Adaptive Model Selection in Deep Learning

May 4, 2017

1 Problem setup

The user has his own dataset D. There are K candidate algorithms $\{1, 2, \cdots, K\}$ (for example, K different network types in deep learning). We use $\mathcal{A}_k(D) \in \mathbb{R}$ to denote the performance of $\mathcal{A}_k$ on D and $\mathcal{A}(D) \in \mathbb{R}^K$ to denote the performance of all algorithms on D. The goal is to find out the best algorithm for his dataset D via as few number of experiments as possible. For simplicity, we can assume that the outcome of applying an algorithm to the data set D is deterministic.

If K algorithms are totally independent, then it is a bandit problem. We need at least K experiments to find out the best algorithm, even all algorithms are deterministic. If we know the correlation of all algorithms, then we could use much fewer number of experiments to identify the best algorithm. Here, we can assume that the performance of K algorithms is a Gaussian process $GP(\mu, \Sigma)$. with

$$\mu := \mathbb{E}_D[\mathcal{A}(D)] \in \mathbb{R}^K \quad \Sigma := \mathbb{E}_D[(\mathcal{A}(D) - \mu)(\mathcal{A}(D) - \mu)^\top] \in \mathbb{R}^{K \times K}.$$

Given a data set D, the performance of K algorithms is a sample generated from this Gaussian process.

2 Algorithm

We apply the GP-UCB algorithm in Srinivas et al. [2009].

Notations:

- $[K] = \{1, 2, \cdots, K\}$.
- $a_{[1:t]} = \{a_1, \cdots, a_t\}$, where $a_i \in [K]$, for $1 \leq i \leq t$.
- $y_{[1:t]} = \{y_1, \cdots, y_t\}$.
- $\Sigma(i, j) = \Sigma_{i,j}$, for $i, j \in [K]$.
- $\Sigma_t(k) = [\Sigma(a_1, k), \cdots \Sigma(a_t, k)]^\top, k \in [K]$.
- $\Sigma_t = [\Sigma(i, j)]_{i,j \in a_{[1:t]}}$.
- $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$.

Algorithm 1 GP-UCB

```
Require: GP prior $\mu_0$, $\sigma$, $\Sigma$, and $\delta \in (0,1)$
Ensure: Return the best algorithm among all algorithms in $a_{[1:T]}$
  Initialize $\sigma_0 = \mathrm{diag}(\Sigma)$
  for $t = 1, 2, \cdots, T$ do
    Update $\beta_t$ by
      $$\beta_t = \log(Kt^2/\delta)$$
    Choose
      $$a_t = \arg\max_{k \in [K]} \mu_{t-1}(k) + \sqrt{\beta_t}\sigma_{t-1}(k)$$
    Observe
      $$y_t = \mathcal{A}_{a_t} + \epsilon_t$$
    Update $\mu_t$ by
      $$\mu_t(k) = \Sigma_t(k)^\top (\Sigma_t + \sigma^2 I)^{-1} y_{[1:t]}$$
    Update $\sigma_t$ by
      $$\sigma_t^2(k) = \Sigma(k,k) - \Sigma_t(k)^\top (\Sigma_t + \sigma^2 I)^{-1} \Sigma_t(k)$$
  end for
```

References

N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.