SCIREA Journal of Mathematics
http://www.scirea.org/journal/Mathematics
May 18, 2022
Volume 7, Issue 2, April 2022
https://doi.org/10.54647/mathematics11319

Variable selection of regularized stochastic gradient descent in logistic regression

Ping Guo

College of Mathematics and Statistics, Guangxi Normal University, Guilin, Guangxi, China

Abstract

In the modern big data environment, Stochastic gradient descent is an important method for training neural networks, processing large scale data sets, optimization, etc. Deeply welcomed in various fields. With regard to SGD, the existing literature considers the stopping condition of parameter iteration. In fact, some unimportant parameters do not always have values of 0 during iteration, and it is not clear whether they are important or not even if the stop condition is reached. We consider variable selection of SGD parameter iteration with L1 regular in generalized linear regression model (taking Logistic regression as an example). Monte Carlo numerical simulation and practical application examples were given to illustrate the consistency of variable selection. The results show that high accuracy can be achieved by using the selected variables to build the model.

Keywords: SGD; Lasso; Logistic regression; Variable selection

1. Introduction

Stochastic Gradient Descent algorithm (SGD) in machine learning because of its large-scale

processing, the characteristics of high efficient calculation, now already the optimization problem has become the dominant generation method. SGD is widely applied in optimization control problems[1] and signal processing[2], such as training neural networks[3], data handling noise[4], but non-significant parameters are not always 0 in the iterative process in optimization problems.

On the other hand, as we know, Lasso was proposed by Tibshirani in 1996[5]. The parameter estimation of this method is the least squares estimation with constraints. In essence, some parameter values are compressed to 0 to realize model selection. However, although Lasso method can realize parameter estimation and variable selection, it is limited by linearity. It is obvious that the relationship between features may not be linear in real life. Inspired by Khalili and Chen[6], this paper combined SGD method with Lasso method under the condition of logistic regression. Monte Carlo simulation shows that this method can select important variables in the feature screening process.

The rest of this article is organized as follows. In the second stanza, we introduce some preliminary knowledge and the idea that it is not significant to judge that the iteration process is not always zero. In Section 3, Monte Carlo numerical simulations are used to demonstrate Consistency of iterative variable selection for L1 regularized SGD parameters. In Section 4, an example of a practical application is given. Finally, the main content is summarized in Section 5.

2. Model Formulation

2.1 Preliminary knowledge

We consider logistic regression model. let $(x_i, y_i), i = 1,2,\cdots,n$ denote an independent and identically distributed dataset, where $x = (x_1, x_2, \cdots, x_d) \in R^{n\times d}$ and $y = (y_1, y_2, \cdots, y_n)^T$. d is the number of covariables.

Without loss of generality, generalized linear models can be expressed as $y = h^{-1}(x^T\beta)$, where $\beta = (\beta_1, \beta_2, \cdots, \beta_d)^T$, and $h(\cdot)$ denote activation function. The activation function selected by logistic regression is sigmoid, y is the dichotomous response variable, the actual formulas as follow:

$$log\, it(y) = x^T\beta\,, \tag{1}$$

where $log\,it(y) = ln(p(y=1)/p(y=0))$.

Then, penalty log conditional likelihood function[6] of logistic regression model can be obtained:

$$l(\beta) = \sum_{i=1}^{n}\left[y_i\, log\left(\frac{1}{1+exp\{-x_i\beta\}}\right) + (1-y_i) log\left(1-\frac{1}{1+exp\{-x_i\beta\}}\right)\right] - \lambda\|\beta\|_1, \tag{2}$$

where $y_i = P(y=1)$. Similar to Wang and Nguyen[7], the partial derivative of β_j was firstly calculated and parameter iteration was carried out using SGD method.

$$\begin{aligned}\frac{\partial}{\partial\beta_j} l(\beta) &= \sum_{i=1}^{n} y_i \frac{exp\{-x_i\beta\}}{1+exp\{-x_i\beta\}} x_{ij} + \sum_{i=1}^{n}(1-y_i)\frac{-1}{1+exp\{-x_i\beta\}} x_{ij} - \lambda sign(\beta_j) \\ &= \sum_{i=1}^{n}\left(y_i - \frac{1}{1+exp\{-x_i\beta\}}\right) x_{ij} - \lambda sign(\beta_j)\end{aligned}$$

The rule of iteration update is

$$\beta_j^{(t+1)} = \beta_j^{(t)} + \alpha\left(\sum_{i=1}^{batch}\left(y_i - \frac{1}{1+exp\{-x_i\beta\}}\right) x_{ij} - \lambda sign\left(\beta_j^{(t)}\right)\right), \tag{3}$$

where α is learning rate, batch denote number of random samples for each training. Stopping condition[7] is $\frac{1}{d}\sum_{j=1}^{d}\left|\beta_j^{(t+1)} - \beta_j^{(t)}\right| < \varepsilon$, the difference is that the threshold ε is 0.05.

