Comparison of different learning rate (step size) on Logistic regression using FR conjugate gradient optimizer

Dada Ibidapo D.[1], Akinwale Adio T.[1], Onashoga Adebukola S.[1] and Osinuga Idowu A.[2]
[1]Department of Computer Science, Federal University of Agriculture, Abeokuta, Nigeria.
[2]Department of Mathematics, Federal University of Agriculture, Abeokuta, Nigeria.
Email: dman4computer@gmail.com

Abstract- Conjugate gradient algorithm is one of the effective optimization algorithms used in solving logistic regression problems. This paper is focused on comparing some existing learning rate methods to reduce the objective function value of the logistic regression model with a limited number of iterations and reduced processing time. Fletcher-Reeves (FR) conjugate gradient method was run in python program using admission and iris flowers dataset to examine the performance of each learning rate. The numerical results of each step size were compared. The result shows that Armijo step size performs better in terms of number of iterations and processing time with good model accuracy.

Keywords- logistic regression; conjugate gradient method; step size

I. INTRODUCTION

Logistic regression is a supervised machine learning-based binary classification algorithm. The model is commonly used in tasks like recommender systems, click rate estimation (CTR) and Computational ads (Yuan *et al.*, 2019). The logistic regression algorithm's key concept is to non-linearize the multiple linear regression equation using the logistic function, i.e. the sigmoid function, to be able to reap the impact of data classification and model generalization. Minimizing error in the optimal parameters of the objective function makes the logistic regression classification as correct as possible. Objective function of logistic regression can be constructed as a nonlinear unconstrained minimization problem which can be solved using the conjugate gradient approach. As long as the current iterate point is not a fixed point, the gradient method search along the negative gradient function will ensure that the objective function is reduced (Yuan, 2008). Many researches have been conducted in the hopes of discovering a better and more appropriate search direction method that will have an effect on minimizing objective functions.

LITERATURE REVIEW

A. LOGISTIC REGRESSION

Logistic regression is a machine learning classification algorithm that is used to predict the chance of a categorical variable. The model is employed to model the probability of a category like pass/fail or win/lose. This could be extended to model many classes of events like determining whether or not an image contains a goat, cat, lion and so on. The logistic regression model is used in a lot of fields such as statistics, mathematics, machine learning, medical fields, social sciences and so on.

The simple logistic regression can be modeled as:

i. The outputs is always either 0 or 1
ii. Hypothesis: Z = wx + B (1)
iii. $h\theta(x) = sigmoid(Z)$ (2)
iv. sigmoid (Z) = $\frac{1}{1+e^{-Z}}$ (3)

B. COST FUNCTION OF LOGISTIC REGRESSION MODEL

The cost function of the logistic regression is called the logistic loss.

$$cost\left(h\theta(x^i), y^i\right) = \begin{cases} -log\left(h\theta(x)\right) & if \quad y = 1 \\ -log\left(1 - h\theta(x)\right) & if \quad y = 0 \end{cases} \tag{4}$$

The cost function of the logistic regression is the summation from all training data samples:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} cost\left(h\theta(x^i), y^i\right) \tag{5}$$

$$J(\theta) = \frac{1}{m}\left[\sum_{i=1}^{m} -y^i log\left(h\theta(x^i)\right) + (1 - y^i) log\left(1 - h\theta(x^i)\right)\right] \tag{6}$$

C. THE FLETCHER-REEVES (FR) CONJUGATE GRADIENT METHOD

Conjugate Gradient Method (CGM) can solve both linear and nonlinear optimization problems (Yu-Hong, 2010).

Unconstrained optimization problem can be modeled as:

$$min\{f(x) | x \in R^n\} \tag{7}$$

where $f: R^n \to R$ is continuously differentiable, $f(x)$ is an objective function and $x \in R^n$ is a vector with independent variables. The Conjugate Gradient Methods are usually solved using an iterative approach which is defined as follows:

$$x_k = x_{k-1} + \propto_{k-1} d_{k-1}, \qquad k = 1,2,3, \dots \tag{8}$$

where x_{k-1} is the present iterative point, $\propto_{k-1}$ is the learning rate and d_k is the search direction of conjugate gradient method. d_k can be defined as follows:

$$d_k = \begin{cases} -g_k & k=0 \\ -g_k + \beta_k d_{k-1} & k=1,2,\dots \end{cases} \tag{9}$$

where g_k is the gradient at point x_k. β_k is FR conjugate gradient (CG) coefficient of $f(x)$ which is given as follows:

$$\beta_k{}^{FR} = \frac{g_k{}^T g_k}{\|g_{k-1}\|^2} \tag{10}$$

$\beta_k \in R$ is a scalar while $g_k = \nabla f(x_k)$ at point x_k.

