6.S897/HST.956 Machine Learning for Healthcare

Lecture 17: Reinforcement Learning (II)

Instructors: David Sontag, Peter Szolovits

1 Lecture overview

First half of the lecture was taught by Prof. David Sontag, followed by a guest lecture by Dr. Barbra Dickerman.

1. Evaluation of policy - causal inference versus reinforcement learning (David Sontag)
2. Evaluating dynamic treatment strategies (Barbra Dickerman)
3. Discussion of the paper "The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care" (faciliated by Barbra Dickerman) [KCB+18]

2 Causal inference versus reinforcement learning

How can we evaluate different policies in causal inference? How is it related to reinforcement learning?

2.1 Reinforcement learning review

In lecture 16, we learn a policy from value-maximization approach. Define the best policy π^* as the policy with the maximum value:

$$\pi^* \leftarrow argmax_\pi V_\pi \quad \text{where } V_\pi = \mathbb{E}_\pi[\sum_t R_t]$$

Implication: We want to find a policy with high expected value of reward average over all patients.

Caveats:

- Under mission critical reward space (patient might die), we might want to capture the worst case reward rather than the average reward.
- Infinite horizon (negative infinity for patients dying), which leads to infeasible optimization problem.
- High variance in reward function. The average reward might be the same in this case compared to the uniform reward scenario. However, it is important that we capture the worst case rewards in different quantiles rather than averaging over the entire population.

2.2 Covariate adjustment: expected reward of a policy

In the previous lecture, we learn that, from the Q-learning algorithm, the expected reward of a policy can be given as:

$$\hat{V}_\pi = max_a Q(s_0, a) \tag{1}$$

We will show that we can come up with a similar formula for the expected value of a policy under causal inference.

Recall that for covariate adjustment, we learn a function $f(X,T) \sim \mathbb{E}[Y_t|X]$ and use it to define:

$$CATE(X) = f(X,1) - f(X,0)$$

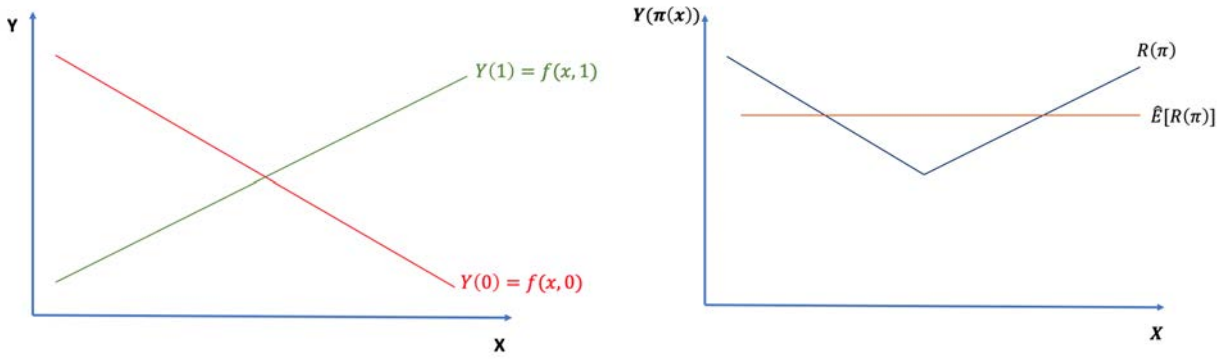

Figure 1:

For the rest of the lecture, consider a policy π such that:

$$\pi(X) = \begin{cases} 1 & \text{if } CATE(X) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

An analogous formula for expected value of a policy under causal inference is as following:

$$\hat{\mathbb{E}}[R(\pi)] = \frac{1}{n} \sum_i \big[\pi(x_i) f(x_i, 1) + (1 - \pi(x_i)) f(x_i, 0)\big] \tag{3}$$

$$= \frac{1}{n} \sum_i max\big(f(x_i, 1), f(x_i, 0)\big) \tag{4}$$

where (3)means that the expected value of reward is the average reward for each individual (patient) given the outcome of the policy and the treatment for that individual, (4) is the derived formula given the policy outlined in (2). Here, we can see the analogy between (2) and the expected reward of a policy under reinforcement learning, (1).

For example: Consider the following case when covariate X is "age" (single covariate) in Figure 1. The figure on the left shows the outcome given treatment/no treatment as a function of X. The figure on the right, on the other hand, shows the value of $\pi(X)f(X, 1) + (1 - \pi(X))f(X, 0)$. The expected reward value, is averaged over all values of X's in the population.

