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: Ran Yang	Supervisor: Qingyao Wu
Student ID： 201721046128	Grade: Graduate

November 14, 2017

Logistic Regression, Linear Classification and Stochastic Gradient Descent

Abstract—The experiments on logistic regression, linear classification and stochastic gradient descent used four different optimization methods to update the model parameters. By selecting suitable thresholds, testing on the validation set and getting the Loss function values of different optimization methods, we can know the Similarities and differences among NAG, RMSProp, AdaDelta and Adam.

I. Introduction

Logistic regression can be seen as a special case of the generalized linear model and thus analogous to linear regression. The model of logistic regression, however, is based on quite different assumptions (about the relationship between dependent and independent variables) from those of linear regression. In particular the key differences between these two models can be seen in the following two features of logistic regression. First, the conditional distribution {\displaystyle y\mid x} y\mid x is a Bernoulli distribution rather than a Gaussian distribution, because the dependent variable is binary. Second, the predicted values are probabilities and are therefore restricted to (0,1) through the logistic distribution function because logistic regression predicts the probability of particular outcomes.

There are three categories of linear classifiers: Perceptual criteria function, SVM, Fisher criterion. Perceptual criteria function: The cost function J = - (W * X + W0), the classification criteria is to minimize the cost function. Perceptron is the basis of neural network (NN). SVM: SVM is also a classic algorithm, the optimization goal is the maximum margin, also known as the maximum interval classifier, is a typical linear classifier. (Using kernel functions to solve nonlinear problems). Fisher's Criterion: The broader term is Linear Discriminant Analysis (LDA), which projects all samples onto a straight line starting from an origin so that the distance between samples of the same type is as small as possible and the distance between samples of different classes is as large as possible, entropy". Linear separable is a more ideal situation, the real world such data is rare. On the one hand, the actual data dimension is generally much larger than two dimensions. The more dimensions, the more complicated the data distribution, the less likely it is to be linearly separable. On the other hand, even if the target function is linear, The noise caused by the process is also very possible nonlinear.

In order to overcome the shortcoming of BGD algorithm training too slowly, the SGD algorithm is proposed. The common BGD algorithm iterates all the samples once per iteration and updates the gradient every training sample. The SGD algorithm is randomly selected from the sample group, updated by a gradient after the training, and then extract a group, and then update the sample size and its large case, you may not have to train all the samples can get a Loss of the model within the acceptable range.

The motivation of this experiment is as follows:

1.Compare and understand the difference between gradient descent and stochastic gradient descent.

2.Compare and understand the differences and relationships between Logistic regression and linear classification.

3.Further understand the principles of SVM and practice on larger data.

II. Methods and Theory

1. Logistic regression

For Logistic Regression, the idea is also based on linear regression (Logistic Regression is a generalized linear regression model). The formula is as follows:

$$\mathrm{h_w}(X) = g(W^T X) = \frac{1}{1 + e^{-W^T X}}$$

The Logistic Regression algorithm maps the result of a linear function to the sigmoid function.Then, probability can be expressed as follows:

$$\mathrm{p} = \begin{cases} h_w(X_I) & y_i = 1 \\ 1 - h_w(X_i) & y_i = 0 \end{cases}$$

Consequently, the loss function is:

$$\mathrm{J(W)} = -\frac{1}{n}\left(\sum_{i=1}^{n} y_i log h_w(X_i) + (1 - y_i)\log(1 - h_w(X_i))\right))$$

2. Linear classification

learning the SVM can be formulated as an optimization:

$$\max_{\mathrm{w,b}} \frac{2}{||\mathrm{w}||}$$

$$\mathrm{s.t.}\quad \mathrm{w^T x_i + b} \begin{cases} \geq 1 & \mathrm{y} = +1 \\ \leq 1 & \mathrm{y} = -1 \end{cases}$$

Therefore, we introduce $\xi_i \geq 0$, for each i, which represents how much example i is on wrong side of margin boundary.

If $\xi_i = 0$ then it is ok.

If $0 < \xi_i < 1$ it is correctly classified, but with a smaller margin than $\frac{1}{||w||}$

If $\xi_i > 1$ then it is in correctly classified.

The optimization problems become:

$$\min_{w,b} \frac{||w||^2}{2} + C \sum_{i=1}^{n} \xi_i$$

$$\text{s.t.} \quad y_i(w^T x_i + b) \geq 1 - \xi_i, \quad i = 1,2,\ldots,n$$

$$\text{Hinge loss} = \xi_i = \max(0, 1 - y_i(w^T x_i + b))$$

The optimization problems become

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

3. Stochastic gradient descent

In the event of a large amount of data, it is difficult not to use the Stochastic gradient descent (SGD). SGD is very intuitive, is to take a random or a few data to do a gradient decline, that is

$$\mathbf{g}_t \leftarrow \nabla J_i(\boldsymbol{\theta}_{t-1})$$
$$\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \eta \mathbf{g}_t$$

As the gradient descent method converges slowly, the stochastic gradient descent method will be much faster

- Calculate weight updates based on error increments for a single sample, resulting in approximate gradient descent searches (take a random sample)
- It can be seen as defining different error functions for each individual training example
- When iteratively iterating over all training examples, the sequence of weights updates gives a reasonable approximation of the gradient decline of the original error function
- By making the rate of descent small enough, the stochastic gradient can be decreased to any degree close to the true gradient.

