Learning Treatment Policies in Mobile Health

S.A. Murphy

AAAI2016

MD2K
Center of Excellence for
Mobile Sensor
Data-to-Knowledge

The Methodology Center
advancing methods, improving health

MICHIGAN
STATISTICS

INSTITUTE FOR SOCIAL RESEARCH
UNIVERSITY OF MICHIGAN

The Dream!

"Continually Learning Mobile Health Intervention"

- Help you achieve and maintain your desired long term healthy behaviors
  - Provide sufficient short term reinforcement to enhance your ability to achieve long term benefit
- The ideal mHealth intervention
  - will engage you when you need it and will not intrude when you don't need it.
  - will adjust to unanticipated life challenges

mHealth

HeartSteps Activity Coach

- Wearable band measures activity, phone sensors measure busyness of calendar, location, weather, ...

- In which contexts should smartphone ping and deliver activity recommendations?

mHealth

MD2K Smoking Cessation Coach

- Wearable bands measure activity, stress, cigarette smoking; smartphone sensors provide location,…….
- In which contexts should the wrist band provide supportive stress-reduction "cue" and smartphone activate to highlight associated stress reduction support?

Data from wearable devices that sense and provide treatments

On each individual:

$$O_1, A_1, Y_2, \ldots, O_t, A_t, Y_{t+1}, \ldots$$

O_t: Observations at t[th] decision time (high dimensional)

A_t : Action at t[th] decision time (treatment)

Y_{t+1} : Proximal Response (aka: Reward, Utility, Cost)

Examples

1) Decision Times (Times at which a treatment can be provided.)
   1) Regular intervals in time (e.g. every 10 minutes)
   2) At user demand

<u>HeartSteps</u>: Approximately every 2-2.5 hours

<u>Smoking Cessation</u>: Every 1 minute during 10 hour day.

Examples

2) Observations O_t
    1) Passively collected (via sensors)
    2) Actively collected (via self-report)

HeartSteps: activity recognition, location, step count, busyness of calendar, usefulness ratings, adherence…….

Smoking Cessation: stress, smoking detection, mood, driving,….

Examples

3) Actions, A_t
   1) Treatments that can be provided at decision time
   2) Whether to provide a treatment

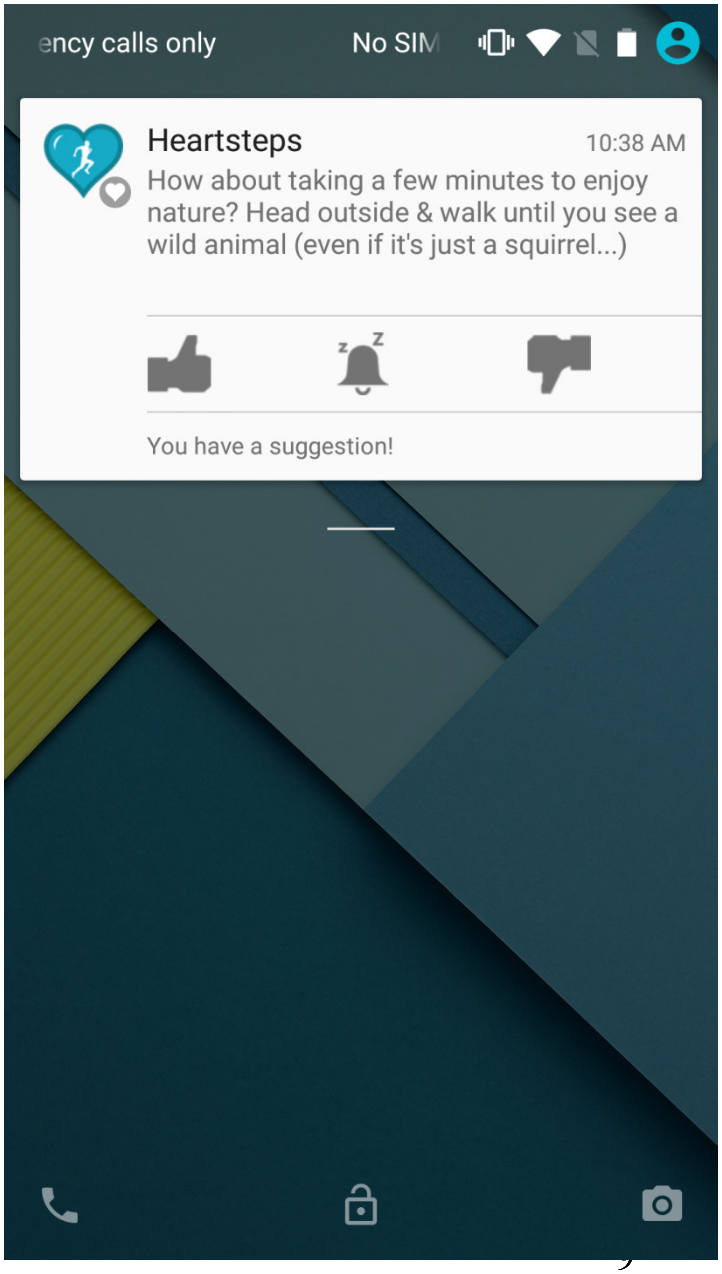
<u>HeartSteps:</u> Activity Recommendation on phone

