Tree-Structured Parzen Estimator with Inequality Constraints for Expensive Hyperparameter Optimization

Zharov Georgiy

MIPT

27 декабря 2023 г.

- Bayesian Optimization
- Standard TPE
- Bayesian Optimization with Unknown Constraints
- Naive TPE extension
- Problems of naive extension
- c-TPE

Bayesian Optimization

Suppose we would like to minimize a validation loss metric $f(x) = L(x, A, D_{train}, D_{val})$ of a supervised learning algorithm A given training and validation D_{train}, D_{val}, then the HPO problem is defined as follows:

$$x_{opt} \in arg \min_{x \in X} f(x)$$

$x \in X$ is a hyperparameter configuration

Bayesian Optimization

A common choice for acquisition functions is expected improvement or probability of improvement:

$$EI_{f^*}[x|D] = \int_{-\infty}^{f^*} (f - f^*) p(f|x, D) df$$

$$P[f \leq f^*|x, D] = \int_{-\infty}^{f^*} p(f|x, D) df$$

Tree-Structured Parzen Estimator

Let

$$p(x|f,D) = \begin{cases} p(x|D^{(l)}) & (f \leq f_\gamma) \\ p(x|D^{(g)}) & (f > f_\gamma) \end{cases}$$

where $D^{(l)}, D^{(g)}$ are the observations with $f_n \leq f_\gamma$ and $f > f_\gamma$. Then

$$EI_{f^*}[x|D] \sim r(x|D) := \frac{p(x|D^{(l)})}{p(x|D^{(g)})}$$

Why Tree-Structured?

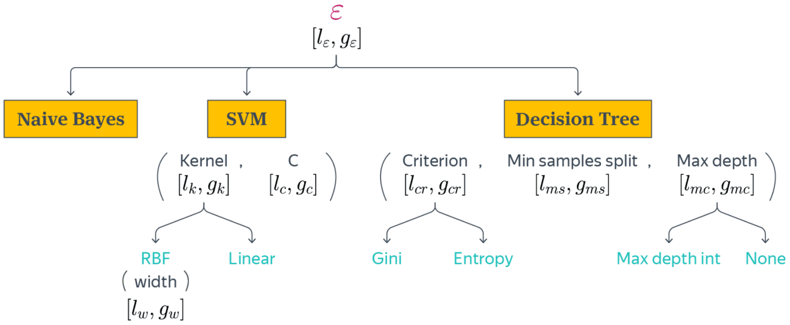

Рис.: Example of HPO process

Bayesian Optimization with Unknown Constraints

Consider unknown constraints $c_i i(x) = C_i(x, A, D_{train}, D_{val})$. Then the optimization is formulated as follows:

$$x_{opt} \in arg \min_x f(x)$$

$$s.t. \forall i \in \{1, ..., C\}, c_i(x) \leq c^*$$

And acquisition function:

$$ECI_{f^*}[x|c^*, D] = EI_{f^*}[x|D] \prod_{i=1}^{C} P(c_i \leq c_i^*|x, D)$$

Naive Extension

Naive extension of TPE can be implemented as the following algorithm:

- Pick the γ-quantile best objective value F^* in D,
- Split D into $D_0^{(l)}$ and $D_0^{(g)}$ at f^*, and D into $D_i^{(l)}$ and $D_i^{(g)}$ for $i \in \{1, ..., C\}$,
- Build kernel density estimators $p(x|D_i^{(l)}), p(x|D_i^{(g)})$ for $i \in \{0, ..., C\}$,
- Take the product of density ratios $\prod_{i=0}^{C} r_i(x|D) := \prod_{i=0}^{C} p(x|D_i^{(l)})/p(x|D_i^{(g)})$ as the AF.

Problems of Naive Extension

1. Vanished Constraints
2. Small Overlaps in Top and Feasible Domains

Constrained TPE

Algorithm 1 c-TPE algorithm (With modifications)

N_{init} (The number of initial configurations), N_s (The number of candidates to consider in the optimization of the AF)

$\mathcal{D} \leftarrow \emptyset$

for $n = 1, \dots, N_{\text{init}}$ **do**

  Randomly pick $\boldsymbol{x}$

  $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\boldsymbol{x}, f(\boldsymbol{x}), c_1(\boldsymbol{x}), \dots, c_C(\boldsymbol{x}))\}$

while Budget is left **do**

  $\mathcal{S} = \emptyset$

  for $i = 0, \dots, C$ **do**

    Split $\mathcal{D}$ into $\mathcal{D}_i^{(l)}$ and $\mathcal{D}_i^{(g)}$, $\hat{\gamma}_i \leftarrow |\mathcal{D}_i^{(l)}|/|\mathcal{D}|$

    Build $p(\cdot|\mathcal{D}_i^{(l)}), p(\cdot|\mathcal{D}_i^{(g)})$

    $\{\boldsymbol{x}_j\}_{j=1}^{N_s} \sim p(\cdot|\mathcal{D}_i^{(l)}), \mathcal{S} \leftarrow \mathcal{S} \cup \{\boldsymbol{x}_j\}_{j=1}^{N_s}$

  ▷ See Appendix D for the hard-constrained version

  Pick $\boldsymbol{x}_{\text{opt}} \in \text{argmax}_{\boldsymbol{x} \in \mathcal{S}} \prod_{i=0}^{C} r_i^{\text{rel}}(\boldsymbol{x}|\mathcal{D})$

  $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\boldsymbol{x}_{\text{opt}}, f(\boldsymbol{x}_{\text{opt}}), c_1(\boldsymbol{x}_{\text{opt}}), \dots, c_C(\boldsymbol{x}_{\text{opt}}))\}$

Some results

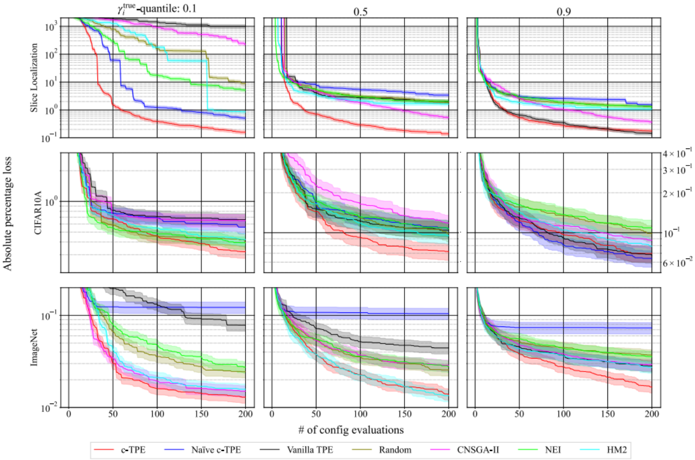

Рис.: Results

Literature

- c-TPE: Tree-Structured Parzen Estimator with Inequality Constraints for Expensive Hyperparameter Optimization
- https://education.yandex.ru/handbook/ml/article/podbor-giperparametrov