2.2 An idea

The values of insignificant parameters of the SGD algorithm are not always 0 in the iterative process. On the other hand, Lasso is used under the linear assumption. Here, we present an idea inspired by probability theory.

We assume that when the iteration parameter reaches the stop condition, the total number of iterations is T. Analyze the parameter value beta obtained according to the iteration rule, and let set K denote index, where $\mathrm{K} = \left\{k : \left|\beta_k^{(T)}\right| \le 0.5\right\} \subset [d]$. Let event A denote $\beta_k^{(t)}, t = 1,2,\cdots,T$ is less than the critical value δ. Iterating over the elements of set K, let B denote the number of times that set A occurs in T iterations. The formula for B is as follows:

$$B = \sum_{t=1}^{T} I\left(\left|\beta_k^{(t)}\right| < \delta\right), \delta = 0.05, k \in \mathrm{K}, \tag{4}$$

We set β_j to 0 if the proportion of B to total times T is greater than 90%, where $Rate = B/T$. Thus, based on the idea that "small probability" events are impossible to happen, we can judge that a certain β_j is an insignificant parameter and select variables.

3. Numerical solutions

In this section, monte Carlo numerical simulation is used to illustrate the consistency of variable selection. Generated data from the following models:

$$log\,it(y) = 3x_1 - 2x_2 + 5x_3 + 0x_4 + \cdots + 0x_{100} + \varepsilon ,$$

where $log\,it(y) = ln(p(y=1)/p(y=0))$, $\varepsilon \sim N(0,1)$. Covariates and random errors are normally distributed.

Yan Sun[8] was referred to for data generation. The difference is that this paper generates a data set containing 10,000 samples. Generated dataset was shuffled, and dataset was divided into training set, test set in a ratio of 7:3. Since this is a convex optimization problem, the iterative initialization is set $\beta_j^{(0)} = 0, j = 1,2,\cdots,d$. In order to avoid the randomness of simulation results, and the same group of training data was repeatedly run for 10 times, and finally selected statistical variables were determined. With the idea that "low probability" events are unlikely to occur, we select three important variables, x_1 , x_2 ,and x_3 , where $\beta_1^{(1000)} = 2.9645$, $\beta_2^{(1000)} = -1.9271$, $\beta_3^{(1000)} = 4.7228$. According to the selected variables, the logistic regression model is established using the test set, and the relevant results are obtained. Analysis of Deviance and confusion matrix shown in Table 1 and Table 2. It was evident that accuracy value is 0.9040, sensitivity is 0.9013, specificity is 0.9067. This shows that the prediction accuracy of the model is very high.

Table 1: Analysis of Deviance

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
x_test	3	2770.7	2986	1374.2	< 2.2e-16 ***

Table 2: confusion matrix

	reference	
	1	0

prediction	1	tp=1343	fp=140
	0	fn=147	tn=1360

Receiver operating characteristic(ROC) curve shown in Figure 1. The value of Area Under Curve(AUC) is 0.9677, the modle has high prediction accuracy.

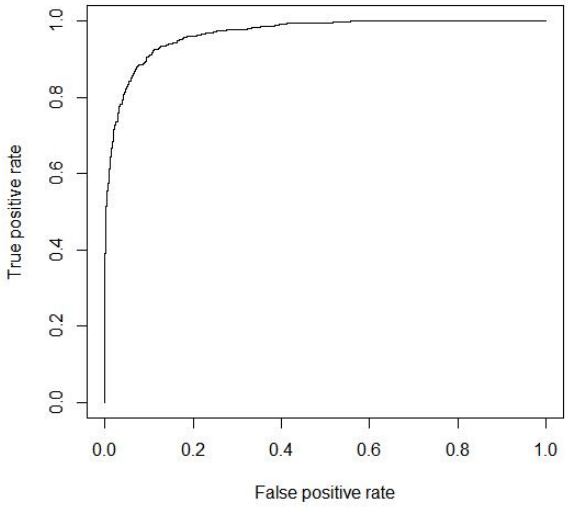

Figure 1: Receiver operating characteristic curve

4. An illustrative example

An example of practical application is given. With the rise of the Internet, three first-generation online recruitment industries were born in China in the 1990s. And Recruitment websites are booming, many enterprises and many job seekers are willing to choose online recruitment. Thus, collecting and using relevant platform recruitment information for study and research with reliability and universality of audience.

