Title: Bayesian optimization

URL: https://en.wikipedia.org/wiki/Bayesian_optimization

PageID: 40973765

Categories: Category:Machine learning, Category:Sequential experiments, Category:Sequential methods, Category:Stochastic optimization

-----

Bayesian optimization is a sequential design strategy for global optimization of black-box functions, that does not assume any functional forms. It is usually employed to optimize expensive-to-evaluate functions. With the rise of artificial intelligence innovation in the 21st century, Bayesian optimizations have found prominent use in machine learning problems for optimizing hyperparameter values .

History

The term is generally attributed to Jonas Mockus [lt] and is coined in his work from a series of publications on global optimization in the 1970s and 1980s.

Early mathematics foundations

From 1960s to 1980s

The earliest idea of Bayesian optimization sprang in 1964, from a paper by American applied mathematician Harold J. Kushner, “A New Method of Locating the Maximum Point of an Arbitrary Multipeak Curve in the Presence of Noise” . Although not directly proposing Bayesian optimization, in this paper, he first proposed a new method of locating the maximum point of an arbitrary multipeak curve in a noisy environment. This method provided an important theoretical foundation for subsequent Bayesian optimization.

By the 1980s, the framework we now use for Bayesian optimization was explicitly established. In 1978, the Lithuanian scientist Jonas Mockus, in his paper “The Application of Bayesian Methods for Seeking the Extremum”, discussed how to use Bayesian methods to find the extreme value of a function under various uncertain conditions. In his paper, Mockus first proposed the Expected Improvement principle (EI) , which is one of the core sampling strategies of Bayesian optimization. This criterion balances exploration while optimizing the function efficiently by maximizing the expected improvement. Because of the usefulness and profound impact of this principle, Jonas Mockus is widely regarded as the founder of Bayesian optimization. Although Expected Improvement principle (EI) is one of the earliest proposed core sampling strategies for Bayesian optimization, it is not the only one, with the development of modern society, we also have Probability of Improvement (PI), or Upper Confidence Bound (UCB) and so on.

From theory to practice

In the 1990s, Bayesian optimization began to gradually transition from pure theory to real-world applications. In 1998, Donald R. Jones and his coworkers published a paper titled “Efficient Global Optimization of Expensive Black-Box Functions ”. In this paper, they proposed the Gaussian Process (GP) and elaborated on the Expected Improvement principle (EI) proposed by Jonas Mockus in 1978. Through the efforts of Donald R. Jones and his colleagues, Bayesian Optimization began to shine in the fields like computers science and engineering. However, the computational complexity of Bayesian optimization for the computing power at that time still affected its development to a large extent.

In the 21st century, with the gradual rise of artificial intelligence and bionic robots, Bayesian optimization has been widely used in machine learning and deep learning, and has become an important tool for Hyperparameter Tuning . Companies such as Google, Facebook and OpenAI have added Bayesian optimization to their deep learning frameworks to improve search efficiency. However, Bayesian optimization still faces many challenges, for example, because of the use of

Gaussian Process as a proxy model for optimization, when there is a lot of data, the training of Gaussian Process will be very slow and the computational cost is very high. This makes it difficult for this optimization method to work well in more complex drug development and medical experiments.

Strategy

Bayesian optimization is used on problems of the form max x ∈ X f (x) {\textstyle \max _{x\in X}f(x)} , with X {\textstyle X} being the set of all possible parameters x {\textstyle x} , typically with less than or equal to 20 dimensions for optimal usage (X → R d ■ d ≤ 20 {\textstyle X\rightarrow \mathbb {R} ^{d}\mid d\leq 20}), and whose membership can easily be evaluated. Bayesian optimization is particularly advantageous for problems where f (x) {\textstyle f(x)} is difficult to evaluate due to its computational cost. The objective function, f {\textstyle f} , is continuous and takes the form of some unknown structure, referred to as a "black box". Upon its evaluation, only f (x) {\textstyle f(x)} is observed and its derivatives are not evaluated.