III. Experiment

Experiment uses a9a of LIBSVM Data, including 32561/16281(testing) samples and each sample has 123/123 (testing) features. Please download the training set and validation set.

Logistic Regression and Stochastic Gradient Descent

1)Load the training set and validation set.

2)Initialize logistic regression model parameters, you can consider initializing zeros, random numbers or normal distribution.

3)Select the loss function and calculate its derivation, find more detail in PPT.

4)Calculate gradienttoward 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)Repeat step 4 to 6 for several times, and drawing graph of and with the number of iterations.

Linear Classification and Stochastic Gradient Descent

1)Load the training set and validation set.

2)Initialize SVM model parameters, you can consider initializing zeros, random numbers or normal distribution.

3)Select the loss function and calculate its derivation, find more detail in PPT.

4)Calculate gradienttoward 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)Repeat step 4 to 6 for several times, and drawing graph of and . with the number of iterations.

The following are the graph of experimental results.

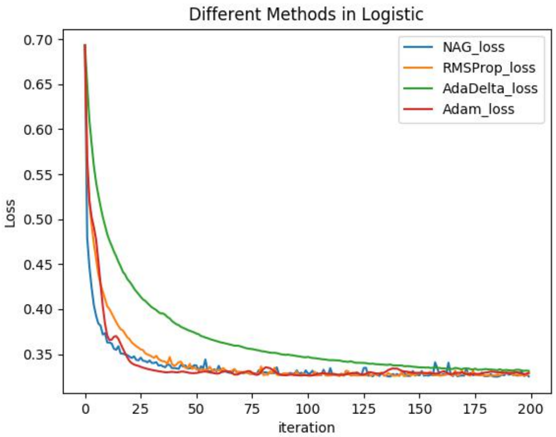

Fig.1 Different Methods in Logistic Regression

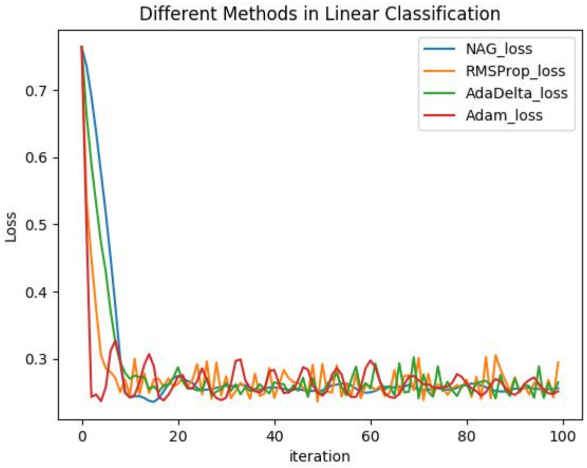

Fig.2 Different Methods in Liner Classification

IV. CONCLUSION

logistic regression and linear classification are both common classification algorithms. As for the objective function, the difference is that, the logistic regression using logistical loss, SVM using hinge loss. The purpose of these two loss functions are both to increase the data points’, that matter classification, weight and reduce the weight of the data points less relevant to the classification. SVM processing method is to consider only support vectors, which is the most relevant and the classification of a few points to learn the classifier. Logistic regression through nonlinear mapping, the weight of the points farther away from the classification plane is reduced and the weight of the data points most relevant to the classification is raised. The basic purpose of both method is the same. In addition, both methods can add different regular terms.So, in many experiments, the results of the two algorithms are very close.

Four different optimization Methods (NAG,RMSProp,AdaDelta and Adam) all have advantage and disadvantage.For NAG,it is possible to accelerate the convergence of SGD by adaptively updating parameters according to the slope of the loss function in each learning process.For AdaDelta,it fully adaptive global learning rate, good performance of acceleration but usually have shock in a small area while in late learning progress.For RMSProp,it good performance of acceleration.For Adam,it is suitable for processing large-scale data.

This experiment is about the comparison of different decent method. After implementing such different optimization methods,I have a better understand of the principle of logistic regression ,linear classification and stochastic gradient descent. Also, by implementing four different optimized methods, I realize the mechanism of NAG，RMSProp，AdaDelta and Adam.I also know that when plotting the graph of loss, using log scale with y dimension can have a better picture of decent progress,he line will be more smooth.,and it’s necessary to master the technique of tuning parameters.