FR conjugate gradient (CG) Algorithm

1: Set initial point $x_0 \in \mathbb{R}^n$, $k = 0$.
2: Compute β_k based on $\beta_k{}^{FR}$ as (10).
3: Compute d_k as (9).
If $\|g_k\| = 0$, then stop, otherwise go to step 4.
4: Compute step size $\propto_k$.
5: Update a new point by (8)
6: Stopping criteria.
If $f(x+1) < f(x)$ and $\|g_k\| < \epsilon$, then stop.
else goto step 1, then set k = k + 1.

III. LEARNING RATE (STEP SIZE)

The aim of any CGM is to find the minimum value of an unconstrained function (Yuan *et al.*, 2019, Hamoda *et al.*, 2015). The learning rate plays a great part in minimizing the objective function. The step size can be solved in two ways using the exact and the inexact line search approaches.

Some step sizes have been proposed by many researchers such as Forshyte (1968), Armijo (1966), Barzilai and Borwein (1988) and Jorge and Stephen (2006) as detailed below:

1. **Cauchy Rule (C step size):** This step size was introduced by Cauchy (1847) which was computed using the exact line search technique [5].

$$\alpha_k = \frac{g_k^T g_k}{g_k^T H_k g_k} \tag{11}$$

2. **Armijo Rule (A step size):** This step size uses the inexact line search technique [6].

Given that $s > 0,\ \beta,\ \sigma \in (0,1)$, let α_k be the largest α in $\{s, s\beta, s\beta^2, \dots\}$ such that

$$f(x_k + \alpha_k d_k) \le f(x_k) + \sigma \alpha_k g_k^T d_k$$

3. **Backtracking Rule (B step size) [8]:**

Given that $\beta, \sigma \in (0,1)$, $\widetilde{\alpha_k} = 1$.

$$\alpha_k = \beta \widetilde{\alpha_k} \tag{12}$$

such that

$$f(x_k + \alpha_k d_k) \leq f(x_k) + \sigma \alpha_k g_k^T d_k$$

4. **Barzilai-Borwein1 Rule (BB1 step size)** [7]**:**

$$\alpha_k = \frac{s_{k-1}^T y_{k-1}}{\|y_{k-1}\|_2^2} \tag{13}$$

where $s_{k-1} = x_k - x_{k-1}$ and $y_{k-1} = g_k - g_{k-1}$.

5. **Barzilai and Borwein 2 Rule (BB2 step size)** [10]**:**

$$\alpha_k = \frac{\|s_{k-1}\|_2^2}{s_{k-1}^T y_{k-1}} \tag{14}$$

where $s_{k-1} = x_k - x_{k-1}$ and $y_{k-1} = g_k - g_{k-1}$.

IV. NUMERICAL EXPERIMENTS

This section is devoted to test and compare all step sizes in (11) - (14) with the procedure of FR conjugate gradient (CG) algorithm. Python programming language is used for the implementation of the logistic regression problem.

A. Description of the problems

i. Problem 1

This data was collected from the admission office of The Gateway (ICT) Polytechnic Saapade, Ogun State, Nigeria for candidates seeking admission into the institution.
From the dataset in table 1, observation shows that the problem is a binary classification problem which contains 10 features.

ii. Problem 2

Iris flowers dataset was downloaded from github.com. The Iris flowers data involves predicting the flower species given measurements (in cm) of the iris flowers. The attributes information are:

1. Sepal length
2. Sepal width
3. Petal length

4. Petal width
5. Class (iris Setosa (1) and iris virginica (0))

sepal_length	sepal_width	petal_length	petal_width	Species
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3	1.4	0.2	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa
4.6	3.1	1.5	0.2	Iris-setosa
5	3.6	1.4	0.2	Iris-setosa
5.4	3.9	1.7	0.4	Iris-setosa
4.6	3.4	1.4	0.3	Iris-setosa
5	3.4	1.5	0.2	Iris-setosa
4.4	2.9	1.4	0.2	Iris-setosa

Table 2: Sample of dataset for problem 2

Logistic regression was built for both problem 1 and problem 2.

B. Parameters settings

The following parameters are stated for some line search conditions:

$s = 1, \beta = 0.0075, \sigma = 0.38$ for the Armijo rule (A) to solve problem 1. $s = 1, \beta = 0.01, \sigma = 0.38$ for the Armijo rule (A) to solve problem 2. $\sigma = 0.001$ for Backtracking rule (B). The initial step size $\alpha_0 = 0.001$ for the Barzilai-Borwein step size 1 (BB1) and Barzilai-Borwein step size 2 (BB2). All other logistic regression parameters are set to 0. The numerical result was compared based on: time of execution, total number of iterations, accuracy and the most decreased value of objective function obtained. The stopping criteria is set to $\|g_k\| \leq 10^{-6}$.

