

【强化学习 87】Nonparametric Bandits



张楚珩

清华大学 交叉信息院博士在读

7 人赞同了该文章

文章讲的是 non-parametric stochastic contextual bandits。

原文传送门

Guan, Melody Y., and Heinrich Jiang. "Nonparametric stochastic contextual bandits." Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

特色

问题设定: stochastic contextual bandits 以及 infinite armed contextual bandits

针对任务: top-arm identification 以及 regret minimization

算法: 前一个任务用 uniform sampling; 后一个任务用 UCB 方法。两者都结合 k-NN 来估计均值。

结果: 前一个任务需要 $\tilde{O}(e^{-(\alpha+\beta)})$ 多的样本, D 是 context 空间的维度; 后一个任务有 sublinear bound, 并且该 bound 可以只和数据真实维度有关。

过程

1. k-NN 方法的估计误差

Theorem 1. (Rate for k-NN (Jiang 2017b)) Let $\delta > 0$. There exists N_0 and universal constant C such that if $n \geq N_0$ and $k = \lfloor n^{2/(2+D)} \rfloor$, then with probability at least $1 - \delta$,

$$\sup_{x \in \mathcal{X}} |f(x) - \hat{f}_{k\text{-NN}}(x)| \leq C \sqrt{\log n \log(1/\delta)} \cdot n^{-1/(2+D)}$$

知乎 @张楚珩

其中 k-NN 估计把最近的 k 个样本做平均, 样本是从真实分布外加一个 sub-Gaussian noise 得到的。 D 表示 context space 的维度, n 表示样本数。

2. Top-arm identification

一个简单的方法, 就是给定 T 步之后, 每个 arm 玩相同的次数。文章证明了, 在这样的操作下, 需要 $\tilde{O}(e^{-(\alpha+\beta)})$ 多的样本使得 k-NN 估计得到的最优 arm 的均值比真实最优 arm 的均值差距大概率小于 ϵ 。

3. Regret for UCB strategy

考虑如下 UCB 算法

Algorithm 2 Upper Confidence Bound (UCB)

- 1: Parameters: M_0, M_1, δ, T .
 - 2: Define $\sigma(n) = M_1 \sqrt{\log n (\log(nK/\delta))} \cdot n^{-1/(2+D)}$.
 - 3: Pull each of the K arms M_0 times.
 - 4: For each round $t = KM_0, KM_0 + 1, \dots, T$:
 - 5: Pull $I_t := \operatorname{argmax}_{i \in [K]} \hat{f}_i(t) + \sigma(T_i(t-1))$
-

其 regret bound 为 $\tilde{O}(\frac{1}{\delta^{1/(2+D)}})$ 。

Theorem 3. Let $\delta > 0$. Suppose that $M_0 \geq N_0$ and $M_1 > C$ in Algorithm 2. Then we have that with probability at least $1 - \delta$,

$$R_T \leq M_1 2^{\frac{1+D}{2+D}} K \sqrt{\log T (\log(TK/\delta))} \cdot T^{\frac{1+D}{2+D}} + KM_0 \max_i \|f_i\|_\infty.$$

4. Contextual bandits on manifold

Context 的表征维度可能比较高，但是可能其真实维度不高，比如能够投影到一个低维的 manifold 上，那么 k-NN 估计的准确度可以和 manifold 的维度 d 相关而不和表征维度 D 相关。

Theorem 4. (Manifold Rate for k-NN (Jiang 2017b)) Let $\delta > 0$. There exists N_0 and universal constant C such that if $n \geq N_0$ and $k = \lfloor n^{2/(2+d)} \rfloor$, then with probability at least $1 - \delta$,

$$\sup_{x \in \mathcal{X}} |f(x) - f_k(x)| \leq C \sqrt{\log n \log(1/\delta)} \cdot n^{-1/(2+d)}$$

相应地，有 manifold 版本的 UCB。

Theorem 6. (UCB Regret Analysis on Manifolds) Let $\delta > 0$. Suppose that $M_0 \geq N_0$ and $M_1 > C$ in Algorithm 2. Then we have that with probability at least $1 - \delta$,

$$R_T \leq M_1 2^{\frac{1+d}{2+d}} K \sqrt{\log T (\log(TK/\delta))} \cdot T^{\frac{1+d}{2+d}} + KM_0 \max_i \|f_i\|_\infty. \quad \text{知乎 @张楚珩}$$

5. Infinite-armed bandits

当遇到 bandits 数量无穷的时候，就不能像之前那样对每个 bandit 单独估计一个 k-NN estimator 了，因此需要估计一个类似 state-action value 的东西：

Definition 4. (Mean Reward function)

$$f : \mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R},$$

where $f(x, a)$ is the expected reward of action $a \in \mathcal{A}$ at context $x \in \mathcal{X}$. 知乎 @张楚珩

考虑如下 infinite-armed 的 UCB 算法（文章还给了解 top arm identification 的算法，也是 uniform sampling，这里略了）

Algorithm 5 Infinite-Armed Upper Confidence Bound (UCB)

- 1: Parameters: M, M_1, T
 - 2: Define $\sigma(n) = M_1 n^{-1/(2+D+D')}$.
 - 3: For $t = 1, \dots, M$:
 - 4: Sample a_t uniformly from \mathcal{A} .
 - 5: Observe context x_t and reward R_t .
 - 6: For $t = M + 1, \dots, T$:
 - 7: Choose $I_t := \operatorname{argmax}_{a \in \mathcal{A}} \hat{f}(x_t, a) + \sigma(t)$. 知乎 @张楚珩
-

其中， D' 为 action space 的维度。则有如下 regret bound。

Theorem 9. *There exists \tilde{C}_1 and \tilde{C}_2 such that the following holds. Let $\delta > 0$. Suppose that M and M_1 are chosen sufficiently large in Algorithm 5 depending on f and σ . Then we have that with probability at least $1 - \delta$,*

$$R_T \leq \tilde{C}_1 \sqrt{\log T (\log(T/\delta))} \cdot T^{\frac{1+D+D'}{2+D+D'}} + \tilde{C}_2$$

Remark 5. *This shows a sub-linear regret of $\tilde{O}(T^{\frac{1+D+D'}{2+D+D'}})$.* @张楚珩

实验

除了一些比较直观的实验外，还把 MNIST 的识别 formulate 成了一个 contextual bandit 问题，把不同的类别当做 arm，把图片输入当做 context。这个想法还挺有意思。

编辑于 2019-08-07

强化学习 (Reinforcement Learning)

▲ 赞同 7 ▼

💬 添加评论

🔗 分享

♥ 喜欢

★ 收藏

...

文章被以下专栏收录



强化学习前沿
读呀读paper

进入专栏