Hyperparameter Tuning

Problem:

- Minimize a black box function $f(x_1, \ldots, x_d)$
- Query mode, no explicit form
- The x_i are hyperparameters, could be discrete or continuous

Different techniques

- Bayesian Optimization
- Gradient descent
- Random Search
- Multi-armed Bandit based algorithms
- Grid Search

Bayesian Optimization

- A sequential algorithm (hard to parallelize, which is very important in hyperparameter tuning)

Procedures:

1. Assume a prior distribution for the loss function
2. Select new samples that balance exploration and exploitation
3. Update the prior with the new samples using Bayes' rule

- Tools: Spearmint
- limitation: does not work well for high dimensional hyperparameters space

Gradient Descent

A simple example for illustration:

- Linear regression: $L(w) = \frac{1}{2}\sum_{i=1}^{n}(w^T x - y)^2$
- Do gradient descent for only two steps:
  - $w_2 = w_1 - \eta \nabla_w L(w_1)$
  - $w_1 = w_0 - \eta_w L(w_0)$
  - $f(w_0, \eta) = L(w_2)$, we need to compute $\nabla_\eta f(w_0, \eta)$
- Define momentum $v_t = \gamma v_{t-1} - (1-\gamma)\nabla_w L(w, \theta, t)$
- v_t store compressed information of $w_1, \ldots, w_T$.

Multi-Armed Bandit

- π arms, each gives a reward (bounded random variable with expectation v_i)

Successive Halving algorithm

Algorithm 1 Successive Halving

Input: budget B

$S_0 \leftarrow [n]$

Per round budget $B' \leftarrow \frac{B}{\log_2(n)}$

for $r = 0$ to $\log_2(n) - 1$ **do**

  Sample each arm $i \in S_r$ for $\frac{B'}{|S_r|}$ times

  Let S_{r+1} be the set of $|S_r|/2$ arms in S_r with the largest empirical average

end for

Output: $S_{\log_2(n)}$

Theoretical Guarantee

- Assume $v_1 > v_2 \ge \ldots \ge v_n$ and $\Delta_i = v_1 - v_i$
- The algorithm finds the optimal solution with probability of $1 - \delta$ within $B = O(H_2 \log n \log(\frac{\log n}{\delta}))$, where $H_2 = \max_{i \ge 1} \frac{i}{\Delta_i^2}$

Proof:

- **Concentration Inequality**: $\frac{B}{|S_r| \log n}$ sampling times for each $i \in S_r$ for round r. Then

$$\Pr(\hat{v}_1 \le \hat{v}_i) \le e^{-\frac{1}{2}\frac{B\Delta_i^2}{|S_i| \log n}} \tag{1}$$

- Let $n_r = \frac{n}{2^{r+2}}$, so in round r we have $4n_r$ left. Denote the smaller $3n_r$ arms by S_r'.
- Let N_r be the number of arms with empirical mean larger than arm 1, and also in S_r'.

$$\mathbb{E}[N_r] = \sum_{i \in S_r'} e^{-\frac{1}{2}\frac{B\Delta_i^2}{|S_r| \log n}} \le |S_r'| e^{-\frac{B}{8 \log n}\frac{\Delta_{n_r}^2}{n_r}} \tag{2}$$

- Then by Markov inequality, with high probability there are not so many bad arms with empirical mean larger than arm 1

$$\Pr[N_r > \frac{1}{3}|S_r'|] \le 3e^{-\frac{B}{8 \log n}\frac{\Delta_{n_r}^2}{n_r}} \tag{3}$$

- This means we will have $\frac{1}{3} \times 3n_r = n_r$ good arms in S_r' and also n_r good arms in $S_r - S_r'$.
- Then the probability that the arm 1 got removed in any round is at most

$$3e^{-\frac{B}{8 \log n}\frac{\Delta_{n_r}^2}{n_r}} \cdot \log n = 3 \log n e^{-\frac{B}{8H_2 \log n}} \tag{4}$$

-

Applications to Hyperparameters tuning

- Each configuration is an arm
- However, we are not drawing random variables, but we only care about the last observed value
- For all $i \in [n]$, $k \ge 1$, let $\ell_{i,k}$ be a sequence for arm i, assuming $v_i = \lim_{\tau \to \infty} \ell_{i,\tau}$.

Algorithm 2 Successive Halving

Require: budget B

$S_0 \leftarrow [n]$

Per round budget $B' \leftarrow \frac{B}{\log_2(n)}$

for $r = 0$ to $\log_2(n) - 1$ **do**

  Pull each arm $i \in S_r$ for $\frac{B'}{|S_r|}$ times, get the current value ℓ_{i,k_i}.

  Let S_{r+1} be the set of $|S_r|/2$ arms in S_r with the smallest ℓ_{i,k_i}

end for

Ensure: $S_{\log_2(n)}$

Theoretical Guarantee

- Let $\gamma_i(t)$ be non-increasing function of t, which gives the smallest value for each t s.t. $|\ell_{i,t} - v_i| \leq \gamma_i(t)$.
    - "envelope" of the curve
- Let $\gamma_i^{-1}(\alpha) = \min\{t \in N : \gamma(t) \leq \alpha\}$
    - First time we are α-close to v_i
    - If $k_i \geq \gamma_i^{-1}(\frac{v_1 - v_i}{2}), k_1 \geq \gamma_1^{-1}(\frac{v_1 - v_i}{2})$, then arm 1 and arm i are separated.