2.3 Propensity score: expected reward of a policy

We just review how we can estimate the expected reward of a policy given the counterfactual outcomes in covariate adjustment. However, we might want to derive a formula which doesn't require having to estimate potential outcomes.
Recall that to estimate ATE, we use the propensity score $e_i = P(T = 1|x_i)$ in the following formula:

$$\hat{ATE} = \frac{1}{n}\Big(\sum_{i:t_i=1} \frac{y_i}{e_i} + \sum_{j:t_j=0} \frac{y_j}{1 - e_j} \Big) \tag{5}$$

The inverse propensity score weighted (IPW) is a way to turn observational study into a pseudo-randomized trial by re-weighting samples. Therefore, we can estimate the expected reward of a policy, $\hat{\mathbb{E}}[R(\pi)]$, by:

$$\hat{R}^{IPW}(\pi) = \frac{1}{n} \sum_i \frac{\mathbb{1}[t_i = \pi(x_i)] y_i}{P(T = t_i|x_i)} \tag{6}$$

where $\mathbb{1}[.]$ is the identity function.

Pros:

- Don't have to impute counterfactual outcome
- Can derive the best policy given a considerably large observational dataset
- In the randomized controlled trial setting, the $P(T = t_i|x_i) = 0.5$. Therefore, (6) gives an unbiased estimator of the policy's reward for this randomized controlled trial.

Cons:

- Need to know the propensity score (for example: need to know the process of the randomized controlled trial or the underlying data generating distribution)
- Large enough observational dataset to prevent overfitting and enable us to estimate the propensity score directly
- If the dataset is small or if it has limited overlap, the estimation will have very high variance.

If the randomized controlled size is large enough, we can derive the best policy without estimating ATE:

$$\pi^* \leftarrow argmax_\pi \hat{R}^{IPW}(\pi)$$

which is a weighted classification problem.
To learn more about this approach, please consult the papers by [SJ15] and [KZ18]. In the first paper, they tackle the above problem by realizing that they might need biased estimator (for example, the propensity score can be very small). And then, they use generalization results from the theory of machine learning to bound the variance of the estimator as a function of propensity score. The second paper deals with the problem of unobserved confounders. As the result, we might not know the propensity score. Therefore, they try to bound how wrong the estimator can be given that how much they don't know about these confounding factors.

2.4 Return to reinforcement learning

Given the above estimator in 1-step causal inference, one can generalize to a t-step reinforcement learning value estimator as following:

$$\hat{V}_\pi = \frac{1}{n} \sum_i \sum_t R_{i,t} \prod_{t'=0}^{t} \frac{\pi(a_{t'}|s_{t'})}{P(a_{t'}|s_{t'})} \tag{7}$$

where $\pi(a_{t'}|s_{t'})$ denotes whether action $a_{t'}$ is taken at time $s_{t'}$ in policy π.

There are many other methods, such as W-robust estimator. More on [TB16].

3 Evaluating dynamic treatment strategies

3.1 What is dynamic treatment strategies?

1. Initiate treatment at baseline and continue over follow-up, unless a contraindication occurs
2. Do not initiate treatement over follow-up, unless an indication occurs

Clinicians encounter these problems in everyday practice. They need to take into consideration a patient's evolving characteristics before making a decision. For example, decisions about prevention, screening, or treatment interventions over time may depend on evolving comorbidities, screening results, or treatment toxicity. Moreover, strategies in clinical guidelines and practice are often dynamic since they often take into account patient's evolving characteristics over time. Likewise, the optimal strategies will be dynamic.

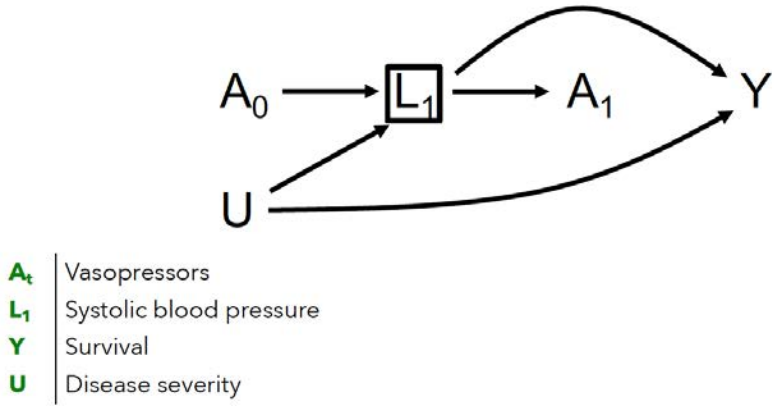

Figure 2: From this causal graph, we are interested in the effects of intervention A (vasopressors) on some outcome Y (survival). We know that the vasopressor will affect blood pressure, which then affect subsequent decisions about using vasopressor. We also know that blood pressure affects survival. In this graph, there is also an unobserved confounder U (disease severity) which affects both blood pressure and survival. If we want to measure the effect of a sustained treatment, we need to measure its effects at every single time point. L1 is a confounder for A1 and Y, and thus, a conventional statistical approach will lead to conditioning on a collider and inducing selection bias. As a result, there might be an association between A and L (maybe through the path A0 to L1 to Y). On the other hand, this association may not be a causal effect because it might be due to the selection bias.

