Journal of Advances in Mathematics and Computer Science

34(6): 1-13, 2019; Article no.JAMCS.54382
ISSN: 2456-9968
(Past name: British Journal of Mathematics & Computer Science, Past ISSN: 2231-0851)

Bayesian Optimization for Parameter of Discrete Weibull Regression

Adesina, Olumide Sunday[1*], Onanaye, Adeniyi Samson[1] and Okewole, Dorcas Modupe[1]

[1]Department of Mathematical Sciences, Redeemer's University, Ede, Osun State, Nigeria.

Authors' contributions

This work was carried out in collaboration among all authors. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/JAMCS/2019/v34i630233
Editor(s):
(1) Dr. Wei-Shih Du, Professor, Department of Mathematics, National Kaohsiung Normal University, Taiwan.
Reviewers:
(1) Irshad Ullah, Pakistan.
(2) Papadakis Stamatios, University of Crete, Greece.
Complete Peer review History: http://www.sdiarticle4.com/review-history/54382

Original Research Article

Received: 17 November 2019
Accepted: 23 January 2020
Published: 28 January 2020

Abstract

This study aim at optimizing the parameter θ of Discrete Weibull (DW) regression obtained by maximizing the likelihood function. Also to examine the strength of three acquisition functions used in solving auxiliary optimization problem. The choice of Discrete Weibull regression model among other models for fitting count data is due to its robustness in fitting count data. Count data of hypertensive patients visits to the doctor was obtained at Medicare Clinics Ota, Nigeria, and was used for the analysis. First, parameter θ and β were obtained using Metropolis Hasting Monte Carlo Markov Chain (MCMC) algorithm. Then Bayesian optimization was used to optimize the parameter the likelihood function of DW regression, given β to examine what θ would be, and making the likelihood function of DW the objective function. Upper confidence bound (UCB), Expectation of Improvement (EI), and probability of Improvement (PI) were used as acquisition functions. Results showed that fitting Bayesian DW regression to the data, there is significant relationship between the response variable, β and the covariate. On implementing Bayesian optimization to obtain parameter new parameter θ of discrete Weibull regression using the known β, the results showed promising applicability of the technique to the model, and found that EI fits the data better relative to PI and UCB in terms of accuracy and speed.

Keywords: Machine learning; Bayesian optimization; Gaussian process; acquisition function; discrete weibull regression; medicine; count data.

**Corresponding author: E-mail: olumidestats@gmail.com;*

Abbreviations

$\mathfrak{D}_1$ =*Initial Trained Data*
$\mathbb{R}^n$ =*Real coordinate space of n dimensions*
I_n =*Identity matrix of* $\mathbb{R}^n$
$x = (x_1, \ldots \ldots . x_n)$, *matrix of dimension* $n \times d$
$x^* = (x_1^*, \ldots, x_1^*)$, *matrix of dimension* $n^* \times d$

1 Introduction

Optimization problem is generally known for maximization or minimization. Objective of optimization may be for profit maximization or accuracy maximization. On the other hand, one may consider cost minimization, or loss minimization depending on the task. Bayesian optimization (BO) is in the body of knowledge of machine learning, and it is used to optimize difficult black-box optimization problems, where the objective function is costly to evaluate. BO can be used in a case where the objective function is vague. [1,2] pointed out that for black box optimization to take place all dimensions should have bounds on the search space. Bayesian optimization broadly looks into right combinations of hyperparamters that will yield maximum accuracy [3,4] or minimize loss such as computational cost. The works of [5] showed the case of minimizing uncertainty using BO technique. Also, [6,7,2] demonstrated the usefulness of machine learning and Bayesian Optimization in finance and portfolio management respectively.

In parametric models, there is possibility for likelihood functions to be intractable and requires numerical computation, among techniques to obtaining parameters from intractable likelihoods is Metropolis Hasting MCMC algorithm based on Bayesian estimation technique. [8], provide some basic and core procedure for implementing Bayesian analysis, one of which include stepwise computer program-based procedure for obtaining target parameters from intractable likelihoods. ([8], pp. 499). Bayesian procedure involves setting a prior distribution combined with likelihood of a given data using Bayes rule to obtain the posterior distribution. Estimates of the parameters of interest are then taken from the posterior distribution. In similar fashion, Bayesian Optimization requires (i) generating objective function $f(x)$ that needs to be maximized or minimized (ii) building Gaussian prior (probabilistic model) for the objective function $f(x)$ (iii) update the probability distribution using samples drawn from the objective function to get a new distribution, called posterior distribution (iv) determine the utility function (or acquisition function), that will be used to solve auxiliary optimization problem, and determine where to make the next measurement. (v) given (iv), update the Gaussian Process posterior probability distribution (vi) repeat step (iii)-(v) until stopping criteria is met, that is, when the objective function is approximately maximized.

