Bayessche Optimierung

Bayessche Optimierung ist eine globale Optimierungsmethode, also eine Methode mit der es möglich ist das globale Optimum einer Funktion zu bestimmen beziehungsweise zu approximieren. Sie ist für Zielfunktionen geeignet, zu denen keine geschlossene Form vorliegt, es jedoch möglich ist einzelne Datenpunkte gegebenenfalls mit Rauscheinfluss zu berechnen. Im Gegensatz zu vielen anderen Optimierungsverfahren ist die Bayessche Optimierung auch anwendbar, wenn es keine Möglichkeit gibt für diese sogenannten Black-Box-Funktionen Ableitungen zu bestimmen oder keine Konvexität vorliegt. Dabei gehört die Bayessche Optimierung zu den effizientesten Optimierungsverfahren in Bezug auf die benötigten Datenpunkte der Zielfunktion, wodurch sich dieses Verfahren besonders für Funktionen eignet, die eine zeitintensive Berechnung aufweisen [1]. Diese Tatsache bedeutet insbesondere, dass sich die Bayessche Optimierung besonders für die Hyperparameteroptimierung von maschinellen Lernverfahren eignet, da die Evaluation einzelner Hyperparameterwerte das gesamte Training des Lernverfahrens beinhaltet. Im Gegensatz zu anderen Optimierungsverfahren wird bei der Bayesschen Optimierung ein probabilistisches Modell für die Zielfunktion erstellt und dieses dann in einem iterativen Verfahren durch Hinzunahme weiterer Datenpunkte aktualisiert und verbessert. Hierbei ist der Grundgedanke, dass die gesamte Information aller berechneten Datenpunkte genutzt werden kann und nicht nur der lokale Gradient des letzten bestimmten Datenpunktes. Für die Erzeugung und Aktualisierung dieses probabilistischen Modells wird der Satz von Bayes verwendet. Dieser besagt in der grundlegenden Form, dass die A-posteriori Wahrscheinlichkeit eines Modells M bei gegebenen Daten D proportional zu der Likelihood von D gegeben M multipliziert mit der A-priori Wahrscheinlichkeit von M ist [1]:

$$P(M|D) \propto P(D|M)P(M) \tag{0.1}$$

Die A-priori Wahrscheinlichkeit beschreibt bei der Bayesschen Optimierung das Vorwissen über den Raum der möglichen Zielfunktionen. Liegt beispielsweise die Annahme vor, dass die Zielfunktion eher glatt verläuft, sollten Zielfunktionen mit hoher Varianz eine geringe A-priori Wahrscheinlichkeit aufweisen. Die A-posteriori

Wahrscheinlichkeit beschreibt dann entsprechend das aktualisierte probabilistische Modell der unbekannten Zielgerade. Ein weiterer entscheidender Aspekt der Bayesschen Optimierung ist die Verwendung einer sogenannten *acquisition function* zur Bestimmung des jeweils nächsten zu verwendenden Datenpunktes. Diese *acquisition function*, die auf unterschiedlichste Weise definiert werden kann, beschreibt eine Form der Nützlichkeit der Datenpunkte für das Modell. Durch eine Bestimmung des Maximums dieser Funktion wird der nächste zu verwendende Datenpunkt ausgewählt, wodurch eine geringe Anzahl benötigter Datenpunkte bei gleichzeitig gutem Ergebnis erreicht werden kann. Zur Veranschaulichung ist dieses Prinzip in Abbildung 0.1 dargestellt.

Um eine Bayessche Optimierung durchführen zu können, sind also zwei wichtige Bereiche zu beachten. Einerseits die Art der Modellierung der A-priori Wahrscheinlichkeit, für die meist Gaußprozesse verwendet werden und andererseits die Wahl der *acquisition function*. Diese beiden Aspekte sollen im Folgenden näher beleuchtet werden.

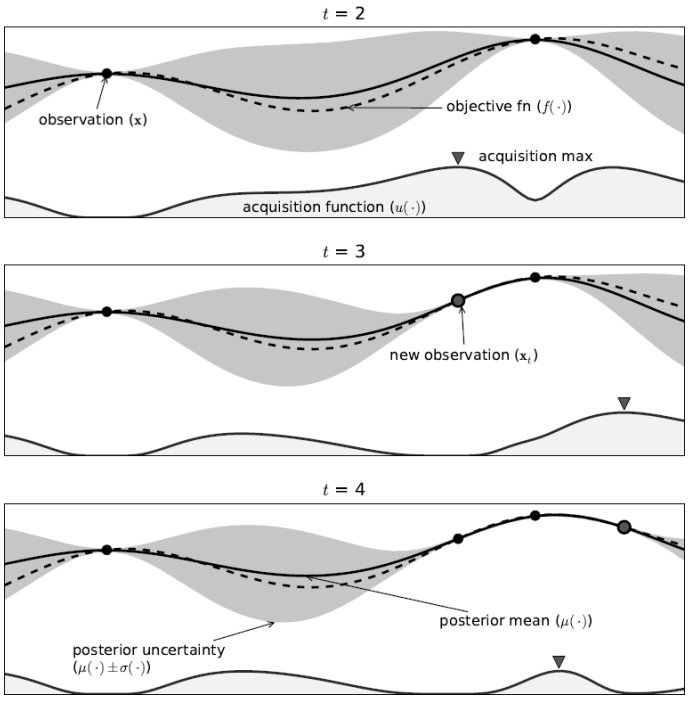

Abbildung 0.1.: Schematische Beispieldarstellung einer Bayesschen Optimierung eines einzelnen Parameters (Maximierung). Die Abbildung zeigt eine Gaußprozess-Approximation des Modells über vier Iterationen. Jeweils im unteren Bereich ist die zugehörige *acquisition function* dargestellt. [1]

