Personalized Heartsteps V2, Pt.2

Guo Yuanxin

CUHK-Shenzhen

November 27, 2019

Challenges

1. Algorithm must adjust for longer term effects of current actions.
2. Algorithm should learn quickly & accommodate noisy data.
3. Algorithm should model mis-specification & non-stationarity.
4. Algorithm should select actions that allows secondary data analyses.

C1: Long-term Effect

- In mobile health, current action may have a large impact on future rewards.
- Intuitively, frequent activity suggestions increases immediate reward, but reduces future rewards due to user habituation or burden.
- Considering future rewards: Akin to using a large discount rate.
- We use a **dosage variable** to do this.

C1: Dosage Variable

- Goal: Model the frequency of suggestions sent at a given time.
- Take action when dosage is large $\Rightarrow$
  - Smaller immediate effect
  - Lower future results
- Initializing the dosage variable to be 0, the transition law is given by:

$$\tau(x'|x,a) = \begin{cases} \mathbf{1}_{\{x'=\lambda x+1\}}, & a=1 \\ p_{sed}\mathbf{1}_{\{x'=\lambda x+1\}} + (1-p_{sed})\mathbf{1}_{\{x'=\lambda x\}}, & a=0 \end{cases}$$

where $p_{sed} = 0.2$ is the probability of 1) no suggestion sent at previous decision time, 2) anti-sedentary suggestion sent after last decision time, and $\lambda = 0.95$.

C2: Role No.2 of Dosage Variable

- An obstacle to achieving high learning rate is high variance.
- To control variance, we use a low-dimensional MDP model to form a **proxy** (value) of the future rewards. This MDP again involves the dosage variable.
- The MDP is reconstructed every night to obtain a new proxy value, which affects the decision in the following day by:

$$\Pr\{f(s)^\top \beta > \eta_d(x);\ \beta \sim \mathcal{N}(\mu_d, \Sigma_d)\} \tag{1}$$

where ν_d is the proxy value for day d.

C2: Control Variance

- Another approach to control variance is using a low-dimensional linear model to model the difference in the reward function under alternate actions (treatment effect).

$$r_t(s,1) - r_t(s,0) = f(s)^\top \beta \tag{2}$$

- This allows us to trade off the bias and the variance to accelerate learning.

C2: Thompson Sampling

- Our application is a descendant from HeartSteps V1, so we actually have some informative prior knowledge.
- This leads us to think in a Bayesian way, in particular, **Thompson sampling**.
- The advantages of using a informative prior distribution (of parameters) are:
  - Speed up learning in the early phase;
  - Reduce variance;
  - Diminish the impact of noisy observation

C3: Baseline Function & Action-centering

- We use the following Bayesian regression model to model the reward and thus estimate β:

$$R_t = g(S_t)^\top \alpha_0 + \pi_t f(S_t)^\top \alpha_1 + (A_t - \pi_t) f(S_t)^\top \beta + \mathcal{N}(0, \sigma^2)$$

- The red term is a low-dimensional model of the "baseline" reward function: $r_t(s, 0)$.
- The blue term is the action-centering term.
- As long as the treatment-effect model (2) is correct, the estimator of β is unaffected by g, even if g is non-stationary (Robustness against mis-specification & non-stationarity)
- In addition, the action-centered term $(A_t - \pi_t)$ enables the Bayesian regression to give a linear approximation to the treatment effect, while without centering, the estimates may not converge to any useful approximation.

C4: Stochastic Policy & Clipping

- We use a stochastic policy (via TS) to ensure the algorithm will continue to explore (as opposed to exploit).
- Clipping restricts the probability of each action in an interval, which enables off-policy data analyses.