbf{i}}}\right)\right)$$

And the gradient of loss function is:

$$\frac{\partial\mathrm{L}(\beta)}{\partial\beta}=-\sum_{\mathrm{i}=1}^{\mathrm{m}}\mathbf{x}_{\mathbf{i}}(\mathrm{y_i}-\mathrm{p1}(\mathbf{x}_{\mathbf{i}};\beta))$$

B. linear SVM

In machine learning, SVM is supervised learning models with learning algorithms that analyze data used for classification and regression.

In this experiment, we simply focus on the binary case,and for all sample ,the loss function is:

$$\frac{1}{2}\|\mathrm{w}\|^2+\mathrm{C}\sum_{\mathrm{i}=1}^{\mathrm{m}}\max\quad(0,1-\mathrm{y_i}(\mathrm{w}^{\mathrm{T}}\mathbf{x}_{\mathbf{i}}+\mathrm{b}))$$

and the gradient of loss function is:

$$\frac{\partial\mathrm{L}}{\partial\mathrm{w}}=\begin{cases}\mathrm{w}^{\mathrm{T}}-\mathrm{C}\mathrm{X}^{\mathrm{T}}\mathrm{y},\ \text{if}\ 1-(\mathrm{y_i}(\mathrm{w}^{\mathrm{T}}\mathrm{x_i}+\mathrm{b})\geq 0\\ \mathrm{w}^{\mathrm{T}},\ \text{else}\end{cases}$$

C. optimization methods

NAG:

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

RMSProp:

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

AdaDelta:

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

Adam:

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

III. EXPERIMENT

A. Dataset

Experiment uses a9a of LIBSVM Data, including 32561/16281 samples and each sample has 123 features,.The label is -1 or 1.

B. Implementation

Logistic regression:

1. Load the training set and validation set.
2. Initalize logistic regression model parameters, you can consider initalizing zeros, random numbers or normal distribution.
3. Select the loss function and calculate its derivation.
4. Calculate gradient G 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 function.
7. Repeate step 4 to 6 for several times, and drawing graph of loss function and with the number of iterations.

Linear classification

1. Load the training set and validation set.
2. Initalize SVM model parameters, you can consider initalizing zeros, random numbers or normal distribution.
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 function.
7. Repeate step 4 to 6 for several times, and drawing graph of loss function and with the number of iterations..

Table 1
The value use in logistic regression

NAG	$\gamma = 0.9$, epoch $= 200$, $\eta = 0.05$
RMSProp	$\gamma = 0.9$, epoch $= 200$, $\eta = 0.005$,$\varepsilon = 1e^{-8}$
AdaDelta	$\gamma = 0.999$, epoch $= 200$,$\varepsilon = 1e^{-8}$
Adam	$\beta = 0.9$,$\gamma = 0.999$, epoch $= 200$,$\varepsilon = 1e^{-8}$

Table 2
The value use in linear SVM

NAG	$\gamma = 0.9$, epoch $= 300$, $\eta = 0.0001$
RMSProp	$\gamma = 0.9$, epoch $= 300$, $\eta = 0.005$,$\varepsilon = 1e^{-8}$
AdaDelta	$\gamma = 0.999$, epoch $= 300$,$\varepsilon = 1e^{-8}$
Adam	$\beta = 0.9$,$\gamma = 0.999$, epoch $= 300$,$\varepsilon = 1e^{-8}$

Fig 1
The loss function of logistic regression

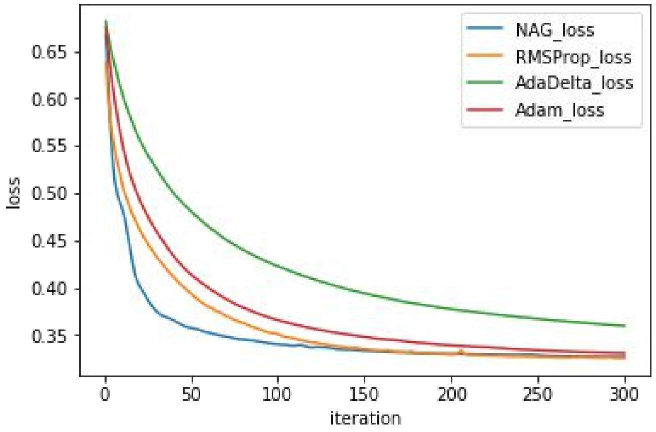
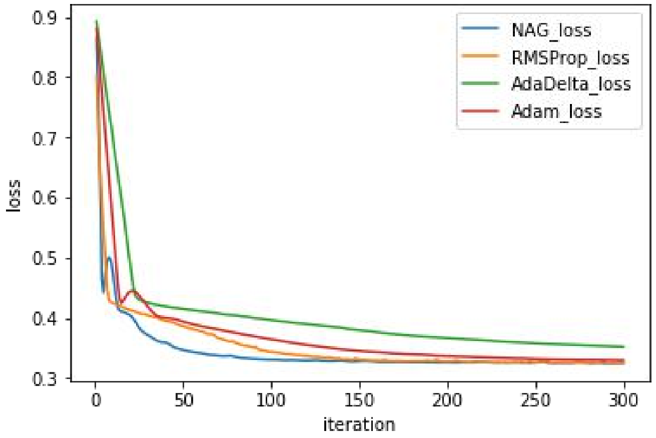


Fig 2
The loss function of SVM



IV. CONCLUSION

In this experiment, we realized logistic regression and linear classification using stochastic gradient decent and tried four different methods to optimize it.