Multi-Task Learning for Contextual Bandits

Aniket Deshmukh[1], Urun Dogan [2], Clayton Scott [1]

1. EECS, University of Michigan Ann Arbor, 2. Microsoft Research, Cambridge, UK

Introduction

Contextual bandits are a form of multi-armed bandit in which the agent has access to predictive side information (known as the context) for each arm at each time step, and have been used to model personalized news recommendation, ad placement, and other applications. In this work, we propose an upper confidence bound-based multi-task learning algorithm for contextual bandits, establish a corresponding regret bound, and interpret this bound to quantify the advantages of learning in the presence of high task (arm) similarity.

Contextual Bandits Setup

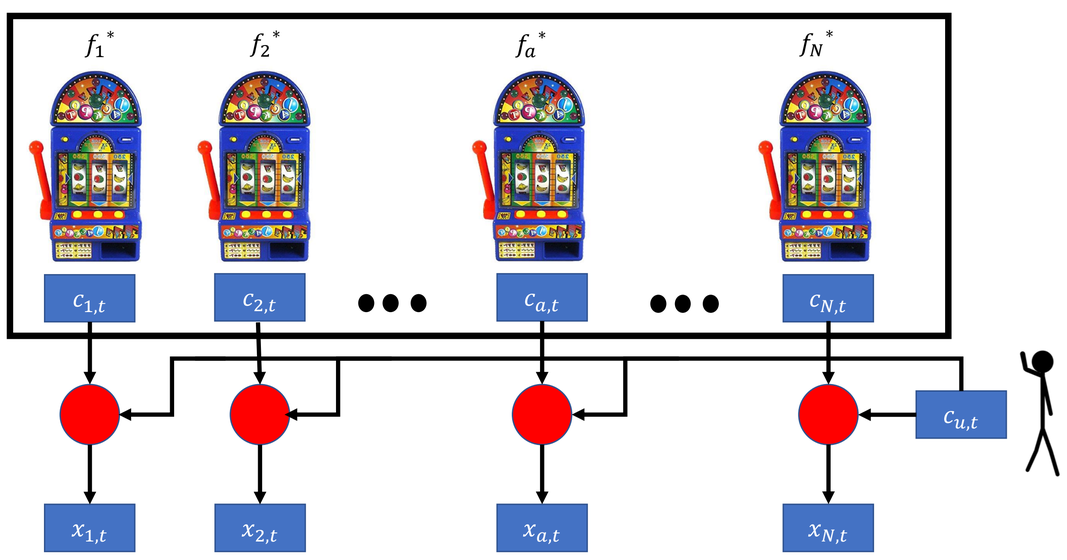

Figure 1: Contextual Bandits

1: **for** $t = 1, ..., T$ **do**
2: Observe context $x_{a,t} \in \mathbb{R}^d$ for all arms $a \in [N]$, where $[N] = \{1, ...N\}$
3: Choose an arm $a_t \in [N]$
4: Receive a reward $r_{a_t,t} \in \mathbb{R}$ s.t. $\mathbb{E}[r_{a_t,t}|x_{a,t}] = f_a^*(\tilde{x}_{a,t})$
5: Improve arm selection strategy based on new observation $(x_{a_t,t}, a_t, r_{a_t,t})$
6: **end for**

Goal: Minimize the T-trial regret, $R(T) = \Sigma_{t=1}^{T} r_{a_t^*,t} - \Sigma_{t=1}^{T} r_{a_t,t}$.

Proposed Multi-Task Learning Algorithm

- Popular algorithm like Lin-UCB [1] or Kernel-UCB [2] estimate f_a separately (independent) or learn a single estimator (pooled) for all $a \in A_t$.
- In the proposed KMTL-UCB approach, we seek to pool some tasks together, while learning others independently.
- We define the estimate on extended context space $\tilde{\mathcal{X}} = \mathcal{Z} \times \mathcal{X}$, $f : \tilde{\mathcal{X}} \to \mathcal{Y}$, where $\mathcal{Z}$ describes some notion of similarity.
- To estimate f, we minimize the following regularized empirical risk in RKHS:
  $\hat{f}_t = \arg\min_{f \in \mathcal{H}_{\tilde{k}}} \frac{1}{N} \Sigma_{a=1}^{N} \frac{1}{n_{a,t-1}} \Sigma_{\tau \in t_a} (f(\tilde{x}_{a,\tau}) - r_{a,\tau})^2 + \lambda \|f\|_{\mathcal{H}_{\tilde{k}}}^2$
- Let $\tilde{k}$ be a SPD kernel on $\tilde{X}$. In this work we focus on kernels of the form

$$\tilde{k}\Big((z,x),(z',x')\Big) = k_{\mathcal{Z}}(z,z')k_{\mathcal{X}}(x,x'), \qquad (1)$$

