Optimal Dynamic Treatment Regime by Reinforcement Learning in Clinical Medicine

Mina Song & David Han, Ph.D. (david.han@utsa.edu)
The University of Texas at San Antonio, TX 78249

Reinforcement ML

- a machine learning (ML) method that takes action in response to the changing environment over time for maximizing rewards, *R*
- application domains of RL
  - **dynamic treatment regime** (**DTR**)
    * chronic diseases: cancer, diabetes, anemia, HIV, mental illness such as epilepsy, depression, Schizophrenia, opioid addiction
    * critical care: sepsis, anesthesia, ventilation, heparin dosing, and so on
  - automated medical diagnosis w/ structured data (medical imaging) and unstructured data (free text)
  - resource scheduling and task allocation, optimal process control, drug discovery (*de novo* design), healthcare management, etc.

Dynamic Tx Regime

- RL approach in *precision medicine* to enable the optimal personalized treatment regime for patients w/ distinct genetic, demographic, clinical characteristics

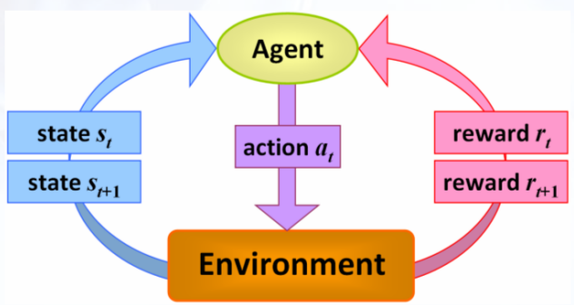

Figure 1. Schematic illustration of RL;
its formulation requires
- policy: map from state to action
- value function: total expected reward over time

Why RL-based DTR?

- incomplete knowledge of environment, usually estimated
  - dynamic programming is often inappropriate.
- limited sample size and costly data collection
- causal association of historical conditions w/ final outcome (*viz.*, no Markov property)
  - exponentially growing state and action space compared to sample size

Q-Learning

- a temporal difference control algorithm to search the optimal DTR from longitudinal data
- backward recursive fitting of linear models based on a dynamic programming algorithm

$$Q_t(h_t, a_t) = E\left[\max_{a_{t+1}} Q_{t+1}(h_{t+1}, a_{t+1}) \middle| h_t, a_t\right]$$

$$\text{with } Q_T(h_T, a_T) = E\left[R | h_T, a_T\right]$$

- easy implementation and interpretation for domain experts
- risk of model misspecification
  - inverse probability weighted estimator (IPWE)
    * non-parametrically estimate mean outcome w/ different weights to the observed outcomes
    * robust but noisy contrast for classification
  - augmented inverse probability weighted estimator (AIPWE)
    * combine info from both propensity score and mean outcome models for smoothing

➔ action space is *binary*; need to implement multi-dimensional action space for combinations of treatment regime.

Illustrative Example

- two-stage treatment w/ multiple (3) treatments in each stage (*n* = 500)

stage 1: 3 covariates (x11, x12, x13)
3 treatments/actions (a11, a12, a13)
stage 2: 3 covariates (x21, x22, x23)
3 treatments/actions (a21, a22, a23)

R = final outcome (reward)
- continuous variable
- higher the value is, better the outcome is.

A = *empirical* treatment decision (action) based on multinomial distribution w/ probability vector by

$$e^{X\beta_i} / \sum_j e^{X\beta_j}$$

optA = *optimal* treatment decision rule
stage 1: treatment 1 if (x11>-0.54) and (x12<0.54)
else treatment 2 if (x11>-0.54) and (x13<0.54)
else treatment 3
stage 2: treatment 1 if (x21>0.3) and (x23<0.46)
else treatment 2 if (x22>0.3) and (x23<0.46)
else treatment 3

Confusion matrix @ stage 2

		optA 1	2	3	
A	1	69	24	84	
	2	31	13	106	(accuracy = 0.388)
	3	23	38	112	

		optA 1	2	3	
DTR	1	108	2	28	
	2	3	46	18	(accuracy = 0.820)
	3	12	27	256	

➔ The optimal DTR at stage 2 has significantly better accuracy than the empirical treatment decision.

Table 1. Snapshot of the dataset

x11	x12	x13	A1	x21	x22	x23	A2	R
-0.02	0.36	0.36	3	0.66	0.39	-0.38	2	1.61
0.49	-0.22	-0.28	2	-1.47	0.66	0.79	3	-0.41
0.75	0.55	-0.29	3	0.22	1.66	0.14	2	1.14
0.28	1.83	0.47	3	0.81	1.78	0.95	2	1.42
0.00	0.30	-0.09	2	0.30	0.33	1.00	1	0.31
				.	.			
				.	.			
				.	.			

Confusion matrix @ stage 1

		optA 1	2	3	
A	1	66	39	44	
	2	120	19	40	(accuracy = 0.336)
	3	63	26	83	

		optA 1	2	3	
DTR	1	241	67	64	
	2	8	0	9	(accuracy = 0.670)
	3	0	17	94	

➔ The optimal DTR at stage 1 also has significantly better accuracy than the empirical treatment decision.

References

- Fernandez, K.C., Fisher, A.J., and Chi, C. (2017). Development and initial implementation of the dynamic assessment treatment algorithm. *PLoS One*, **12:** e0178806.
- Laber, E.B. and Davidian, M. (2017). Dynamic treatment regimes, past, present, and future. *Statistical Methods in Medical Research*, **26:** 1605–1610.
- Laber, E.B., Lizotte, D.J., Qian, M., Pelham, W.E., and Murphy, S. (2014). Dynamic treatment regimes: technical challenges and applications. *Electronic Journal of Statistics*, **8:** 1225–1272.
- Murphy, S. (2003). Optimal dynamic treatment regimes. *Journal of Royal Statistical Society B*, **65:** 331–366.
- Wallace, M.P. and Moodie, E.E. (2015). Doubly-robust dynamic treatment regimen estimation via weighted least squares. *Biometrics*, **71:** 636–644.
- Zhang, Y., Laber, E.B., Davidian, M., and Tsiatis, A.A. (2018). Interpretable dynamic treatment regimes. *Journal of the American Statistical Association*, **113:** 1541–1549.
- Zhang, Z. (2019). Reinforcement learning in clinical medicine: a method to optimize dynamic treatment regime over time. *Annals of Translational Medicine*, **7:** e345.