Topic 4: Making decisions

Eric B. Laber

Department of Statistical Science, Duke University

Statistics 561

On choices

Dad always thought laughter was the best medicine, which I guess is why several of us died of tuberculosis.
—Robert Koch

On decisions

When you're ten years old, and a car drives by and splashes a puddle of water all over you, it's hard to decide if you should go to school like that or try to go home and change and probably be late. So while he was trying to decide, I drove by and splashed him again..
—Bertrand Russell

Warm-up (5 minutes)

- Explain to your group
  - What is a randomized clinical trial? Why do we randomize?
  - What is confounding?
  - What is a one-armed bandit?
- True or false
  - Regression + randomization = causality
  - Bandit problems were invented by computer scientists
  - Laber has had food poisoning from Pizza Hut twice

Decision problems

- Nearly all statistical analyses drive decisions
  - Estimate treatment effect $\Rightarrow$ treatment recommendations
  - Identify gene associated with disease $\Rightarrow$ follow-up study
  - Model click-through-rate as function of customer + ad attributes $\Rightarrow$ ad selection for website
  - Forecast product demand $\Rightarrow$ manufacturing decisions
  - Model wins-above-replacement $\Rightarrow$ contract decisions
  - ...

Roadmap

- One-stage decision problems
- K-stage decision problems
- Contextual bandits
- Markov Decision Problems
- Freedom!

One-stage setup

- Observe $\{(\boldsymbol{X}_i, A_i, Y_i)\}_{i=1}^n$ iid from P
  - $\boldsymbol{X} \in \mathcal{X} \subseteq \mathbb{R}^p$ covariates (decision context)
  - $A \in \mathcal{A} = \{-1, 1\}$ action (treatment, intervention, decision, etc.)
  - $Y \in \mathbb{R}$ utility (outcome, output, reward, etc.) higher is better
- Goal: select actions to maximize expected utility

Policies

- $\psi : \mathcal{X} \to 2^{\mathcal{A}}$ is set of allowable actions, i.e., $\psi(\boldsymbol{x}) \subseteq \mathcal{A} \setminus \emptyset$
- Policy $\pi : \mathcal{X} \to \mathcal{A}$ such that $\pi(\boldsymbol{x}) \in \psi(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{X}$
  - Under π decision maker will select action $\pi(\boldsymbol{x})$ in context $\boldsymbol{x}$
  - Define $V(\pi) \triangleq \mathbb{E}^{\pi} Y$ to be expected utility if actions are selected according the policy π
  - Optimal policy satisfies $V(\pi^{\mathrm{opt}}) \geq V(\pi)$ for all π

Formalizing the optimal policy

- Potential outcome $Y^*(a)$ under action a, i.e., the outcome under action a (which may be contrary to what was observed)
  - Imagine each individual has two potential outcomes $Y^*(1)$ and $Y^*(-1)$, one associated with each action
  - The potential outcome under policy π is

$$Y^*(\pi) = Y^*(1)1_{\pi(\boldsymbol{X})=1} + Y^*(-1)1_{\pi(\boldsymbol{X})=-1}$$

    formally, the value of a policy is $V(\pi) = \mathbb{E}Y^*(\pi)$

laber draws a table illustrating potential outcomes

Identifying the optimal policy

- Optimal policy defined in terms of potential outcomes, need to link to data-generating model
- Standard causal assumptions
  - No unmeasured confounders: $\left\{Y^*(1), Y^(-1)\right\} \perp A|\boldsymbol{X}$
  - Consistency: $Y = Y^*(A)$, i.e., outcome is potential outcome under action taken
  - Positivity: there exists $\epsilon > 0$ such that $P(A = a|\boldsymbol{X} = \boldsymbol{x}) \geq \epsilon$ for (almost) all $\boldsymbol{x} \in \mathcal{X}$

  No interference: the action assigned to one unit do affect outcomes of others

No unmeasured confounders

- Actions may be selected according to $\boldsymbol{X}$ and the perceived impact on the outcome Y, e.g., clinical decisions
  - No unmeasured confounders says we captured all factors affecting action selection and the outcome
  - One minute: construct example in which this assumption is violated

Spillover effects

- One minute: generate three examples where this is violated
  -
  -
  -

A sad fact about your life

- Requisite causal assumption are not testable using observed data
  - There is no test, procedure, etc. that can be applied to the observed data (no matter how much there is) to evaluated needed causal conditions
  - Randomization ensures no unmeasured confounders by construction
  - Must use external information: i.e., knowledge of underlying science, richness of the features $\boldsymbol{X}$, etc.