KMTL-UCB

Lemma 1 *Suppose the rewards* $[r_{a_\tau,\tau}]_{\tau=1}^{T}$ *are independent random variables with means* $\mathbb{E}[r_{a_\tau,\tau}|x_{a_\tau,\tau}] = f^*(\tilde{x}_{a_\tau,\tau})$, *where* $f^* \in \mathcal{H}_{\tilde{k}}$ *and* $\|f^*\|_{\mathcal{H}_{\tilde{k}}} \leq c$. *Let* $\alpha = \sqrt{\frac{\log(2TN/\delta)}{2}}$ *and* $\delta > 0$. *With probability at least* $1 - \frac{\delta}{T}$, *we have that* $\forall a \in [N]$

$$|\hat{f}_t(\tilde{x}_{a,t}) - f^*(\tilde{x}_{a,t})| \leq w_{a,t} := (\alpha + c\sqrt{\lambda})s_{a,t}$$

where $s_{a,t} = \lambda^{-1/2}\sqrt{\tilde{k}(\tilde{x}_{a,t},\tilde{x}_{a,t}) - \tilde{k}_{a,t}^T(\eta_{t-1}\tilde{K}_{t-1} + \lambda I)^{-1}\eta_{t-1}\tilde{k}_{a,t}}$.

UCB Algorithm: $a_t = \arg\max_{a \in A_t} \hat{f}_t(\tilde{x}_{a,t}) + \beta s_{a,t}$.

Regret Bound

Theorem 1 *Assume that* $r_{a,t} \in [0,1], \forall a \in [N]$, $T \geq 1$, *and the task similarity matrix* K_Z *is known. With probability at least* $1 - \delta$,

$$R(T) \leq \tilde{O}\left(\sqrt{T \log(g([T]))}\right)$$

where $g([T]) = \frac{\det(\tilde{K}_{T+1} + \lambda I)}{\lambda^{T+1}}$.

Theorem 2 *Let the rank of matrix* $K_{X_{T+1}}$ *be* r_x, *the rank of matrix* K_Z *be* r_z *and* $\tilde{k}(\tilde{x},\tilde{x}) \leq c_{\tilde{k}}, \forall \tilde{x} \in \tilde{X}$. *Then* $\log(g([T])) \leq r_z r_x \log\left(\frac{(T+1)c_{\tilde{k}} + \lambda}{\lambda}\right)$

Results

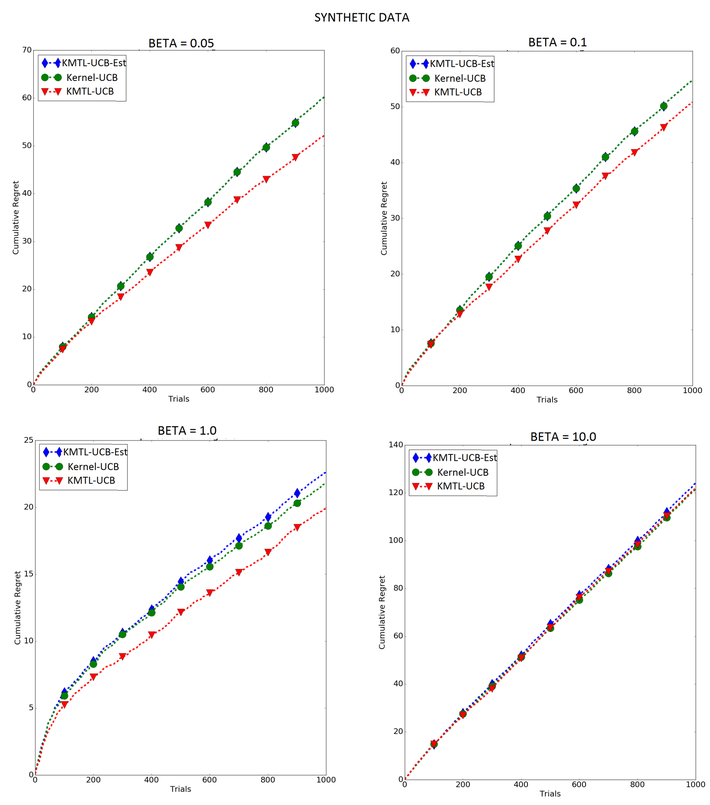

Figure 2: Cumulative Regret for Synthetic Data

References

[1] Chu, Wei, et al. "Contextual Bandits with Linear Payoff Functions." AISTATS. Vol. 15. 2011.

[2] Valko, Michal, et al. "Finite-Time Analysis of Kernelised Contextual Bandits." Uncertainty in Artificial Intelligence. 2013..

[3] Deshmukh, Aniket Anand, Urun Dogan, and Clayton Scott. "Multi-Task Learning for Contextual Bandits." arXiv preprint arXiv:1705.08618 (2017).