Since the objective function is unknown, the Bayesian strategy is to treat it as a random function and place a prior over it. The prior captures beliefs about the behavior of the function. After gathering the function evaluations, which are treated as data, the prior is updated to form the posterior distribution over the objective function. The posterior distribution, in turn, is used to construct an acquisition function (often also referred to as infill sampling criteria) that determines the next query point.

There are several methods used to define the prior/posterior distribution over the objective function. The most common two methods use Gaussian processes in a method called kriging . Another less expensive method uses the Parzen-Tree Estimator to construct two distributions for 'high' and 'low' points, and then finds the location that maximizes the expected improvement.

Standard Bayesian optimization relies upon each x ∈ X {\displaystyle x\in X} being easy to evaluate, and problems that deviate from this assumption are known as exotic Bayesian optimization problems. Optimization problems can become exotic if it is known that there is noise, the evaluations are being done in parallel, the quality of evaluations relies upon a tradeoff between difficulty and accuracy, the presence of random environmental conditions, or if the evaluation involves derivatives.

Acquisition functions

Examples of acquisition functions include

probability of improvement

expected improvement

Bayesian expected losses

upper confidence bounds (UCB) or lower confidence bounds

Thompson sampling

and hybrids of these. They all trade-off exploration and exploitation so as to minimize the number of function queries. As such, Bayesian optimization is well suited for functions that are expensive to evaluate.

Solution methods

The maximum of the acquisition function is typically found by resorting to discretization or by means of an auxiliary optimizer. Acquisition functions are

maximized using a numerical optimization technique , such as Newton's method or quasi-Newton methods like the Broyden–Fletcher–Goldfarb–Shanno algorithm .

Applications

The approach has been applied to solve a wide range of problems, including learning to rank , computer graphics and visual design, robotics , sensor networks , automatic algorithm configuration, automatic machine learning toolboxes, reinforcement learning , planning, visual attention, architecture configuration in deep learning , static program analysis, experimental particle physics , quality-diversity optimization, chemistry, material design, and drug development.

Bayesian optimization has been applied in the field of facial recognition. The performance of the Histogram of Oriented Gradients (HOG) algorithm, a popular feature extraction method, heavily relies on its parameter settings. Optimizing these parameters can be challenging but crucial for achieving high accuracy. A novel approach to optimize the HOG algorithm parameters and image size for facial recognition using a Tree-structured Parzen Estimator (TPE) based Bayesian optimization technique has been proposed. This optimized approach has the potential to be adapted for other computer vision applications and contributes to the ongoing development of hand-crafted parameter-based feature extraction algorithms in computer vision.

See also

Multi-armed bandit

Kriging

Thompson sampling

Global optimization

Bayesian experimental design

Probabilistic numerics

Pareto optimum

Active learning (machine learning)

Multi-objective optimization

References

v

t

e

Golden-section search

Powell's method

Line search

Nelder–Mead method

Successive parabolic interpolation

Trust region

Wolfe conditions

Berndt–Hall–Hall–Hausman

Broyden–Fletcher–Goldfarb–Shanno and L-BFGS

Davidon–Fletcher–Powell

Symmetric rank-one (SR1)

Conjugate gradient

Gauss–Newton

Gradient

Mirror

Levenberg–Marquardt

Powell's dog leg method

Truncated Newton

Newton's method

Barrier methods

Penalty methods

Augmented Lagrangian methods

Sequential quadratic programming

Successive linear programming

Cutting-plane method

Reduced gradient (Frank–Wolfe)

Subgradient method

Affine scaling

Ellipsoid algorithm of Khachiyan

Projective algorithm of Karmarkar

Simplex algorithm of Dantzig

Revised simplex algorithm

Criss-cross algorithm

Principal pivoting algorithm of Lemke

Active-set method

Approximation algorithm

Dynamic programming

Greedy algorithm

Integer programming Branch and bound / cut

Branch and bound / cut

Bor■vka

Prim

Kruskal

Bellman–Ford SPFA

SPFA

Dijkstra

Floyd–Warshall

Dinic

Edmonds–Karp

Ford–Fulkerson

Push–relabel maximum flow

Evolutionary algorithm

Hill climbing

Local search

Parallel metaheuristics

Simulated annealing

Spiral optimization algorithm

Tabu search

Software