Bayesian Optimization aim at optimizing hyperparameters, in machine learning a hyperparameter is always set before the learning, while the values of other parameters are obtained in the process of the training. One reason why Bayesian optimization is attractive is because it can be applied to popular black-box functions, as it only requires input-output process [1,9,10]. Another reason why BO is attractive is that one can carefully select next design to evaluate, thereby reducing computational cost.

In the work of [11], the authors proposed Bayesian optimization for likelihood-free inference models, which are simulator-based. The authors adopted a strategy of combining probabilistic modeling with optimization to obtain inference for models that are likelihood-free. Some other authors who applied Bayesian optimization in various fields is [3,4], the authors applied BO to robot gait parameters that optimize, and showed reliability of BO in maximization, as it outperformed existing techniques. Another area of application can be found in [12]. Just like in the case of [3,4], [5] also applied Bayesian optimization to find a policy for robot path that would minimize uncertainty about its location and heading. In contrast with [11] who carried out Bayesian optimization likelihood-free inference models, we propose Bayesian optimization for likelihood based inference of a two parameter Discrete Weibull regression.

The remaining part of this paper is organized as follows, Material and Methods in section 2, which include details on Gaussian process, Bayesian optimization, Utility function, and hyperparamter selection, Section 3 contains results and discussion. Finally, conclusion was drawn in section 4.

2 Materials and Methods

2.1 Gaussian process

A Gaussian process (GP) has infinite dimension stochastic process of multivariate Gaussian distribution, and any finite combination of dimensions will be a Gaussian distribution. GP helps in improving global optimization by estimating the confidence region of a sample test. Regular method cannot be used to obtain the solution of intractable objective function of black-box optimization problems, owing to difficulty in evaluating the gradient of the object function. Therefore, Bayesian optimization can replace the unknown objective function, as well as difficult to evaluate problem by sequence of optimization problem, following Gaussian process. [13] defined Gaussian process as follows

Definition 1: For a set X in $\mathbb{R}^d$, a Gaussian process can be defined as a collection of $\{f(x), x \in \mathrm{X}\}$ such that for $x_1, \dots, x_n \in \mathrm{X}$ and for any $n \in \mathbb{N}$, the random vector $(f(x_1), \dots, f(x_n))$ has a joint multivariate Gaussian distribution. From definition 1, GP process can be characterized by its mean function

$$m(x) = E[f(x)] \tag{1}$$

Its covariance function can be expressed as

$$\kappa(x, x') = cov(f(x), f(x')) = E\left[\left(f(x) - m(x)\right)\left(f(x') - m(x')\right)\right] \tag{2}$$

The covariance function is $\kappa(x, x')$, called kernel function is pivotal in analyzing GP. The popularly used kernel function is the squared exponential function, and given by

$$\kappa_{SE}(x, x') = exp\left(-\frac{1}{2} \parallel x - x' \parallel^2\right) \tag{3}$$

Where $x \; and \; x' \in \mathbb{R}^d$

2.1.1 Gaussian process regression (GPR)

For a set of training set $\{(x_i, y_i)\}_{i=1}^n$, the objective of GPR aim at predicting $f(x^*)$ for new inputs $x^* \in \mathbb{R}^{n \times d}$. Bayesian technique would be used to obtain the posterior distribution for Gaussian process on $x = (x_1, \dots, x_n) \in \mathbb{R}^{n \times d}$. With the assumption that observations are noise-free, that is, $f(x) = y = (f(x_1), \dots f(x_n))$, where f is the Gaussian process.

If the new inputs are represented by x^* and n^* and conditional prediction $y^* = E[f(x^* \mid x, y)]$. Then

$$f(x, x^*) \sim \mathcal{N}\left(0_{n+n^*} \begin{pmatrix} \kappa(x, x) & \kappa(x, x^*) \\ \kappa(x^*, x) & \kappa(x^*, x^*) \end{pmatrix}\right) \tag{4}$$

Where $\kappa(x^*, x)$ is $n^* \times n$ matrix and entries $\kappa_{i,j}(x^*, x) = \kappa\left(x_i^*, x_j^*\right)$, [2] showed that the vector $y^* = E[f(x^* \mid x, y)]$ is equally Gaussian:

$$f(x^* \mid x, y) = \mathcal{N}(m(x^* \mid x, y), \kappa(x^*, x^* \mid x, y))$$

Where $m(x^* \mid x, y)$ represents the mean vector of the posterior distribution, given as

$$m(x^* \mid x, y) = \kappa(x^*, x)\kappa(x, x)^{-1} y$$

2.2 Bayesian Optimization (BO)

Bayesian optimization for unconstrained optimization, as well as the algorithm is presented in this section. The general Bayesian optimization technique for a maximization problem aim at optimizing the following problem:

$$x^* = \underset{x \in \mathrm{X}}{\operatorname{argmax}} f(x) \tag{5}$$

The objective is to compute a maximizer x^* of expensive-to-evaluate function f. In order to apply BO to solve the optimization problem in equation (1), we start with an initial training set $\mathcal{D}_1 = \{(x_1, f(x_1))\}$, for a design x_1 with the corresponding objective function $f(x_1)$. If our prior belief is the objective function, provided that the data is noiseless, then Bayesian principle of combining prior belief with likelihood of the data using Bayes rule to obtain posterior distribution is expressed as follows.