Gaußprozesse

Ein Gaußprozess ist ein stochastischer Prozess, bei dem jede endliche Teilmenge von Zufallsvariablen mehrdimensional normalverteilt ist. Damit sind Gaußprozesse eine natürliche Generalisierung der mehrdimensionalen Normalverteilung. Analog zu einer mehrdimensionalen Normalverteilung, die durch den Erwartungswert und die Kovarianzmatrix vollständig und eindeutig bestimmt ist, ist ein Gaußprozess durch eine Erwartungswertfunktion $m(x)$ und eine Kovarianzfunktion $k(x, x')$ vollständig und eindeutig bestimmt [3]. Definiert werden sie mit

$$m(x) = \mathrm{E}[(f(x)] \tag{0.2}$$

und

$$k(x, x') = \mathrm{Cov}(f(x), f(x')) = \mathrm{E}\left[(f(x) - m(x))(f(x') - m(x'))\right] . \tag{0.3}$$

Der Gaußprozess wird dann notiert als

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')) . \tag{0.4}$$

Hierbei ist zu beachten, dass ein Gaußprozess im Gegensatz zu mehrdimensionalen Normalverteilungen keine Verteilung von Zufallsvariablen, sondern eine Verteilung von Funktionen darstellt. Für ein Verständnis dieses Konzepts ist es hilfreich, den Gaußprozess als Funktion zu betrachten, die für ein zufälliges x keinen Skalar, sondern den Erwartungswert und die Varianz einer Normalverteilung zurückgibt. Diese Normalverteilung beschreibt dann die Verteilung der möglichen Werte von f an der Stelle x. Ein Gaußprozess kann als A-priori Wahrscheinlichkeit des probabilistischen Modells der Bayesschen Optimierung verwendet werden. Mithilfe von Datenpunkten, zu denen die Funktionswerte bekannt sind, ist es dann möglich die A-posteriori Wahrscheinlichkeit des Modells zu berechnen und somit das Modell zu aktualisieren. Seien x_i, $i = 1, ..., n$ die bekannten Datenpunkte, $\boldsymbol{f}$ ein Vektor aller zugehörigen Funktionswerte $f(x_i)$ und $\boldsymbol{m}$ ein Vektor aller Erwartungswerte $m(x_i)$, dann ist der A-posteriori Gaußprozess gegeben durch [3]

$$\begin{aligned} f(x)|D \sim \mathcal{GP}(m_D(x), k_D(x, x')) , \\ m_D(x) = m(x) + \Sigma(X, x)^T \Sigma^{-1}(\boldsymbol{f} - \boldsymbol{m}) \\ k_D(x, x') = k(x, x') - \Sigma(X, x)^T \Sigma^{-1} \Sigma(X, x') , \end{aligned} \tag{0.5}$$

wobei $\Sigma(X, x)$ ein Vektor der Kovarianzen zwischen jedem bekannten Datenpunkt und x ist. Zur Erzeugung eines Gaußprozesses, der als A-priori Wahrscheinlichkeit genutzt werden soll, ist es also nötig eine Erwartungswertfunktion und eine Kovarianzfunktion zu definieren. Für die Erwartungswertfunktion wird dabei in den meisten Fällen eine Konstante gewählt [1]. Falls ein bekannter Trend vorliegt, ist es auch möglich diese anwendungsspezifische Struktur durch die Wahl eines niederdimensionalen Polynoms als Erwartungswertfunktion in dem Gaußprozess zu berücksichtigen. Für die Kovarianzfunktion wird in der Regel eine vordefinierte Kernelfunktion verwendet. Ein grundlegender Kernel ist der *automatic relevance determination* (ARD) *squared exponential* Kernel, der definiert ist durch [4]

$$k_{SE}(x, x') = \theta_0 \exp\left\{-\frac{1}{2} r^2(x, x')\right\} \tag{0.6}$$

mit

$$r^2(x, x') = \sum_{d=1}^{d'} \frac{(x_d - x'_d)^2}{\theta_d^2} \,. \tag{0.7}$$

Dabei ist d' die Anzahl der Dimensionen und θ_0 ein Parameter für die Amplitude der Varianz. Je niedriger der Wert von θ_0 ist, desto stärker ist die Begünstigung von Funktionen die nah an der Erwartungsfunktion liegen und umgekehrt. Die Parameter θ_1 bis $\theta_{d'}$ sind ein Maß für die Glätte der Funktionen, wobei jeder Parameter das Verhalten in einer Dimension beeinflusst. Da dieser Kernel in praktischen Optimierungsverfahren teilweise unrealistisch glatte Funktionen generiert, ist ein weiterer oft verwendeter Kernel der ARD Matérn-Kernel, bei dem dieses Problem einen geringeren Stellenwert einnimmt [4]. Der ARD Matérn-Kernel ist definiert durch [1]

$$k_M(x, x') = \theta_0 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu r^2}\right)^{\nu} K_\nu \left(\sqrt{2\nu r^2}\right) \,. \tag{0.8}$$