3.2 What is G-methods? When is it needed?

Conventional statistical methods cannot appropriately compare dynamic strategies in the presence of treatment-confounder feedback. In other words, this is when the time-varying confounders are affected by previous treatment effect (Figure 2). This problem can be solved with G-methods, which provide an estimation of structural nested models and inverse probability weighting of marginal structural models.

Specifically, parametric g-formula is:

- Generalization of standardization to time-varying exposures and confounders
- Conceptually, the g-formula risk is a weighted average of risks conditional on a specified intervention history and observed confounder history
  - The weights are the probability density functions of the time-varying confounders, estimated using parametric regression models
  - The weighted average is approximated using Monte Carlo simulation

The detailed steps for estimating parametric g-formula in this study is presented in Figure 3.

To elaborate on the steps of the study:

1. First, we make copies of our dataset such that everyone is adhering to the strategy in each copy. We need to build each copy from the ground-up at time 0. The values of the covariates are sampled from the empirical distribution. Then, we fit regression models for these covariates.
2. At time step t, we use the regression models fitted earlier and force the level of treatment to be the level of treatment specified by the strategy. Then, we estimate the outcome using the regression models for the outcome values.
3. Repeat 2 over all time period.
4. Average the subject risks.

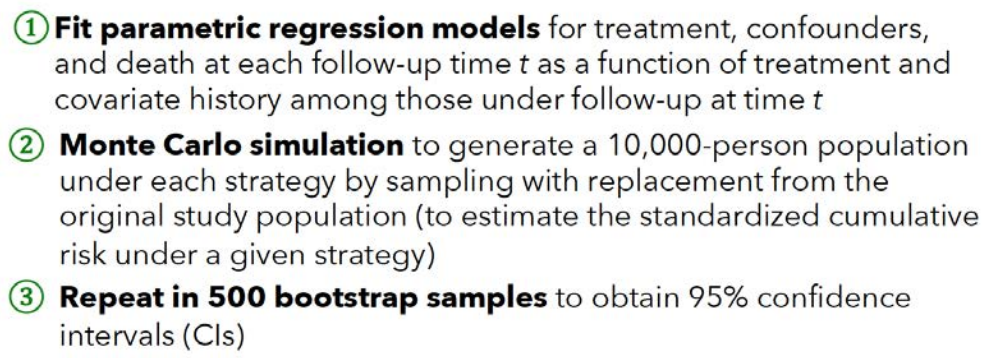

Figure 3: Detailed steps of parametric g-formula

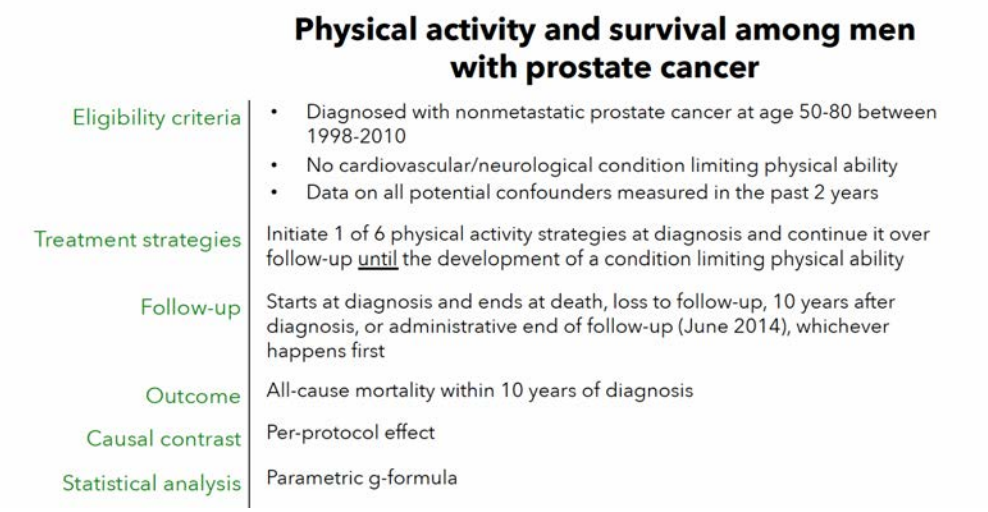

Figure 4: Workflow of the study of physical activity and survival among men with prostate cancer. The strategies and conditions have been pre-specified.