<u>Smoking Cessation:</u> Cue on wrist band

Tailored Activity Recommendation

No Message

or



Examples

4) Proximal Response (reward) Y_{t+1}

HeartSteps: Activity (step count) over next 30 minutes.

Smoking Cessation: Stress over next x minutes

Continually Learning Mobile Health Intervention

1) Trial Designs: Do the actions affect the proximal response? *experimental design & causal inference*

2) Data Analysis Methods for use with trial data: Are there delayed effects of the actions? Do effects vary by context? *causal inference*

3) Learning algorithms for use with trial data: Construct a "warm-start" treatment policy. *batch RL*

4) Online training algorithms that will result in a Continually Learning, Personalized mHealth Intervention. *online RL*

Micro-Randomized Trial

Randomize between actions at decision times → Multiple individuals, each randomized 100's or 1000's of times.

- These are sequential, "full factorial," designs.
- Design trial to detect main effects.

Extension of A/B testing & Single Case Designs

Micro-Randomized Trial Elements

1. Record outcomes
   - Distal (scientific/clinical goal) & Proximal Response
2. Record context (sensor & self-report data)
3. Randomize among treatment actions at decision points
4. Use data after study ends to assess treatment effects, learn warm-start treatment policy

Micro-Randomized Trial

How to justify the trial costs?

- Address a question that can be stated clearly across disciplinary boundaries and be able to provide guarantees.
- Design trial so that a variety of further interesting questions can be addressed.

<u>First Question to Address</u>: Do the treatment actions impact the proximal response? (aka, is there a signal?)

Micro-Randomized Trial for HeartSteps

- 42 day trial

- Whether to provide an Activity Recommendation? $A_t \in \{0, 1\}$

- Randomization in HeartSteps
  $$P[A_t = 1] = .4 \quad t = 1, \ldots, T$$

Micro-Randomized Trial

Time varying potentially intensive/intrusive treatment actions → potential for accumulating habituation and burden

→

Allow main effect of the treatment actions on proximal response to vary with time

Availability & the Treatment Effect

- Treatment actions can not be delivered at a decision time if an individual is *unavailable*.

- The effect of treatment at a decision time is the difference in proximal response between *available* individuals assigned an activity recommendation and *available* individuals who are not assigned an activity recommendation.

Availability

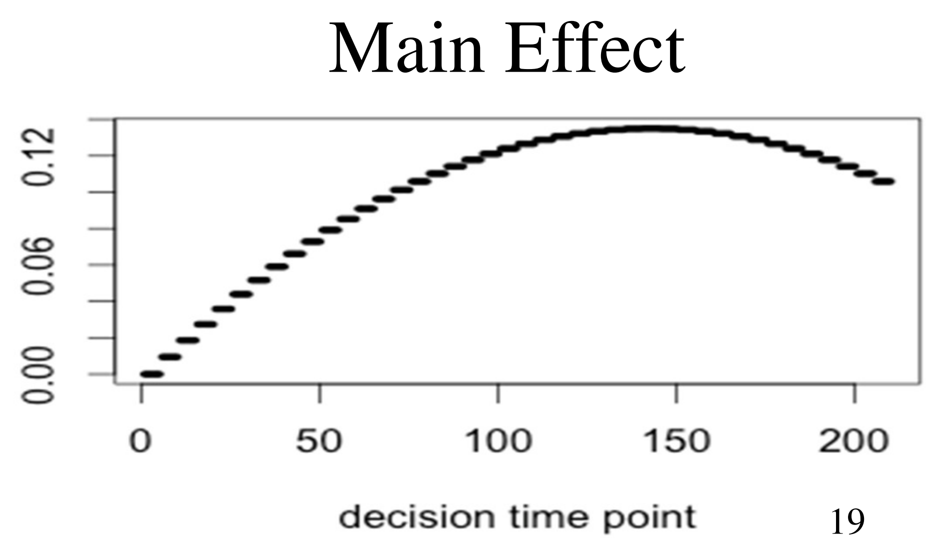
- Treatment actions can only be delivered at a decision time if an individual is *available*
- Set $I_t=1$ if the individual is available at decision time t, otherwise, $I_t=0$
- Availability is not the same as adherence.

