and Bayesian Optimization

参考文献 [1] に基づいて parameter 最適化のアルゴリズムについてまとめる。

1 Sequential Model-based Global Optimization (SMBO)

評価に非常に時間がかかる予測モデル $f : \mathcal{X} \to \mathbb{R}$ の hyper parameter $\boldsymbol{x} \in \mathcal{X}$ の最適化を考えた場合、直感的な解決方法は予測モデル f の $\boldsymbol{x}$ に対する performance を低コストに予測できる適当な回帰モデル (response surface model と呼ばれることが多い) を作成し、それを用いて $\boldsymbol{x}$ の評価を行うことで最適化を行うことである。このような手法は model-based optimization と呼ばれるが、このモデル作成と予測値の収集の過程を繰り返していく手法が Sequential model-based optimization (SMBO) と呼ばれるものである。

以上の話からわかるように、model-based な最適化では以下の 2 点を適当に定めなければならない。

- Hyper parameter に対する予測にどのようなモデルを仮定するか
- そのモデルの評価にどのような指標 (関数) を用いるか

hyperopt や optuna では、予測確率モデルとして Tree-structured Parzen Estimator (TPE) を、評価指標として Expected Improvement [2] (EI)

$$\mathrm{EI}_{y^*}(x) \equiv \int_{-\infty}^{\infty} \max\left(y^* - y, 0\right) p_M\left(y|x\right) dy \tag{1}$$

を用いている。これは、f の予測モデル M の下での $f(x)$ の threshold y^* からの期待スコア改善量を表している*1。

なお、その他の評価指標として minimizing the Conditional Entropy of the Minimizer [3] や the bandit-based criterion [4] などがあるが、ここでは EI のみを考える。

2 Tree-structured Parzen Estimator Approach (TPE)

TPE approach は Bayes の定理

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}, \tag{2}$$

を用いて $p(y|x)$ を計算をする手法である。簡単のため、

$$p(x|y) = \begin{cases} l(x) & (y < y^*) \\ g(x) & (y \geq y^*) \end{cases} \tag{3}$$

とし、y^* を y の γ 分位点 $p(y < y^*) \equiv \gamma$ とおく。$l(x), g(x)$ はそれぞれ予測値が y^* 以下、以上になったものの parameter の分布を表す。

*1 一般に評価指標 y は小さいほどよいものが想定されている。

Input: Target algorithm $f : \boldsymbol{x} \to \mathbb{R}$, inital SMBO model M_0, number of trials T, surrogate function S
Output: Observation history $\mathcal{H}$ of parameters and performance $([(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots])$

$\mathcal{H} \leftarrow \emptyset$;
for t *in* T **do**
  $x^* \leftarrow \mathrm{argmin}_x S(\boldsymbol{\theta}, M_{t-1})$;
  $y \leftarrow f(\boldsymbol{x}^*)$ (Expensive step);
  $\mathcal{H} \leftarrow \mathcal{H} \cup (\boldsymbol{x}^*, y)$;
  $M_t \leftarrow \mathrm{FitModel}(\mathcal{H})$
end
return $\mathcal{H}$

Algorithm 1: Algorithm framework of generic Sequential Model-Based Optimization (SMBO)

このとき EI は

$$\begin{aligned}
\mathrm{EI}_{y^*}(x) &= \int_{-\infty}^{y^*} (y^* - y) p\,(y|x)\,dy, && (4)\\
&= \int_{-\infty}^{y^*} (y^* - y) \frac{p(x|y)p(y)}{p(x)} dy, && (5)\\
&= \frac{l(x)\left[\gamma y^* - \int_{-\infty}^{y^*} y p(y) dy\right]}{\gamma l(x) + (1-\gamma) g(x)}, && (6)\\
&\propto \left(\gamma + (1-\gamma)\frac{g(x)}{l(x)}\right)^{-1}. && (7)
\end{aligned}$$

と変形できることがわかる。従って、EI を最大化するには良い予測値の parameter による分布の値 $l(x)$ を大きく、逆に悪いものの $g(x)$ の値を小さくなるような x を取れば良い。