$$p(f \mid \mathcal{D}_1) \propto p(\mathcal{D}_1 \mid f) \cdot p(f)$$

Let the training data from initial level to n be represented by $\mathcal{D}_1, \ldots \mathcal{D}_n$, the training set $\mathcal{D}_n$ is used to build statistical model which is basically a Gaussian process [14]. The posterior mean and posterior variance for any given design of interest x is $\bar{\mu}(x;\ \mathcal{D}_n, f)$, and $\bar{\sigma}^2(x;\ \mathcal{D}_n, f)$. The posterior mean is the surrogate function of f at x, provided $\mathcal{D}_n$ can be cheaply calculated so that it will serve as surrogate for the objective function f, while $\bar{\sigma}^2(x;\ \mathcal{D}_n, f)$ serves as surrogate function of x.

The acquisition function $\mathcal{U}_n(x, \mathcal{D}_n)$ is maximized at each iteration, and it also helps to identify where next to sample from the objective function. The acquisition function measures the benefits of estimating a new design x based on the surrogate model generated with $\mathcal{D}_n$ .The next model to estimate x_{n+1} is determined by solving auxiliary optimization problem that will maximize the utility function. Iteration continues until when the objective function approximately maximized.

2.3 Hyperparameters

Hyperparameters needs to be selected, and the rule for selecting Hyperparameters is maximizing the likelihood of the parameters of the model, say φ. The likelihood is represented by $L(\varphi) = p(y \mid \varphi)$, where y is the sample data, which its probability needs to be maximized. Let $g = f(z)$ represents a GP, then,

$$p(y \mid \varphi) = \int p(y \mid \varphi, g) p(y \mid \varphi) dg \tag{6}$$

The log marginal likelihood is usually maximized in most cases by integrating out the latent values of g, so as to solve

$$\hat{\varphi} = \underset{\varphi}{\operatorname{argmax}}\, \ell(\varphi)$$

where $\ell(\varphi) = ln\, p(y \mid \varphi)$. When Gaussian noise is considered, it results to $Y \sim \mathcal{N}(0_n, \kappa(\varphi_\kappa) + \sigma_\varepsilon^2 I_n)$, where $\kappa(\varphi_\kappa)$ is the kernel matrix subject to kernel parameters φ_κ. With Probability density function of Gaussian distribution the following is obtained:

$$\ell(\varphi) = \frac{n}{2} ln\, (2\pi) - \frac{1}{2} \mid \kappa(\varphi_\kappa) + \sigma_\varepsilon^2 I_n \mid - \frac{1}{2} Y^T (\kappa(\varphi_\kappa) + \sigma_\varepsilon^2 I_n)^{-1} Y \tag{7}$$

However, [15] mentioned that $\ell(\varphi)$ is not always convex, and can have problem of local maxima.

2.4 Acquisition functions

In Bayesian Optimization, acquisition function is needful because it helps the search to active optimum results. When acquisition is maximized, it is used to determine the next point at which function will be evaluated. Simply put, the objective is to sample function f at $argmax_x \mathcal{U}(x \mid \mathcal{D})$, where $\mathcal{U}(\cdot)$ which is generic stands for acquisition function. In the works of [1], the authors presented three types of acquisition functions, the acquisition functions, which include (i) the probability of improvement (PI) (ii) Expected improvement, and (iii) Upper confidence bound (UCB). which are considered in this paper. Once estimation is done with f, the training set is augmented with the new training point $\left(x_{n+1}, f(x_{n+1})\right)$.