Treatment Effect

- The Main Effect at time j is

$$\beta(t) = E\left[Y_{t+1} | I_t = 1, A_t = 1\right] - E\left[Y_{t+1} | I_t = 1, A_t = 0\right]$$

- What does this main effect $\beta(t)$ mean?



Sample Size for Trial

- We calculate the number of subjects to test H_0: no effect of the action, i.e.,

$$H_0 : \beta(t) = 0, t = 1, 2, ...T$$

- Size to detect a low dimensional, smooth alternate H_1.
  - Example: H_1: $\beta(t)$ quadratic with intercept, β_0, linear term, β_1, and quadratic term β_2 and test

$$\beta_0 = \beta_1 = \beta_2 = 0$$

Sample Size Calculation

- Our test statistic uses estimators from a “generalization” of linear regression.
- The test statistic is quadratic in the estimators of the β terms.
- Given a specified power to detect the smooth alternative, H_1, a false-positive error prob., and the desired detectable signal to noise ratio, we use standard statistics to derive the sample size.

Sample Size Calculation

Alternative hypothesis is low dimensional $\longrightarrow$ assessment of the effect of the activity recommendation uses contrasts of *between subject responses* **+** contrasts of *within subject responses*.

--The required number of subjects will be small.

HeartSteps Sample Sizes
Power=.80, False-positive error=.05

Standardized Average Main Effect over 42 Days	#Subjects for 70% availability or 50% availability
0.06 standard deviations	81 or 112
0.08 standard deviations	48 or 65
0.10 standard deviations	33 or 43

A Micro-Randomized Trial

The micro-randomized trial is a sequential factorial trial with multiple factors, e.g.

Factor 1: Activity recommendation is randomized 5 times per day

Factor 2: Daily activity planning is randomized each evening

42 day study

Experimental Design Challenges

Micro-randomized trials are a new type of factorial design

i. Time varying factors→ time varying main effects, time-varying two-way interactions, different delayed effects
ii. Randomization that depends on an outcome of past actions
iii. Design studies specifically to detect interactions between factors.

Continually Learning Mobile Health Intervention

1) Trial Designs: Do the actions affect the proximal response? *experimental design & causal inference*

2) Data Analysis Methods for use with trial data: Are there delayed effects of the actions? Do effects vary by context? *causal inference*

3) Learning algorithms for use with trial data: Construct a "warm-start" treatment policy. *batch RL*

4) Online training algorithms that will result in a Continually Learning & Personalized mHealth Intervention. *online RL*

Treatment policies

- Most current treatment policies are constructed using behavioral theory, clinical experience, observational data analyses and expert opinion.
- We aim to develop algorithms that use trial data in constructing treatment policies.

  -- treatment policy should be interpretable.

  -- treatment policy can act as a "warm-start" in future implementation of an online algorithm.

Stochastic Treatment Policy

Construct a parameterized policy, $\pi_\theta(a|s)$

- Ensure $\pi_\theta(a|s)$ probabilities bounded away from 0 and 1: variation in actions can help retard habituation and maintain engagement.
- Parameterized $\pi_\theta(a|s)$ can be interpreted/vetted by domain experts

Setup

1) On each of *n* individuals, data set contains:

$$S_1, A_1, Y_2, \ldots, S_T, A_T, Y_{T+1}$$

-- S_t is a summary of $O_1, A_1, Y_2, \ldots, Y_t, O_t$ that permits the Markovian property; this is a modeling assumption.