$\Gamma(\cdot)$ beschreibt dabei die Eulersche Gammafunktion und $K_\nu(\cdot)$ steht für die modifizierte Bessel-Funktion. ν stellt einen zusätzlichen Parameter dar, der als allgemeiner Glattheitsparameter interpretiert werden kann. Um eine leicht berechenbare Form des ARD Matérn-Kernels zu erzeugen, werden hierfür meist Standardwerte wie zum Beispiel $\nu = 3/2$ oder $\nu = 5/2$ gewählt.

Acquisition function

Bei dem Einsatz von Bayesscher Optimierung für die Hyperparameteroptimierung eines Lernverfahrens ist das Ziel immer eine Minimierung, da dabei stets versucht wird den Fehler des Lernverfahrens zu minimieren. Daher soll die *acquisition function* an Punkten hohe Werte aufweisen, an denen sehr wahrscheinlich ein niedriger Wert der zu optimierenden Funktion zu finden ist. Aus diesem Grund sind typische *acquisition functions* so formuliert, dass sich ein hoher Wert ergibt, wenn das geschätzte Modell an dieser Position einen geringen Wert aufweist, eine große Ungenauigkeit an dieser Position vorliegt, oder wenn beides zutrifft [1]. Sehr häufig wird für die *acquisition function* das *expected improvement* (EI) verwendet. Neben einem besonders wünschenswerten Verhalten bei Optimierungen, ist ein weiterer großer Vorteil des EI gegenüber anderen möglichen Funktionen, dass für die Definition keine festzulegenden Parameter benötigt werden [4]. Außerdem umfasst die Definition des EI nur wenige Schritte. Ist $f' = \min \boldsymbol{f}$ der bisher niedrigste beobachtete Wert, lässt sich die Verbesserung an einem Punkt x mit der Funktion

$$u(x) = \max\ (0, f' - f(x)) \tag{0.9}$$

beschreiben. Das EI ist dann die erwartete Verbesserung als eine Funktion von x und lässt sich notieren als [2, 5]

$$\begin{aligned} \mathrm{EI}(x)|D = \mathrm{E}\left[u(x)\right] &= \int_{-\infty}^{f'} (f' - f)\, \mathcal{N}\left(f; m_D(x), k_D(x,x)\right)\, \mathrm{d}f \\ &= (f' - m_D(x))\, \Phi(f'; m_D(x), k_D(x,x)) \\ &\quad + k_D(x,x)\, \mathcal{N}(f'; m_D(x), k_D(x,x))\,. \end{aligned} \tag{0.10}$$

Diese Notation veranschaulicht auch die Funktionsweise des EI. Der erste Term wird durch eine Verringerung der Erwartungswertfunktion $m_D(x)$ vergrößert und der zweite Term durch eine Erhöhung der Varianz $k_D(x,x)$. Hierbei wird also der Trade-off zwischen *exploitation* (Auswahl von Punkten mit einem niedrigen geschätzten Wert) und *exploration* (Auswahl von Punkten mit einer hohen Ungenauigkeit) explizit dargestellt. Dadurch wird deutlich, dass dieser Trade-off bei der Verwendung des EI automatisch berücksichtigt wird und keiner Einstellung durch einen Parameter bedarf. Der Punkt mit dem höchsten EI wird dann als nächster zu beobachtender Punkt ausgewählt. Um dieses Maximum der *acquisition function* zu bestimmen gibt es verschiedene effizient durchzuführende Methoden.

Literaturverzeichnis

[1] E. Brochu, V. M. Cora und N. de Freitas. *A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning*. 2010.

[2] D. R. Jones, M. Schonlau und W. J. Welch. „Efficient Global Optimization of Expensive Black-Box Functions“. In: *Journal of Global Optimization* 13.4 (1998), S. 455–492.

[3] C. E. Rasmussen und C. K. I. Williams. *Gaussian processes for machine learning*. Adaptive computation and machine learning. Cambridge, Massachusetts und London, England: The MIT Press, 2006.

[4] J. Snoek, H. Larochelle und R. P. Adams. „Practical Bayesian Optimization of Machine Learning Algorithms“. In: *Advances in Neural Information Processing Systems*. Neural Information Processing Systems 2012. 2012, S. 2951–2959.

Online-Ressourcen

[5] R. Garnett. *Bayesian Methods in Machine Learning – Spring 2019. Lecture notes - 06.03.2019*. Washington University in St. Louis. URL: `https://www.cse.wustl.edu/~garnett/cse515t/spring_2019/files/lecture_notes/12.pdf` (besucht am 16. 09. 2019).