C. Results and Discussion

Abbreviations:
F1 = Fixed learning rate, set as 0.0001
F2 = Fixed learning rate, set as 0.001
A = Armijo learning rate
B = Backtracking learning rate
BB1 = Barzilai and Borwein learning rate 1
BB2 = Barzilai and Borwein learning rate 2

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (15)$$

$$Precision = \frac{TP}{TP + FP} \quad (16)$$

True Positive (TP): Correctly predict positive (1)

True Negative (TN): Correctly predict negative (0)
False Positive (FP): Predict negative (0) class as positive
False Negative (FN): Predict positive (1) class as negative

Problem No.	Learning rate methods	Numbers of Iteration	Processing Time μs	Accuracy %	Precision %	f
1.	F1	541	6.57	88.10	87.57	1259.106
	F2	169	1.87	88.10	87.57	1259.106
	A	19	0.11	87.23	87.63	1260.809
	B	22	0.14	88.01	87.87	1264.87
	BB1	-	-	-	-	-
	BB2	-	-	-	-	-
2.	F1	1304	10.57	92	92.59	7.42248
	F2	440	3.66	92	92.59	7.42248
	A	109	0.9	92	92.59	7.42248
	B	147	1.55	92	92.59	7.42248
	BB1	-	-	-	-	-
	BB2	-	-	-	-	-

Table 3: Numerical Results

The numerical result obtained in all experiments shows that Armijo method reaches the optimal values (minimum cost) faster than other methods with 19 iterations and 109 iterations for problem 1 and 2 respectively. Table 3 also show that Armijo method has the least processing time with $0.11\mu s$ and $0.9\mu s$ for problem 1 and problem 2 respectively. All methods are highly competitive in terms of accuracy and precision except for BB1 and BB2 methods which fails to solve both problem 1 and 2. Therefore, from the experiment Armijo rule is a better learning rate method than others in terms of number of iterations and processing time.

V. CONCLUSION

This paper, we applied different learning rate (step size) methods to solve real-life binary logistic regression problems. According to the results Armijo (A) and Backtracking (B) rules perform well in solving the problems in terms of number of iterations and processing time. A and B methods are also competitive with the fixed learning rate in terms of accuracy and precision. A and B rules are better with the initial conditions given for each problem.

References

Armijo L. (1966) Minimization of function having Lipschitz continuous first partial derivatives Pacific J. of Mathematics 6 pp 1-3.

Barzilai J and Borwein J 1988 Two-point step size gradient methods IMA J. of Numer. Anal. 8 141-148.

Forshyte G. E (1968) On the Asymptotic directions of the s-dimensional optimum gradient method Numerische Mathematik, 11 pp 57-76.

Hamoda. M., Rivaie. M., Mamat. M. & Salleh, Z. (2015) "A new nonlinear conjugate gradient coefficient for unconstrained optimization", Applied Mathematical Sciences, 9(37), 1813-1822.

Jorge Nocedal and Stephen J Wright (2006) Line search methods. Numerical Optimization, pages 30-65.

Yuan Sun; Zhihao Zhang; Zan Yang; Dan Li (2019). "Application of logistic regression with fixed memory step gradient descent method in multi-class classification problem", the 2019 6th International Conference on Systems and Informatics.

Yuan Y (2008) Step-sizes for the gradient method In Third International Congress of Chinese Mathematicians. Part 1, 2 AMS/IP Stud. Adv. Math., 42, Pt. 1 2 785–796. Amer. Math. Soc., Providence, RI.

Yu-Hong Dai (2010). "Nonlinear Conjugate Gradient Methods". Wiley Encyclopedia of Operations Research and Management Science.

Appnum	SexName	Age	StateName	PName	SessionName	JambNumber	JambScore	PUTMEScore
GTS1912706	MALE	19	ABIA	ACCOUNTANCY	2019/2020	96479912BJ	192	26
GTS1915164	FEMALE	17	ABIA	ACCOUNTANCY	2019/2020	96460260JG	173	26
GTS1919950	MALE	23	ABIA	ACCOUNTANCY	2019/2020	96655844JH	168	23
GTS1921099	FEMALE	21	ABIA	ACCOUNTANCY	2019/2020	96937039CE	186	21
GTS1920264	FEMALE	22	AKWA-IBOM	ACCOUNTANCY	2019/2020	95137630AH	182	19
GTS1918711	MALE	24	AKWA-IBOM	ACCOUNTANCY	2019/2020	96590011AH	165	32
GTS1915094	FEMALE	21	AKWA-IBOM	ACCOUNTANCY	2019/2020	96400526EC	172	25
GTS1915576	FEMALE	22	AKWA-IBOM	ACCOUNTANCY	2019/2020	99999999AA	150	28
GTS1916734	MALE	23	AKWA-IBOM	ACCOUNTANCY	2019/2020	96936124FD	188	30
GTS1919901	FEMALE	20	ANAMBRA	ACCOUNTANCY	2019/2020	96527252AH	201	25
GTS1920829	MALE	24	BENUE	ACCOUNTANCY	2019/2020	96911414HF	186	18
GTS1915872	MALE	24	BENUE	ACCOUNTANCY	2019/2020	96910259EI	169	27
: .								

Table1: Sample of dataset for problem 1