2.4.1 Improvement-based acquisition function

Given that $x^+ = argmax_{x_{i \in x_i}} f(x_i)$, maximizing the probability of improvement over first n steps $f(x^+)$ is of great value, so that we can express Probability of improvement (PI) of x as

$$PI(x) = P\big(f(x) \geq f(x^+)\big) = \Phi\left(\frac{\mu(x) - f(x^+)}{\sigma(x)}\right) \tag{8}$$

The function $\Phi(\cdot)$, sometimes called "maximum probability of Improvement" (MPI) is the normal cumulative distribution function (CDF). [1] mentioned that the formulation is primarily exploitation, that is, (sampling from areas of high uncertainty), and will leave out exploration, that is, (sampling areas likely to offer improvement over the current best observation), and that constitute a disadvantage of the formulation. To make up for the drawback associated with PI is to modify equation (8) by adding an adjustment parameter $\varphi \geq 0$, leading to:

$$PI(x) = P(f(x) \geq f(x^+) + \varphi) = \Phi\left(\frac{\mu(x) - f(x^+) + \varphi}{\sigma(x)}\right) \tag{9}$$

The user is at liberty of choosing the exact value of φ. What makes this formulation attractive is that in the process of maximizing $PI(\cdot)$, it chooses the point that will most likely give an improvement of at least φ. Performance of $PI(\cdot)$ can be found in the work carried out by [16]. Following the shortcomings identified with $PI(\cdot)$, a more satisfying acquisition function that would capture both probability of improvement, and also capture the magnitude of improvement that a point can possibly yield. This would be done by minimizing expected deviation from the true maximum $f(x^*)$. For a new sample point we'll have:

$$x_{t+1} = \underset{x}{\operatorname{argmin}}\, \mathbb{E}(\| f_{t+1}(x) - f(x^*)\| \mid \mathcal{D}_{1:t})$$

Integrating we have

$$= \underset{x}{\operatorname{argmin}} \int \| f_{t+1}(x) - f(x^*)\| P(f_{t+1} \mid \mathcal{D}_{1:t})\, df_{t+1} \tag{10}$$

The equation given in (10) has its limitation because it can only consider one-step-ahead choices. Therefore, [17] proposed another way of maximizing the expected improvement so that one can have many steps ahead as desired with respect to $f(x^+)$. The improvement function by [17] is defined as

$$I(x) = max\{0, f_{t+1}(x) - f(x^+)\}$$

Meaning that $I(x)$ is positive when prediction is higher than the best known value so far. Otherwise, $I(x)$ would be set to zero. Maximizing the expected improvement follows that

$$x = \underset{x}{\operatorname{argmin}}\, \mathbb{E}(\max\{0, f_{t+1}(x) - f(x^+)\} \mid \mathcal{D}_t)$$

The likelihood I of improvement on a normal posterior distribution $\mathcal{N}(\mu(x), \sigma^2(x))$can be computed from normal PDF.

$$f(x, \sigma\ |I) = \frac{1}{\sqrt{2\pi\sigma(x)}} exp\left(-\frac{(\mu(x) - f(x^+) - I)^2}{2\sigma^2}\right)$$

The expected of improvement $\mathbb{E}(I)$ is given as

$$\mathbb{E}(I) = \int_{I=0}^{I=\infty} I \cdot \frac{1}{\sqrt{2\pi\sigma(x)}} exp\left(-\frac{(\mu(x) - f(x^+) - I)^2}{2\sigma^2}\right) dI \tag{11}$$

[18] provided the result of evaluation of (7) and given as

$$\mathbb{E}(I) = \begin{cases} \big(\mu(x) - f(x^+)\big)\Phi(Z) + \sigma(x)\phi(Z) & if \quad \sigma(x) > 0 \\ 0 & if \quad \sigma(x) = 0 \end{cases}, Z = \frac{\mu(x) - f(x^+)}{\sigma(x)} \tag{12}$$

$\phi(\cdot)$ and $\Phi(\cdot)$ represent the PDF and CDF of the standard normal distribution, $\mathcal{N}(0,1)$ respectively.

Another type of acquisition function introduced by [19,20], is the "Sequential Design for optimization" (SDO), the SDO chooses points to be evaluated based on either the lower confidence bound:

$$LCB(x) = \mu(x) - \kappa\sigma(x),$$

Where $\kappa \geq 0$, which aim is to minimize.

Or on the upper confidence bound in order to maximize

$$UCB(x) = \mu(x) + \kappa\sigma(x),$$

The user is left to decide the value of parameter κ. In order to make Bayesian optimization to be applicable to solving diverse problems, [21] proposed another acquisition function, which is considered as instantaneous regret function.

$$r(x) = f(x^*) - f(x)$$

It is designed in a way to optimize:

$$min \sum_t^n r(x_n) = max \sum_t^T f(x_t),$$

Where n is the number of iterations that would run in other to optimize the function.

Using the upper confidence bound selection criterion wit $\kappa_t = \sqrt{\omega\tau_t}$ and the hyperparameter $\omega > 0$, it follows that [21] defined Gaussian process-UCB as

$$\text{GP-}UCB(x) = \mu(x) + \kappa\sqrt{\omega\tau_t}\sigma(x) \tag{13}$$

Another acquisition functions was discussed by [2]; they are entropy-based acquisition functions in the study by [22], the earlier works of [9] has proven to have taken care of the challenges identified in the works of [22]. In achieve that, [9] proposed technique called *predictive entropy search* (PES). *Knowledge gradient-based acquisition function* (KG) was also proposed by [23], KG is a closed form of the expected

improvement. What differentiate KG from EI is that KG takes account of noise and does not limit the final solution to a previously estimated point. Therefore, if a given data is noiseless, and the final solution is limited to the previous sampling, the KG acquisition function would be reduced to EI acquisition function.

