Bayesian Optimization[1][2]

Introduction

1. Consider hyperparameter tuning for a ML model
2. Find hyperparameters that minimizes the test error
3. Can think of this as minimizing the test error function w.r.t hyperparameters
   - we do not know the function(black box access)
   - therefore cannot use the gradient, Hessian
4. Each **evaluation is expensive** in the sense that the number of evaluations that may be performed is limited, typically to a few hundred. This limitation typically arises because each evaluation takes a substantial amount of time (typically hours), but may also occur because each evaluation bears a monetary cost (e.g., from purchasing cloud computing power, or buying laboratory materials), or an opportunity cost (e.g., if evaluating f requires asking a human subject questions who will tolerate only a limited number)

Formulation

Suppose we have a function $f : \mathcal{X} \longrightarrow R$ that we wish to minimize on some domain $X \subseteq \mathcal{X}$. That is, we wish to find

$$x^* = argmin_{x \in X} f(x)$$

A common approach to optimization problems is to make some assumptions about f. *Bayesian Optimization* procees by maintaining a probabilistic belief about f and designing an *acquisition function* to determine where to evaluate the function next.
Bayesian optimization is particularly well-suited to global optimization problems where f is an expensive black-box function. In brief

- BO use past queries and function estimate + uncertainty to decide where to query next.
- Next query pt. is chosen using an acquisition fn. $a(x)$.

For this BO requires 2 ingridents

- **Regression Model**: to learn the fn. given the current set of query-value pairs $\{x_n, f(x_n)\}_{n=1}^{N}$. Although not strictly required, BO almost always reasons about f by choosing an appropriate Gaussian process prior:

  $$p(f) = \mathcal{GP}(f; \mu, K)$$

  Given observations $\mathcal{D}$, we can condition our distribution on $\mathcal{D}$ as usual:

  $$p(f|\mathcal{D}) = \mathcal{GP}\left(f; \mu_{f|\mathcal{D}}, K_{f|\mathcal{D}}\right)$$

- **Acquisition fn**: $a(x)$ tells the utility of any future pt. x. The acquisition function is typically an inexpensive function that can be evaluated at a given point that is commensurate with how desirable evaluating f at x is expected to be for the minimization problem.

Acquisition Functions

Some acquisition functions

1. **Probability Improvement**: Suppose

$$f' = \min f$$

is the minimal value of f observed so far. PI evaluates f at the point most likely to improve upon this value. This corresponds to the following utility function associated with evaluating f at a given point x:

$$u(x) = \left\{ \begin{array}{ll} 0, & \text{for } f(x) > f' \\ 1, & \text{for } f(x) \leq f' \end{array} \right\}$$

That is, we receive a unit reward if $f(x)$ turns out to be less that f', and no reward otherwise. The PI acquisition function is then the expected utility as a function of x

$$\begin{aligned} a_{PI}(x) &= E\left[u(x)|x, \mathcal{D}\right] \\ &= \int_{-\infty}^{f'} \mathcal{N}(f; \mu(x), K(x,x))df \\ &= \Phi(f'; \mu(x), K(x,x)) \end{aligned}$$

The point with the highest probability of improvement is selected.

2. **Expected Improvement**: The loss function associated with probability of improvement is somewhat odd: we get a reward for improving upon the current minimum independent of the size of the improvement! This can sometimes lead to odd behavior, and in practice can get stuck in local optima and underexplore globally.
We define a new acquisition fn.

$$u(x) = \left\{ \begin{array}{ll} 0, & \text{for } f(x) > f' \\ f' - f(x), & \text{for } f(x) \leq f' \end{array} \right\}$$

The EI acquisition function is then the expected utility as a function of x

$$\begin{aligned} a_{EI}(x) &= E\left[u(x)|x, \mathcal{D}\right] \\ &= \int_{-\infty}^{f'} (f' - f)\mathcal{N}(f; \mu(x), K(x,x))df \\ &= (f' - \mu(x))\Phi(f'; \mu(x), K(x,x)) + K(x,x)\mathcal{N}(f; \mu(x), K(x,x)) \end{aligned}$$

The point with the highest expected improvement is selected.

> **Significance** The expected improvement has two components. The first can be increased by reducing the mean function $\mu(x)$. The second can be increased by increasing the variance $K(x,x)$. These two terms can be interpreted as explicitly encoding a tradeoff between exploitation (evaluating at points with low mean) and exploration (evaluating at points with high uncertainty).

3. **Lower Confidence Bound(LCB)**: takes into account exploitation vs exploration. Assuming the posterior predictive for a new pt. $x = \mathcal{N}(\mu(x), K(x,x))$ then LCB is

$$a_{LCB}(x) = -(\mu(x) - \alpha\sqrt{K(x,x)})$$

where α is parameter of trade-off b/w mean and variance. **Theoretic result**: Under certain conditions, the iterative application of this acquisition fn. will converge to global optima of f.

References

[1] Eric Brochu, Vlad M Cora, and Nando De Freitas. “A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning”. In: *arXiv preprint arXiv:1012.2599* (2010).

[2] Peter I Frazier. “A tutorial on Bayesian optimization”. In: *arXiv preprint arXiv:1807.02811* (2018).