Regression-based characterization of optimal regime

- Define $Q(\boldsymbol{x}, a) = \mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x}, A = a)$
- Under standard causal assumptions

$$V(\pi) = \mathbb{E}Q\left\{\boldsymbol{X}, \pi(\boldsymbol{X})\right\}$$

given this expression suggest an estimator of π^{opt}

Derivation of regression-based estimator

Q-learning: part I

- Bound: for any policy π it follows that

$$V(\pi) = \mathbb{E} Q\left\{\boldsymbol{X}, \pi(\boldsymbol{X})\right\} \leq \mathbb{E} \sup_{a \in \psi(\boldsymbol{X})} Q(\boldsymbol{X}, a)$$

- Note that the policy

$$\pi^{\text{opt}}(\boldsymbol{x}) = \arg \max_{a \in \psi(\boldsymbol{x})} Q(\boldsymbol{x}, a)$$

attains this bound and is thus optimal

Q-learning: part I cont'd

- Idea: estimate $Q(\boldsymbol{x}, a)$ by regressing Y on $\boldsymbol{X}, A$ to obtain $\widehat{Q}_n(\boldsymbol{x}, a)$ and subsequently $\widehat{\pi}_n(\boldsymbol{x}) = \arg \max_{a \in \psi(\boldsymbol{x})} \widehat{Q}_n(\boldsymbol{x}, a)$
- Ex. suppose $\psi(\boldsymbol{x}) = \{-1, 1\}$ and posit linear model $Q(\boldsymbol{x}, a; \beta) = \boldsymbol{x}_0^\intercal \beta_0 + a \boldsymbol{x}_1^\intercal \beta_1$ where $\beta = (\beta_0^\intercal, \beta_1^\intercal)^\intercal$ and $\boldsymbol{x}_0, \boldsymbol{x}_1$ features of $\boldsymbol{x}$
  - $\widehat{\beta}_n \triangleq \arg\min_{\beta} \mathbb{P}_n \{Y - Q(\boldsymbol{X}, A; \beta)\}^2$
  - $\widehat{\pi}_n(\boldsymbol{x}) = \arg\max_a Q(\boldsymbol{x}, a; \widehat{\beta}_n) = \operatorname{sign}(\boldsymbol{x}_1^\intercal \widehat{\beta}_{1,n})$

Q-learning: flexible models

- No need to restrict to linear models
- Construct estimator $\widehat{Q}_n(\boldsymbol{x}, a)$ of $Q(\boldsymbol{x}, a) = \mathbb{E}(Y|\boldsymbol{X} = x, A = a)$ using random forest and then take $\widehat{\pi}_n(\boldsymbol{x}) = \arg \max_{a \in \psi(\boldsymbol{x})} \widehat{Q}_n(\boldsymbol{x}, a)$
- Quantum theory of decision rules: if a regression method exists, someone, somewhere has published a paper applying it in Q-learning
  - Boosting
  - Neural nets
  - Nearest neighbors
  - Gaussian processes
  - NP-Bayes
  - ...

Advantage learning v1

- Note that we can always write

$$Q(\boldsymbol{x}, a) = \mu(\boldsymbol{x}) + a\Delta(\boldsymbol{x})$$

where

$$\mu(\boldsymbol{x}) = \frac{Q(\boldsymbol{x}, 1) + Q(\boldsymbol{x}, -1)}{2}, \text{ and } \quad \Delta(\boldsymbol{x}) = \frac{Q(\boldsymbol{x}, 1) - Q(\boldsymbol{x}, -1)}{2}$$

only need to estimate Δ to identify optimal policy

Advantage learning v1 cont'd

- Write $Q(\boldsymbol{x}, a) = \widetilde{\mu}(\boldsymbol{x}) + \{(1+a)/2 - q(\boldsymbol{x})\}\,\Delta(\boldsymbol{x})$, where $\widetilde{\mu}(\boldsymbol{x}) = \mu(\boldsymbol{x}) - q(\boldsymbol{x})\Delta(\boldsymbol{x})$, where $q(\boldsymbol{x}) = P(A = 1|\boldsymbol{X} = \boldsymbol{x})$