As a way of illustration, if the objective is to obtain the global minimum of $f(x) = (3x-2)^2 \sin(12-4)$, it simply imply that the value of x that would result in global minimum of $f(x)$ is required. The graph of $(minf(x) = (3x-2)^2 sin\ (12-4))$, is shown in Figure 1. The global minimum is 0.1845.

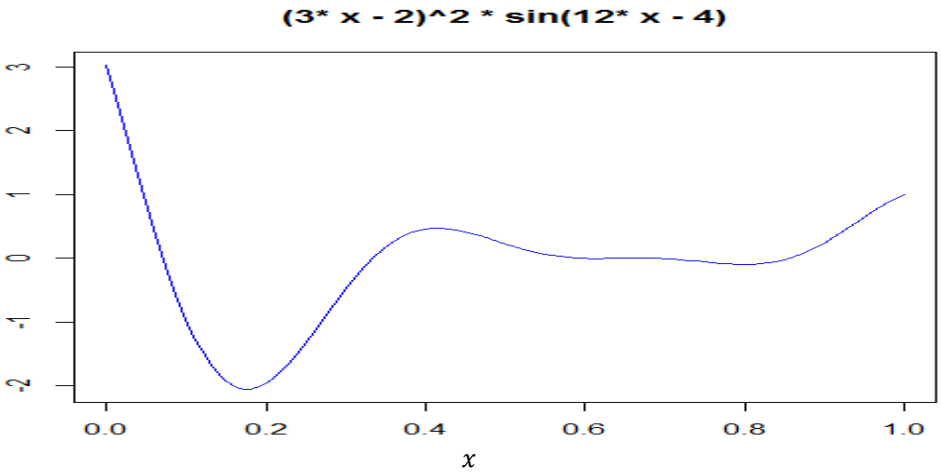

Fig. 1. Graph of $\boldsymbol{minf(x) = (3x-2)^2 sin\ (12-4)}$

2.5 Application to parameter optimization of discrete weibull distribution

2.5.1 Discrete weibull regression

With respect to Discrete Weibull distribution of type III identified in the work of [24], if a random variable Y follows a Discrete Weibull distribution with parameters $q, and\ \beta$ then the probability density function (pdf) is:

$$G(y;q,\beta) = \begin{cases} 1 - q^{(1+y)^{\beta}}, & y = 0,1,2 \ldots \\ 0, & otherwise \end{cases} \tag{14}$$

While its probability mass function would be:

$$G(y;q,\beta) = \begin{cases} q^{y^{\beta}} - q^{(1+y)^{\beta}}, & y = 0,1,2 \ldots \\ 0, & otherwise \end{cases} \tag{15}$$

The log-likelihood be given as

$$\ell(y,\beta) = \prod_{i=1}^{n} \log\left(q^{y^{\beta}} - q^{(1+y)^{\beta}}\right) \tag{16}$$

Which precedes estimation of parameters of Discrete Weibull, from which the MLEs of q and β can be easily obtained by directly maximizing the log-likelihood. Discrete Weibull regression has both log link and logit link [25,26], for $X_1,\ X_2 \ldots X_p,\ p$ covariates, logit link follows that

$$\log(q/1-q) = \theta_0 + \theta_1 x_1 + \cdots + \theta_p x_p \tag{17}$$

From (17) q becomes,

$$q = \frac{e^{x_i\theta}}{1+e^{x_i\theta}} \tag{18}$$

Substituting (18) into (16), we have,

$$\ell(\theta,\beta|x,y) = \prod_{i=1}^{n} \log\left(\left(\frac{e^{x_i\theta}}{1+e^{x_i\theta}}\right)^{y^\beta} - \left(\frac{e^{x_i\theta}}{1+e^{x_i\theta}}\right)^{(1+y)^\beta}\right) \tag{19}$$

2.5.2 Data

Parameter of Discrete Weibull was optimized using health data of high blood pressure patients. The data of total number of 181 of hypertensive patients was obtained from Medicare health facility in Ota ogun State, the data covers visits from July 2016 to July 2017. The response variable is count of visits of hypertensive patients to the doctor while the predictor is the Blood pressure control (0=poor, 1=good). The count regression is to determine if pressure control is responsible for visit to the doctor using Bayesian Discrete Weibull model.