-- known randomization

$$P[A_t = a | S_t = s] = \mu(a|s)$$

2) Optimality criterion to maximize: Average Reward resulting from use of policy π_θ

Markov Decision Process

Markovian Assumptions

$$P[S_{j+1} = s' | S_1, A_1, \ldots, S_j, A_j] = P[S_{j+1} = s' | S_j, A_j]$$

and

$$P[Y_{j+1} = r | S_1, A_1, \ldots, S_j, A_j] = P[Y_{j+1} = r | S_j, A_j]$$

Stationarity Assumptions

$$P[S_{j+1} = s' | S_j = s, A_j = a] = p(s' | s, a)$$

and

$$E[Y_{j+1} | S_j = s, A_j = a] = r(s, a)$$

Optimality Criterion (to maximize)

Average Reward, η_θ, for policy π_θ:

$$\begin{aligned}\eta_\theta &= \lim_{T\to\infty}\frac{1}{T}E_\theta\left[\sum_{t=0}^{T-1}Y_{t+1}\middle|S_0=s_0\right]\\ &= \sum_s d_\theta(s)\sum_a \pi_\theta(a|s)r(s,a)\end{aligned}$$

E_θ denotes expectation under the stationary distribution, d_θ, associated with π_θ.

Background: Differential Value

V_θ is the Differential Value

$$V_\theta(s) = \lim_{T\to\infty} E_\theta\left[\sum_{t=0}^{T}\Big(Y_{t+1} - \eta_\theta\Big)\Big| S_0 = s\right].$$

$V_\theta(s)$ - $V_\theta(s')$ reflects the difference in sum of centered responses accrued when starting in state *s* as opposed to state *s*′.

(η_θ is the average reward)

Background: Bellman Equation

Oracle Temporal Difference:

$$\delta_t = Y_{t+1} - \eta_\theta + V_\theta(S_{t+1}) - V_\theta(S_t)$$

Bellman Equation:

$$E_\theta \left[\delta_t \middle| S_t \right] = 0$$

$S_t, A_t, Y_{t+1}, S_{t+1}$

Background: Bellman Equation

Bellman's equation implies that

$$E\left[\frac{\pi_\theta(A_t|S_t)}{\mu(A_t|S_t)}\Big(Y_{t+1}-\eta+V(S_{t+1})-V(S_t)\Big)\begin{pmatrix}1\\ f(S_t)\end{pmatrix}\right]$$

will be, for all *t*, for any vector, *f(.)*, of appropriately integrable functions, and expectation over data generating distribution, *E*, equal to 0 if $\eta=\eta_\theta,\ V=V_\theta$

Estimating Function

- Construct a flexible model for, $V_\theta(s)$, say $f(s)^T v_\theta$ for *f(s)* a *p* by *1* vector of basis functions evaluated at *s* (*p* is large)

- Solve

$$\mathbb{P}_n\left[\sum_{t=1}^{T}\frac{\pi_\theta(A_t|S_t)}{\mu(A_t|S_t)}\Big(Y_{t+1}-\eta+f(S_{t+1})^T v-f(S_t)^T v\Big)\begin{pmatrix}1\\ f(S_t)\end{pmatrix}\right]$$

=0 for $\hat{\eta}_\theta$, $\hat{v}_\theta$

Overview of Algorithm

- The resulting η and v are functions of θ, denote by $\hat{\eta}_\theta, \ \hat{v}_\theta$
    - $\hat{\eta}_\theta, \ \hat{v}_\theta$ are the output of the Critic
- The Actor maximizes $\hat{\eta}_\theta$ over θ to obtain $\hat{\theta}$.
    - this will require repeated calls to the Critic
    - $\hat{\theta}$ is the output of the Actor

Actor

- The objective function for the actor is given by

$$\hat{\eta}_\theta = \mathbb{P}_n \left[\sum_{t=1}^{T} \frac{\pi_\theta(A_t|S_t)}{\mu(A_t|S_t)} \Big(Y_{t+1} + f(S_{t+1})^T \hat{v}_\theta - f(S_t)^T \hat{v}_\theta \Big) \right]$$

- We want to construct a policy, π_θ that is bounded away from 0, 1.