This paper adopts the recruitment data of development engineers in Beijing, Shanghai, Guangzhou and Shenzhen from a recruitment platform up to April 10, 2022, with a total of 3391 pieces of data. Take the monthly average salary of the advertised position as the independent variable, let the sample above the total average wage be 1, otherwise 0. There are 17 covariables, Job title, workplace, company attributes, Job category, Company size, Experience requirements, education requirements, social welfare, food and housing benefits, working system, vacation benefits, regular physical examination, employee travel benefits,

company culture, other benefits, employee training, where $x=(x_1,x_2,\cdots,x_{17})\in R^{3391\times 17}$.

To evaluate the model, 3,391 recruitment data were shuffled and then divided into training set, and test set in an 7:3 ratio, with 2,374 and 1,017 samples, respectively. Since the unit of measurement of each characteristic variable is not consistent, normalization is considered. The rules of normalization is

$$x'_{i,t} = \frac{x_{i,t} - \min\limits_{1\le i\le 3391}(x_{i,t})}{\max\limits_{1\le i\le 3391}(x_{i,t}) - \min\limits_{1\le i\le 3391}(x_{i,t})}, t = 1,2, \cdots ,17. \tag{5}$$

Initialize the parameter $\beta_j^{(0)}=0, j=1,2,\cdots,d$. Also apply the idea of low probability events that will not happen. After repeated debugging program, when the number of iterations reached 10000, $\frac{1}{17}\sum_{j=1}^{17}\left|\beta_j^{(10000)}-\beta_j^{(9999)}\right|=0.0136$. According to the iteration results, 11 important variables were selected. The significant parameters are as follows: $\beta_1^{(10000)}=0.7880$, $\beta_6^{(10000)}=2.4506$, $\beta_7^{(10000)}=4.6368$, $\beta_{10}^{(10000)}=-1.0925$ and so on. These variables were used to establish logistic regression model, ROC curve shown in Figure 2, the value of AUC is 0.7746 and get high predicting accuracy.

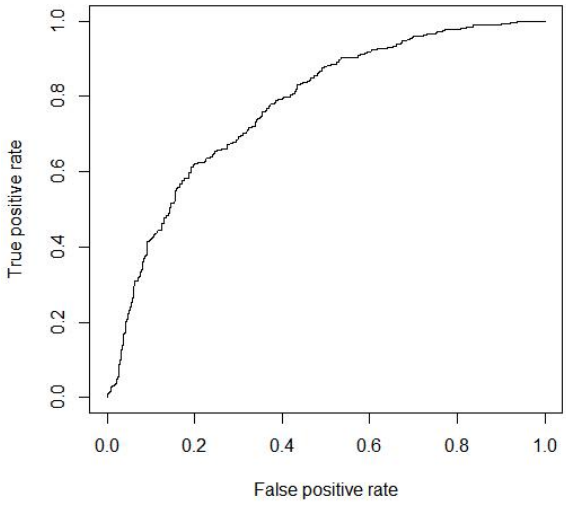

Figure 2: ROC curve

5. Conclusion

Stochastic gradient algorithm has become an indispensable tool for machine learning. In this paper, we propose an idea that judgment parameters are not significant based on Khalili and Chen's punishment likelihood framework. Monte Carlo simulation verifies that the results of

variable selection of this idea are consistent and perform well in the predictive analysis of logistic regression model. Future work can be extended from generalized linear model to finite mixed regression model, to determine insignificant parameters as soon as possible and reduce the amount of calculation.

References

[1] Kushner, H. and Yin, G.(1997). Stochastic Approximation Algorithms and Applications. Springer Verlag, New York.

[2] Nemirovski, A., Juditsky, A., Lan, G.H., et al(2009). Robust stochastic approximation approach to stochastic programming. SIAM Journal on Optimization, 19:1574–1609.

[3] Du, S.S., Zhai, X.Y., Póczos, B., et al(2018). Gradient descent provably optimizes over parameterized neural networks. Statistics, 1467-5463.

[4] Bottou,L. and Bousquet, O.(2007). The tradeoffs of large scale learning. In Advances in Neural Information Processing Systems (NeurIPS), 161–168.

[5] Tibshirani, R.(1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B:Methodological, 58(1):267-288.

[6] Khalili, A., and Chen, J.H.(2007). Variable Selection in Finite Mixture of Regression Models. Journal of the American Statistical Association, 102(479): 1025-1038.

[7] Wang, P.Q. and Nguyen, P.X.(2012). Variations of Logistic Regression with Stochastic Gradient Descent.

[8] Sun,Y., Song Q.F.,and Liang, F.M.(2021). Consistent Sparse Deep Learning: Theory and Computation. Journal of the American Statistical Association, 1-42.