$l(x), g(x)$ を予測する手法の一つとして Parzen Estimator (カーネル密度推定法とも呼ぶ) がある。カーネル密度推定法は未知の母集団からの sampling とみなせるデータ点の集合から母集団の確率密度分布を推定するノンパラメトリック手法の一つで、$x_i (i = 1, \cdots, n)$ を確率密度 $p(x)$ からの独立な標本としたとき、$p(x)$ は以下のように表される。

$$p(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{8}$$

つまり、標本まわりにある band width h でカーネル関数 K を仮定し、それを足し合わせて平均して全体の確率密度関数とする方法である。カーネルとしては標準ガウス関数

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \tag{9}$$

が使われることが多い。

付録 A　The Gaussian Process Approaches (GP)

ある入力 $\boldsymbol{x}$ から出力 y を予測する最も一般的な確率モデルはガウス過程 (Gaussian process) である。

定義. 確率過程 $\{\boldsymbol{X}_t; t \in T\}$ がガウス過程であるとは、任意の finite subset $\boldsymbol{X}_{t_1,\cdots,t_k} = (\boldsymbol{X}_{t_1}, \cdots, \boldsymbol{X}_{t_k})$ の同時確率分布が多変量正規分布に従うことをいう。これは $(\boldsymbol{X}_{t_1}, \cdots, \boldsymbol{X}_{t_k})$ の任意の線形結合がガウス分布に従うことと同義である。

ガウス過程を用いた回帰では、以下のような手順である入力 $\boldsymbol{x}$ に対する出力の予測を行う。

1. 予測モデルの出力 $\boldsymbol{f}(X) = (f(\boldsymbol{x}_1), f(\boldsymbol{x}_2) \cdots)$ がガウス過程に従うと仮定する
2. 平均 $\boldsymbol{m}_0(X)$、分散共分散行列 $K(X,X) = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ を適当に定めることで $\boldsymbol{f}$ の事前分布 $p(\boldsymbol{f}|X) = N(\boldsymbol{f}|\boldsymbol{m}_0(X), K(X,X))$ を定める
3. 定めた事前分布からランダムに $\boldsymbol{f}$ を生成する
4. 生成した $\boldsymbol{f}$ を用いて $\boldsymbol{y} = \boldsymbol{f}(X)(+\epsilon)$ を計算する
5. Bayes の定理を用いて $\boldsymbol{y}$ が観測された元での $\boldsymbol{y}^*$ の事後分布を求める

$$p(\boldsymbol{y}^*|\boldsymbol{y}) = \frac{p(\boldsymbol{y}, \boldsymbol{y}^*)}{p(\boldsymbol{y})}$$

6. 新たな parameter X^* を事後分布に代入することで次の予測値の分布を計算する

なお、典型的な分散共分散行列としてはガウスカーネル

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\beta|\boldsymbol{x}_i - \boldsymbol{x}'_j|^2\right), \tag{10}$$

を用いることが多い。

EI を用いる場合は、y^* に observation 内の最小値

$$y^* = \min\left(f(x_i), 1 \leq i \leq n\right), \tag{11}$$

を用いる。

参考文献

[1] J. Bergstra, R. Bardenet, Y. Bengio and B. Kégl, *Algorithms for hyper-parameter optimization*, in *Advances in Neural Information Processing Systems* (J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira and K. Q. Weinberger, eds.), vol. 24, Curran Associates, Inc., 2011.

[2] D. R. Jones, *A taxonomy of global optimization methods based on response surfaces*, *Journal of Global Optimization* **21** (2001) 345–383.

[3] J. Villemonteix, E. Vázquez and E. Walter, *An informational approach to the global optimization of expensive-to-evaluate functions*, *CoRR* **abs/cs/0611143** (2006) [`arXiv:cs/0611143`].

[4] N. Srinivas, A. Krause, S. M. Kakade and M. W. Seeger, *Gaussian process bandits without regret: An experimental design approach*, *CoRR* **abs/0912.3995** (2009) [`arXiv:0912.3995`].