An Actor-Critic Contextual Bandit Algorithm for Personalized Interventions using Mobile Devices

Huitian Lei, Ambuj Tewari and Susan Murphy

STATISTICS

Introduction and motivation

- Mobile health (mHealth): the practice of medicine and public health supported by mobile devices (wikipedia).
- Just in time adaptive intervention (JITAI) are interventions that are delivered in real-time, and are adapted to address the immediate and changing needs of individuals as they go about their daily lives (Nahum-Shani et al 2014), by employing the real-time data collection and communication capabilities that modern mobile devices provide.
- JITAI utilizes a policy that takes patient information as input and outputs recommended intervention.
- We make a first attempt to bridge the methodological gap between the increasing popularity JITAIs receive from clinical and behavioral science community and the lack of methodological guidance in constructing data-based high quality JITAI.
- We provide an online actor-critic algorithm to learn the optimal JITAI.

Framework: contextual bandits and parametrized stochastic policy

Consider a series of pre-chosen decision points $\{1, 2, ..., t, ...\}$ where user-device interaction happens. Upper case letter to denote random variable, lower case letter to denote realization.

1. Finite action space $\mathcal{A}$: all suitable interventions at the time (different types, dose, modalities of intervention, etc). Denote A_t the action at time t.
2. Context space $\mathcal{S}$: vector space of intervention relevant summary information (GPS location, self-reported measures, etc). Denote S_t the context vector at time t.
3. Linear expected reward: Given context s and action a, reward is a linear function of a d dimensional reward feature vector $f(s, a)$ with an unknown coefficient vector μ^* plus a noise term ϵ with standard deviation σ.

$$R = f(s, a)^T \mu^* + \epsilon \tag{1}$$

The choice of a_t is based on the algorithm's analysis of context-action-reward tuples from the previous $t-1$ decision points. Upon acquiring the new information at decision point t, the algorithm updates its analysis in order to improve its choice of action at decision point $t+1$.

4. Our focus of the policy is a class of parameterized stochastic policies $\pi_\theta(s, a)$ where $\theta \in \mathbb{R}^p$ is a p dimensional vector.

$$\pi_\theta(s, a) = \frac{e^{g(s,a)^T\theta}}{1 + e^{g(s,a)^T\theta}} \tag{2}$$

1. Contribution of each element in context is reflected by θ.
2. Creating intervention variety (exploration): preventing intervention burden/habituation/boredom.
3. Most likely $p \ll d$.

Comparing to existing contextual bandit literature

1. We focus on low dimensional policy space. Scientists to choose a small set of contextual variables that potentially moderate the effect of interventions.
2. We encourage a minimal amount of exploration primarily to improve intervention adherence and enhance intervention effects. Algorithmic benefits are secondary.
3. We quantify data uncertainty, thus the uncertainty of the learned optimal policy, by by performing statistical inference on policy parameters, such as creating asymptotic confidence intervals.

For References, Questions and comments please contact ehlei@umich.edu

Regularized average reward

The average reward of a policy $\pi_\theta(s, a)$ is the expected reward $E(R|s, a)$ weighted by the policy-specified probability distribution over action space and the distribution over context space.

$$V(\theta) = \sum_{s\in\mathcal{S}} d(s) \sum_{a\in\mathcal{A}} E(R|s, a)\pi_\theta(s, a)$$

Naturally, we define the optimal policy $\pi_{\theta^{**}}$ as the policy that maximizes the average reward, $\theta^{**} = \text{argmax}_\theta V(\theta)$. We argue for the need to penalize

- When $E(R|S = s, A = a)$ is a (near) constant for all a, maximizing $V(\theta)$ is an ill-posed problem since solution is not unique.
- When $E(R|S = s, A = a)$ is not a constant, it is easy to provide example where the maximizer of $V(\theta)$ has one or more entries equal to ∞. One or more infinitely large θ_i's render other θ_i's unidentifiable and the problem ill-posed (Mahoney 2012)

Our solution to circumvent ill-posed problems is to regularize the average reward by subtracting a L_2 penalty term.

- The regularized average reward is

$$J(\theta) = \sum_{s\in\mathcal{S}} d(s) \sum_{a\in\mathcal{A}} E(R|s, a)\pi_\theta(s, a) - \lambda\|\theta\|_2^2 \tag{3}$$

where λ is a tuning parameter that controls the amount of penalization. The optimal policy is $\pi_{\theta^*}(s, a)$ with $\theta^* = \text{argmax}_\theta J(\theta)$.

- Maximizing $J(\theta)$ becomes a well-posed problem and convergence to a pure stochastic policy is guaranteed in the absence of intervention effects.
- Guarantees a minimum exploration probability.