2.5.3 Implementation

Metropolis Hasting Monte Carlo Markov chain [26] was used to sample from the posterior distribution of Bayesian Discrete Weibull with 20,000 iterations using Laplace prior, because Discrete Weibull does not have a conjugate prior and Laplace prior is considered a suitable prior distribution [25]. Bayesian procedure outlined by [26,27,28] was used to obtain parameters of interest. From the likelihood function given in (19), estimation for parameters $\theta_0, \theta_1, and\ \beta$ were obtained, and reported using algorithm 1. Given the parameter β obtained and the data, (19) was taken to be objective function, with the aim of optimizing parameter θ , so as to obtain new values of θ using Upper confidence bound (UCB), Expectation of Improvement (EI), and probability of Improvement (EI) as acquisition functions. Software [29] was used to carry out the analysis. Package "BDWreg" by [30] was used to fit Bayesian Discrete Weibull regression, while package "rBayesianOptimization" by [31] was used to implement for the analysis and 5, and 50 iterations were carried out respectively.

Algorithm 1: Metropolis Hasting MCMC algorithm
Initialization: Choose a starting point, say x^1
$i = 2, \ldots., L$ **do**
1. Given x^{i-1}, generate $\tilde{x} \sim q(x^{i-1}, x)$
2. Generate $a = \frac{(\pi(\tilde{x})/q(x^{i-1},\tilde{x}),1)}{\pi(x^{i-1})/q(\tilde{x},x^{i-1})}$
3. Compute $\phi(x^{i-1},\tilde{x}) = min(a,1)$
4. With probability $\phi(x^{i-1},\tilde{x})$, accept $\tilde{x}$ and set x^i=$x^{(i-1)}$
else
draw a random value u uniformly from the unit interval $[0,1]$
if $u < a$, then $x^i = \tilde{x}$ (accept)
else
$x^i = x^{i-1}$ (accept)
end if
end

Initial point for both 5 and 50 iterations was 10, making a total of 60 rounds. The tunable parameter of GP Upper Confidence Bound was taken to be 2.576, so as to balance exploitation against exploration. -0.0315 and 0.4729 constitute the lower and upper bounds of each hyperparaneter, from the values theta obtained from algorithm. The tunable parameter of Expected Improvement and Probability of Improvement was taken to be 2. Square exponential was used as the Kernel (correlation function) for the Gaussian Process. The data is noise-free hence, using knowledge gradient-based acquisition function would be the same as using expectation of improvement acquisition function. Algorithm for Bayesian optimization is described as follows:

Algorithm 2: Bayesian Optimization
Input: Initial training set $\mathcal{D}_1$, hyperparameters φ_1, utility function $\mathcal{U}_n$
for $n = 1, 2, \ldots . N$ ***do***
build GP using $\mathcal{D}_n$
solve the auxiliary optimization problem for $x_{n+1} = argmax_{x \in X} \mathcal{U}_n(x, \mathcal{D}_n)$
Evaluate $f(x_{n+1})$
Update the data:
$\mathcal{D}_{n+1} \longleftarrow \mathcal{D}_n \cup \left\{\left(x_{n+1}, y_{n+1,} \widehat{f}_{n+1}(x_{n+1})\right)\right\}$
Update the hyperparamter vector φ_{n+1} kernel function
end for
return $\mathcal{D}_n$ and φ_n

3 Results and Discussion

Table 1. Result of parameter estimation of discrete Weibull model, using count data

Parameter	Lower	Estimate	Upper
θ_0	1.7671	1.9509	2.3137
θ_1	0.0295	0.4729	0.7778
β	1.5059	1.6709	1.8365

Table 2. Sampler for Bayesian discrete weibull

Sampler		
Iterations: 30000	Logit : TRUE	Scale : 0.01
Rev.Jump: FALSE	RegQ :TRUE	RegB : FALSE
Penalized: FALSE	Fixed.penalty :FALSE	
Model Summary		
AIC : 729.5721	AICc : 729.7062	BIC : 739.1676
QIC : 4.034054	CAIC : 742.1676	LogPPD : -371.5198
DIC : 729.0273	PBIC : 732.0322	df : 3

Result for Bayesian estimation for Discrete Weibull model is given in Table 1 using metropolis Hasting MCMC to sample from the posterior distribution. This response variable was count, while the covariate is blood pressure control (θ_1). The result shows that there is significant relationship between response variable, β, and θ_1. Table 2 contains model selection criteria. Figure 2 shows that θ_0, θ_1 and β (with red unbroken line) are significant.