3.3 Case study: Physical activity and survival among men with prostate cancer [DGP+19]

What is the effect of adhering to guideline-based physical activity strategies on survival among men with nonmetastatic prostate cancer? To evaluate this, one possible idea is to conduct a randomized controlled trial. However, prostate cancer progresses very slowly and such a trial may take up to 10 years. Therefore, they feel the need to estimate this effect by combining high-quality observational data and G-formula. The data is leveraged from Health Professionals Follow-up Study (HPFS).

In order to apply the G-formula, they follow the three steps listing below:

1. Establish a protocol of the targeted trial that would have been helpful to conduct, if feasible
2. Measure enough covariates to adjust for confounding and achieve conditional exchangeability
3. Choose a statistical method to compare the specified treatment strategy under the assumption of conditional exchangeability intervals (CIs)

The overall workflow is detailed in 4.

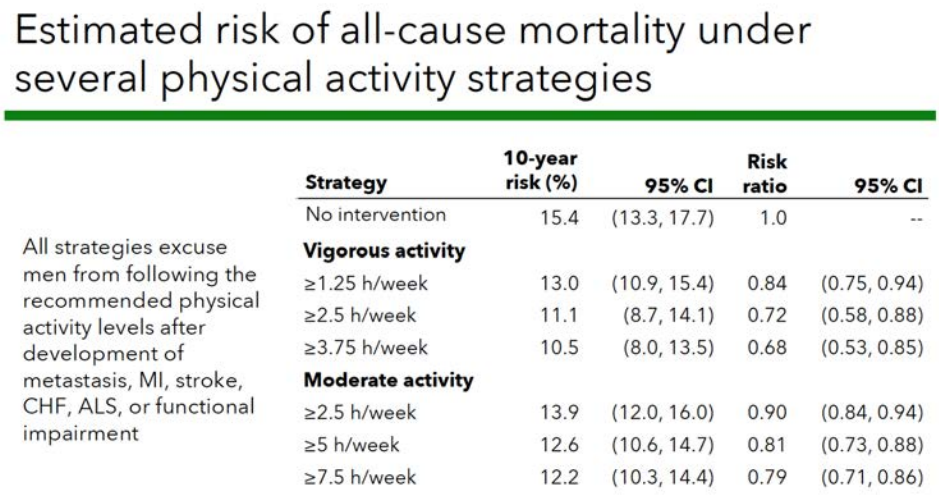

Figure 5: Main results of the study.

3.4 Primary results

The main result is presented in Figure 5. More importantly, the two main takeaways are:

1. Weekly dose/level of the intervention has been specified and based on current guidelines.
2. Covariates that make the strategy dynamic have also been prespecified. For example, all strategies excuse men from following the recommended physical activity levels after development of metastasis, MI, stroke, CHF, ALS, or functional impairment, angina pectoris, pulmonary embolism, heart rhythm disturbance, diabetes, chronic renal failure, rheumatoid arthritis, gout, ulcerative colitis or Crohns disease, emphysema, Parkinsons disease, and multiple sclerosis (all of these conditions have been "prespecified").

3.5 Sensitivity analysis

Here, we present one of the sensitivity analyses for unmeasured confounding by lag and negative outcome control.

1. Lagged physical activity and covariate data by two years
2. Negative outcome control to detect potential unmeasured confounding by clinical disease (Questionnaire non-response)

In conclusion:

G-methods let them validly estimate the effect of pre-specified dynamic strategies.

4 Discussion of "THE AI CLINICIAN" [KCB+18]

The key takeways from the discussion are:

1. The paper failed to prove whether the AI policy really has a causal effect on survival or it has an advantage due to bias.

2. The features did not capture the causal interpretation. Because the covariates are binned in 4 hours, they may overlook or may not capture the intervention at all.

3. The system also failed to capture the confounders (again, because the covariates are binned in 4-hour bins).

4. The evaluation might favor the AI policy more than other policies.

References

[DGP+19] Barbra A Dickerman, Edward Giovannucci, Claire H Pernar, Lorelei A Mucci, and Miguel A Hernán. Guideline-based physical activity and survival among us men with nonmetastatic prostate cancer. *American journal of epidemiology*, 188(3):579–586, 2019.

[KCB+18] Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716, 2018.

[KZ18] Nathan Kallus and Angela Zhou. Confounding-robust policy improvement. In *Advances in Neural Information Processing Systems*, pages 9269–9279, 2018.

[SJ15] Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*, pages 814–823, 2015.

[TB16] Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148, 2016.

6.S897 / HST.956 Machine Learning for Healthcare

Spring 2019