An actor-critic algorithm with linear expected reward

Algorithm 1: Actor critic algorithm for linear expected reward

T_{max} is the total number of decision points.
Critic initialization: $B(0) = I_{d\times d}$, a $d \times d$ identity matrix. $A(0) = 0_d$, $\mu_0 = 0_d$, are $d \times 1$ column vectors.
Actor initialization: θ_0 to be the theory based policy parameter provided by behavioral scientists.
Start from $t = 0$.
while $t < T_{max}$ **do**
  At decision point t, observe context s_t ;
  Draw an action a_t according to probability distribution $\pi_{\theta_{t-1}}(s_t, a)$;
  Observe an immediate reward r_t ;
  Critic update:
  $B(t) = B(t-1) + f(s_t, a_t)f(s_t, a_t)^T$, $A(t) = A(t-1) + f(s_t, a_t)r_t$, $\mu_t = B(t)^{-1}A(t)$. ;
  Actor update:

$$\theta_t = \underset{\theta}{\text{argmax}} \frac{1}{t} \sum_{\tau=1}^{t} \sum_{a} f(s_\tau, a)^T \mu_t \pi_\theta(s_\tau, a) - \lambda\|\theta\|_2^2 \tag{4}$$

  Go to decision point $t+1$.
end

Theory: consistency, rate of convergence and asymptotic distributions

Theorem 1. *The critic's estimate μ_t converges to μ^* in probability. The convergence rate is $O(1/\sqrt{t})$, the optimal parametric convergence rate. Furthermore, $\sqrt{t}(\mu_t - \mu^*)$ converges in distribution to multivariate normal with mean 0_d and covariance matrix $[\mathbb{E}_{\theta^*}(f(s, a)f(s, a)^T)]^{-1}\sigma^2$, where $\mathbb{E}_\theta(f(s, a)f(s, a)^T = \sum_s d(s)\sum_a f(s, a)f(s, a)^T\pi_\theta(s, a)$. The plug-in estimator of the asymptotic covariance is consistent.*

Theorem 2. *The critic's estimate θ_t converges to θ^* in probability. The convergence rate is $O(1/\sqrt{t})$. Furthermore, $\sqrt{t}(\theta_t - \theta^*)$ converges in distribution to multivariate normal with mean 0_p and covariance matrix $(g^*_{\theta\theta})^{-1}V^*(g^*_{\theta\theta})^{-1}$, where $V^* = \sigma^2 g^*_{\theta\mu}\mathbb{E}_\theta(f(s, a)f(s, a)^T g^*_{\mu\theta} + \sum_s d(s)G_\theta(\mu^*, \theta^*, s)G_\theta(\mu^*, \theta^*, s)^T$. The plug-in estimator of the asymptotic covariance is consistent, where*

$$G(\mu, \theta, s) = \sum_a f(s, a)^T \mu\pi_\theta(s, a) - \lambda|\theta|_2^2$$

$$g(\mu, \theta) = \sum_s d(s) \sum_a f(s, a)^T \mu\pi_\theta(s, a) - \lambda|\theta|_2^2$$

Simulation: An example of reducing college students smoking

- Test the actor critic algorithm using an example in the context of reducing smoking among college students (Witkiewitz et al. 2014).
- Relevant contextual information, $s = [s_1, s_2, s_3, s_4]$. s_1 = smoking urge, s_2 = fixed mind, s_3 = indicator of low smoking mood(≤ 2), s_4 = indicator of user not busy.
- Three decision points per day. Action: $a = 0$ provide informational intervention, $a = 1$ provide behavioral intervention.
- Given context s and action a, the negative reward, or the cost is the smoking rate C. Smoking rate defined to be the average cigarettes smoked per hour between two decision points.

$$C = 6.54 + 1.16s_1 - 1.97a - 2.06s_2 - 1.47s_3 - 2.87s_4 + 1.4215s_2a + 1.5060s_3a + 1.25s_4a + \epsilon$$

where ϵ is a truncated normal random variable.

- Consider the class of policies: $\pi_\theta(s, a = 1) = \frac{e^{\theta_0+\theta_1 s_2+\theta_2 s_3+\theta_3 s_4}}{1+e^{\theta_0+\theta_1 s_2+\theta_2 s_3+\theta_3 s_4}}$

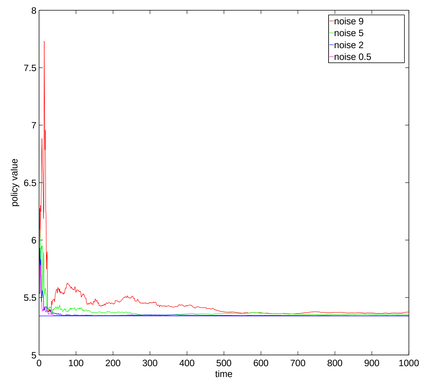

Figure 1: convergence of θ in terms of policy value.

Acknowledgement

We acknowledge the support of University of Michigan Mcubed grant for Methodology for Developing Real-Time Mobile Phone Adaptive Interventions, NIH grants P50DA010075, and Grant U54EB020404 (NIDCR, NIAMS, NHLBI, NIBIB, NIA).