Table 3 contains the value of best parameter of θ resulting from posterior distribution after maximizing the likelihood function given in equation (19). Different acquisition functions were used at 5 and 50 iterations respectively. As number of iteration increases, results become more accurate. Best parameter of θ was obtained afterwards. The values obtained differ significantly from the one that was initially obtained with

Bayesian Discrete Weibull model (0.4729). The absolute difference of best parameter for 5 and 50 using Upper confidence bound (UCB) is 0.2626, and the absolute difference of best parameter for 5 and 50 using probability of improvement (PI) is 0.019, while that of Expectation of Improvement (EI) is 0.016. This shows that EI is more consistent among the three methods. Column 4 of Table 3 represents the round at which θ was optimized; the rounds is between 1 and 15 for 5 iterations, while for 50 iterations it was between1 and 60 since initial point was specified was 10.

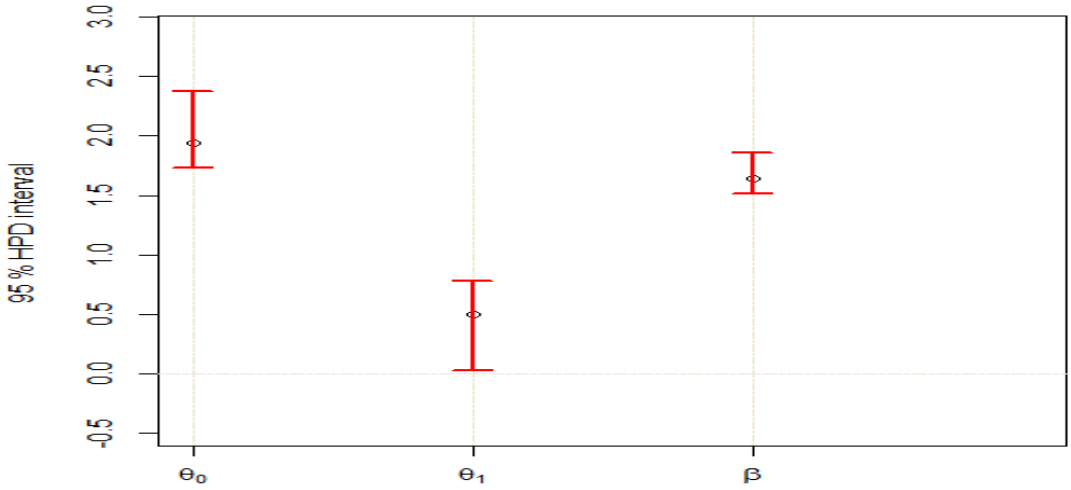

Fig. 2. Graph showing posterior estimation of the relationship between the covariate and the response variable

Table 3. Parameter tuning of discrete weibull model, given the data

No of iterations	Acq. function	Best parameter	Opt. round
5	UCB	0.3855	1
5	POI	0.1039	4
5	EI	0.1086	1
50	UCB	0.1226	36
50	POI	0.1229	14
50	EI	0.1252	54

Generally, the result takes some time to return using package "rBayesianOptimization" in R. Also the running time for running UCB takes approximately 3 times as much as EI and UCB with "rBayesianOptimization" package in R.

4 Conclusion

In this study, Bayesian optimization procedure for likelihood-based function was given. The study aimed at obtaining a new parameter θ, given β. Therefore, making the likelihood function of Discrete Weibull the objective function. Three (EI, UCB, and PI) acquisition functions were used for 5 and 50 iterations respectively. The results obtained show practical applicability of Bayesian optimization to tuning a target parameter having obtained the likelihood function, which is a deviation from the works of [11]. On applying Bayesian technique to the resulting regression model, the results show that there is significant relationship, between blood pressure control of hypertensive patients and number of visits to the doctor with estimate of $\theta = 0.4729$. After subjecting the likelihood function to BO different values of θ were obtained and presented in Table 3.

On comparing the three acquisition functions, the EI appear to be more consistent relative to UCB and PI based on computation speed and accuracy. Future work may consider implementing BO using various statistical software and compare based on speed, computational cost, and accuracy. The case of more than one covariate may also be considered in future works.

Disclaimer

The consent of Medical Director of the health facility where data was collected and he gave approval for using the data.

Competing Interests

Authors have declared that no competing interests exist.

References

[1] Brochu E, Cora VM, De Freitas N. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. arXiv. 2010;1012:2599.

[2] Gonzalvez L, Lezmi E. Financial applications of Gaussian processes and Bayesian optimization. arXiv. 2019;1903:04841.

[3] Lizotte D, Wang T, Bowling M, Schuurmans D. Automatic gait optimization with Gaussian process regression. In IJCAI; 2007.

[4] Lizotte D. Practical Bayesian optimization. PhD Thesis, University of Alberta, Edmonton, Alberta, Canada; 2008.

[5] Martinez-Cantin R, de Freitas N, Brochu E, Castellanos J, Doucet A. A Bayesian exploration-exploitation approach for optimal online sensing and planning with a visually guided mobile robot. Autonomous Robots. 2009;27(2):93-103.