Binary action: $$\pi_\theta(a|s) = \frac{e^{\theta^T g(s) a}}{1 + e^{\theta^T g(s)}}$$

Actor

Chance constraint on θ:

$$\min_a P^* \left[p_0 \leq \pi_\theta(a|S) \leq 1 - p_0\right] \geq 1 - \alpha$$

given α, p_0 and P^*, a reference distribution over states, S.

This constraint is nonconvex; we relax via Markov inequality.

CRITIC

Write

$$\mathbb{P}_n\left[\sum_{t=1}^{T}\frac{\pi_\theta(A_t|S_t)}{\mu(A_t|S_t)}\Big(Y_{t+1}-\eta+f(S_{t+1})^Tv-f(S_t)^Tv\Big)\begin{pmatrix}1\\ f(S_t)\end{pmatrix}\right]$$

$$=\hat{A}_\theta\begin{pmatrix}\eta\\ v\end{pmatrix}-\hat{b}_\theta$$

The critic minimizes

$$||\hat{A}_\theta\begin{pmatrix}\eta\\ v\end{pmatrix}-\hat{b}_\theta||^2+\lambda_c||v||^2$$

to obtain

$$\hat{\eta}_\theta,\ \ \hat{v}_\theta$$

ACTOR

- The actor obtains $\hat{\theta}$ by maximizing

$$\hat{\eta}_\theta = \mathbb{P}_n \left[\sum_{t=1}^{T} \frac{\pi_\theta(A_t|S_t)}{\mu(A_t|S_t)} \Big(Y_{t+1} + f(S_{t+1})^T \hat{v}_\theta - f(S_t)^T \hat{v}_\theta \Big) \right]$$

subject to the constraint, $\theta^T \Sigma_g \theta \leq k_{max}$

$$\Sigma_g = T^{-1} \sum_{t=1}^{T} E^* \left[g(S_t) g(S_t)^T \right]$$

BASICS Mobile

- Smartphone-based intervention to reduce heavy drinking and smoking in college students
  - 14 day study
  - Self-report 3x/day (morning, afternoon, evening)
  - Intervention 2x/day (afternoon, evening)
    - Mindfulness-based intervention ($A_t=1$) vs general health information ($A_t=0$)
- <u>Question</u>: Should a mindfulness-based intervention (vs general health info) be provided when there is an increase in need to self-regulate?

BASICS Mobile

- n subjects = 27, T decision points = 28
- Availability: To be available to receive a treatment, the student must complete self-report questions ($I_t = 1$). If the student is available then the student is provided a treatment with probability 2/3.
- Reward is (-)smoking rate

BASICS Mobile

- S_t is 8 dimensional composed of 5 discrete and 3 continuous valued features.
- Differential value approximated by B-splines and two way products of B-splines constructed from entries in S_t.
- Parameterized policy:

$$\pi_\theta(1|s) \quad = \quad I_t \frac{e^{\theta_0+\theta_1 g_1+\theta_2 g_2}}{1+e^{\theta_0+\theta_1 g_1+\theta_2 g_2}}$$

BASICS Mobile

- g_1 is indicator for an increase in self-regulation demands (1 if yes, 0 if no)
- g_2 is indicator for no burden (1 if yes, 0 if no)
- $\hat{\theta}_0 = .74,\ \hat{\theta}_1 = -.95,\ \hat{\theta}_2 = 2.26 \rightarrow$ An available student with no increase in self-regulation demands and who is not indicating burden is recommended treatment with probability 0.85

$$\pi_\theta(1|s) \quad = \quad I_t \frac{e^{\theta_0+\theta_1 g_1+\theta_2 g_2}}{1+e^{\theta_0+\theta_1 g_1+\theta_2 g_2}}$$

Challenges

- Bandit vs Average Reward vs Discounted Reward?
  - Burden → disengagement raises the need to pay attention to future.
  - In batch setting and/or online setting?
- Disengagement is a terminal event: Safe exploration?
- Method should provide confidence intervals/permit scientists to test hypotheses.

General Challenges

- How to reduce the amount of self-report data (How might you do this?)
- Non-stationarity: Transfer learning within a user?
- Measuring burden without causing burden.
- How to accommodate/utilize the vast amount of missing data, some of which will be informative……

Collaborators