Tree-structured Parzen estimator

2020-07-30

하창우

Sequential model-based optimization (SMBO) methods

Sequential Model-Based Optimization (SMBO)

1. Gaussian Process
2. Random forest regression
3. ***Tree-structured Parzen estimator***

Acquisition function

Expected Improvement

Tree-structured Parzen estimator

Expected improvement is the expectation under some model M of f : X → R^N that f(x) will exceed (negatively) some threshold y∗

$$\mathrm{EI}_{y^*}(x) := \int_{-\infty}^{\infty} \max(y^* - y, 0) p_M(y|x) dy.$$

The TPE defines p(x|y) using two such densities:

$$p(x|y) = \begin{cases} \ell(x) & \text{if } y < y^* \\ g(x) & \text{if } y \geq y^*, \end{cases}$$

Tree-structured Parzen estimator

현재 데이터의 metric에 대해 정렬 후 y^*를 기준으로 분리해서 분포 l(x), g(x) 생성

$$p(x|y) = \begin{cases} \ell(x) & \text{if } y < y^* \\ g(x) & \text{if } y \geq y^*, \end{cases}$$

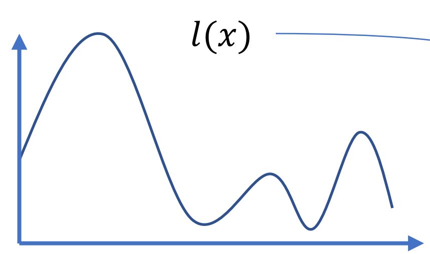
Sorted by y

X (parameter)	Y (target)
0.54	0.01
0.26	0.02
...	...
0.67	0.73

$(\gamma = p(y < y^*))$

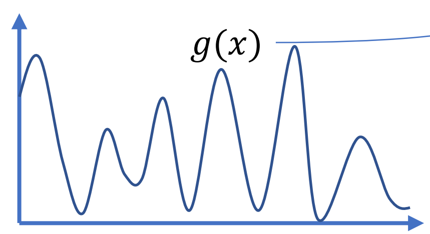
γ

x	y
0.54	0.01
0.26	0.02
...	...



$1-\gamma$

x	y
...	...
...	...
0.67	0.73



Expected improvement에서 next x^*를 구할 때 사용

Tree-structured Parzen estimator

$$\mathrm{EI}_{y^*}(x) = \int_{-\infty}^{y^*} (y^* - y) p(y|x) dy = \int_{-\infty}^{y^*} (y^* - y) \frac{p(x|y)p(y)}{p(x)} dy$$

By construction, $\gamma = p(y < y^*)$ and $p(x) = \int_{\mathbb{R}} p(x|y)p(y)dy = \gamma\ell(x) + (1-\gamma)g(x)$.

$$\int_{-\infty}^{y^*} (y^* - y) p(x|y) p(y) dy \quad = \quad \ell(x) \int_{-\infty}^{y^*} (y^* - y) p(y) dy = \gamma y^* \ell(x) - \ell(x) \int_{-\infty}^{y^*} p(y) dy,$$

$$\Rightarrow \; EI_{y^*}(x) = \frac{\gamma y^* \ell(x) - \ell(x) \int_{-\infty}^{y^*} p(y) dy}{\gamma \ell(x) + (1-\gamma) g(x)} \propto \left(\gamma + \frac{g(x)}{\ell(x)}(1-\gamma)\right)^{-1}$$

⇒ g(x)가 낮고 l(x)가 큰 x를 다음 탐색 후보로 지정。

Tree structured Parzen estimator

References

Algorithms for Hyper-Parameter Optimization
https://papers.nips.cc/paper/4443-algorithms-for-hyper-parameter-optimization.pdf

TPE: how hyperopt works
https://www.youtube.com/watch?v=tdwgR1AqQ8Y

Appendix

4.1 Optimizing EI in the TPE algorithm

The parametrization of $p(x, y)$ as $p(y)p(x|y)$ in the TPE algorithm was chosen to facilitate the optimization of EI.

$$\mathrm{EI}_{y^*}(x) = \int_{-\infty}^{y^*} (y^* - y)p(y|x)dy = \int_{-\infty}^{y^*} (y^* - y)\frac{p(x|y)p(y)}{p(x)}dy \tag{3}$$

By construction, $\gamma = p(y < y^*)$ and $p(x) = \int_{\mathbb{R}} p(x|y)p(y)dy = \gamma\ell(x) + (1-\gamma)g(x)$. Therefore

$$\int_{-\infty}^{y^*} (y^* - y)p(x|y)p(y)dy \quad = \quad \ell(x)\int_{-\infty}^{y^*} (y^* - y)p(y)dy = \gamma y^* \ell(x) - \ell(x)\int_{-\infty}^{y^*} p(y)dy,$$

so that finally $EI_{y^*}(x) = \frac{\gamma y^* \ell(x) - \ell(x)\int_{-\infty}^{y^*} p(y)dy}{\gamma\ell(x) + (1-\gamma)g(x)} \propto \left(\gamma + \frac{g(x)}{\ell(x)}(1-\gamma)\right)^{-1}$. This last expression shows that to maximize improvement we would like points x with high probability under $\ell(x)$ and low probability under $g(x)$. The tree-structured form of ℓ and g makes it easy to draw many candidates according to ℓ and evaluate them according to $g(x)/\ell(x)$. On each iteration, the algorithm returns the candidate x^* with the greatest EI.

Appendix

Surrogate model

The surrogate function, also called the response surface, is the probability representation of the objective function built using previous evaluations.
(https://towardsdatascience.com/a-conceptual-explanation-of-bayesian-model-based-hyperparameter-optimization-for-machine-learning-b8172278050f)

Nonparametric Density

Instead of assuming a parametric model for the distribution (e.g. Normal distribution with unknown expectation and variance), we rather want to be “as general as possible”: that is, we only assume that the density exists and is suitably smooth (e.g. differentiable).
https://stat.ethz.ch/education/semesters/SS_2006/CompStat/sk-ch2.pdf