[6] McKinsey. FinTechnicolor: The new picture in finance. McKinsey Report; 2016.

[7] Bourgeron T, Lezmi E, Roncalli T. Robust asset allocation for Robo-advisors; 2018. SSRN. Available:www.ssrn.com/abstract=3261635

[8] Barber D. Bayesian reasoning and machine learning; 2010. Available:http://web4.cs.ucl.ac.uk/staff/D.Barber/textbook/270212.pdf

[9] Hernandez-Lobato JM, Gelbart MA, Homan MW, Adams RP, Ghahramani Z. Predictive entropy search for Bayesian optimization with unknown constraints. Proceedings of the 32nd International Conference on Machine Learning, Lille, France; 2015.

[10] Frazier PI. A tutorial on Bayesian optimization. Entropy search for efficient global optimization of black-box functions. In Ghahramani, Z., Welling, M, Cortes, C., Lawrence, N.D., and Weinberger, K.Q. (Eds). Advances in Neural Information Processing Systems. 2018;27:918-926. arXiv. 1807.02811.

[11] Gutmann MU, Corander J. Bayesian optimization for likelihood-free inference of simulator-based statistical models. Journal of Machine Learning Research. 2015;16.

[12] Frean M, Boyle P. Using Gaussian processes to optimize expensive functions. In W. Wobcke and M. Zhang, Editors, AI 2008: Advances in Artificial Intelligence, Volume 5360 of Lecture Notes in Computer Science, Springer Berlin / Heidelberg. 2008;258-267.

[13] Rasmussen CE, Williams CKI. Gaussian processes for machine learning, adaptive computation and machine learning. MIT Press; 2006.

[14] Simpson TW, Mauery TM, Korte JJ, Mistree F. Kriging models for global approximation in simulation based multidisciplinary design optimization. AIAA Journal. 2001;39(12):2233-2241.

[15] Duvenaud D, Lloyd JR, Grosse R, Tenenbaum JB, Ghahramani Z. Structure discovery in nonparametric regression through compositional kernel search. Proceedings of the 30th International Conference on Machine Learning. 2013;28(3):1166-1174.

[16] Jones DR. A taxonomy of global optimization methods based on response surfaces. J. Global Optimization. 2001;21:345-383.

[17] Mockus J, Tiesis V, Zilinskas A. Toward global optimization, Volume 2, Chapter the application of Bayesian methods for seeking the extremum. Elsevier. 1978;117-128.

[18] Jones DR, Schonlau M, Welch WJ. Efficient global optimization of expensive black-box functions. J. Global Optimization. 1998;13(4):455-492.

[19] Cox DD, John S. A statistical method for global optimization. In Proceedings of the IEEE Conference on Systems, Man and Cybernetics; 1992.

[20] Cox DD, John S. SDO: A statistical method for global optimization. In Multidisciplinary Design Optimization: State-of-the-Art; 1997.

[21] Srinivas N, Krause A, Seeger M, Kakade SM. Gaussian process optimization in the bandit setting: No regret and experimental design. arXiv. 2010;0912.3995v2.

[22] Cover TM, Thomas JA. Elements of information theory. John Wiley & Sons; 2012.

[23] Frazier PI, Powell W, Dayanik S. The knowledge-gradient policy for correlated normal beliefs. Informs Journal on Computing. 2009;21(4):599-613.

[24] Nagakawa T, Osaki S. The discrete Weibull distribution. IEEE Transactions on Reliability. 1975;24(5).

[25] Haselimashhadi H, Vinciotti V, Yu K. A new Bayesian regression model for counts in medicine. arXiv. 2016;1601.02820 [stat.ME].

[26] Adesina OS, Olatayo TO, Agboola OO, Oguntunde PE. Bayesian Dirichet process mixture prior for count data. International Journal of Mechanical Engineering and Technology (IJMET). 2018;9(12): 630-646.

[27] Hastings W. Monte Carlo sampling methods using Markov chains and their application. Biometrika. 1970;57:97-109.

[28] Adesina OS, Agunbiade DA, Oguntunde PE, Adesina TF. Model for Bayesian zero truncated count data. Asian Journal of Probability and Statistics. 2019;4(1):1-12.

[29] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria; 2019.
Available:https://www.R-project.org

[30] Haselimashhadi H. BDWreg: Bayesian inference for discrete weibull regression. R Package Version 1.2.0; 2017.
Available:https://CRAN.R-project.org/package=BDWreg

[31] Yan Y. rBayesianOptimization: Bayesian optimization of hyperparameters. R package version 1.1.0; 2016.
Available:https://CRAN.R-project.org/package=rBayesianOptimization.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
http://www.sdiarticle4.com/review-history/54382