- Advantage learning solves

$$\widehat{\mu}_n, \widehat{\Delta}_n = \arg\min_{\widetilde{\mu},\Delta} \mathbb{P}_n \left[Y - \widetilde{\mu}(\boldsymbol{X}) + \{(A+1)/2 - q(\boldsymbol{X})\}\,\Delta(\boldsymbol{X})\right]^2$$

so that the estimated optimal policy is given by
$\widehat{\pi}_n(\boldsymbol{x}) = \operatorname{sign}\left\{\widehat{\Delta}_n(\boldsymbol{x})\right\}$ if $\psi(\boldsymbol{x}) = \{-1, 1\}$

Why A-learning v1 works

Why A-learning v1 works

Advantage learning v2

- Let $A \in \mathcal{A} \subset \mathbb{R}$ be more general action space and for simplicity assume that $\psi(\boldsymbol{x}) = \mathcal{A}$ for all $\boldsymbol{x}$
- Define $\Gamma(\boldsymbol{x}, a) = Q(\boldsymbol{x}, a) - \max_a Q(\boldsymbol{x}, a)$ so that $\Gamma(\boldsymbol{x}, a) \leq 0$ and $\Gamma(\boldsymbol{x}, a) = 0$ if $a = \pi^{\text{opt}}(\boldsymbol{x})$ then

$$Q(\boldsymbol{x}, a) = \omega(\boldsymbol{x}) + \Gamma(\boldsymbol{x}, a),$$

  where $\omega(\boldsymbol{x}) = \max_a Q(\boldsymbol{x}, a)$
- $\Gamma(\boldsymbol{x}, a)$ is the advantage of selecting action a in context $\boldsymbol{x}$ and $\pi^{\text{opt}}(\boldsymbol{x}) = \arg\max_a \Gamma(\boldsymbol{x}, a)$

Advantage learning v2

- Goal: estimate Γ without estimating ω
- Claim: Γ satisfies

$$\Gamma = \arg\min_{\gamma} \mathbb{E}\left\{ Y - \gamma(\boldsymbol{X}, A) + \int \gamma(\boldsymbol{X}, a) p(a|\boldsymbol{X}) d\eta(a) \right\}^2,$$

  where η is a dominating measure

- Note there's no ω in the above expression!

Why A-learning v2 works

Why A-learning v2 works

Classification-based representation: quiz

- Warm-up: discuss with your stats group (3.25 minutes)
  - What is sampling bias? When does it occur?
  - What is the Horvitz-Thompson estimator?
  - What is cost-sensitive classification?
- True or false
  - Double sampling is when you use the same spoon twice in the same trough of bean dip (I'm looking at you Chad)
  - Survey sampling is mostly relegated to the census and marketing
  - Laber discovered 'Cart Narcs' on YouTube at 2AM and binged every episode could find (let's keep the tough questions to a minimum)

Classification-based representation overview

- Define the propensity score $P(A = 1|\boldsymbol{X} = \boldsymbol{x})$
- Under our standard causal conditions

$$V(\pi) = P\left\{\frac{Y 1_{A=\pi(\boldsymbol{X})}}{P(A|\boldsymbol{X})}\right\}$$

  this is the classic Horvitz-Thompson (HT) estimator from survey sampling!

- HT more commonly known as inverse probability weighted (IPW) representation; What's the intuition behind this estimator?

HT representation details

IPWE

- Inverse probability weighted estimator (IPWE) of $V(\pi)$

$$\widehat{V}_n^{\mathrm{IPWE}}(\pi) = \mathbb{P}_n \left\{ \frac{Y 1_{A=\pi(\boldsymbol{X})}}{\widehat{P}_n(A|\boldsymbol{X})} \right\},$$

  where $\widehat{P}_n(a|\boldsymbol{x})$ is the estimated propensity score, e.g., estimated using logistic regression, nnet, etc.

- Estimated optimal decision rule

$$\widehat{\pi}_n = \arg\max_{\pi \in \Pi} \mathbb{P}_n \left\{ \frac{Y 1_{A=\pi(\boldsymbol{X})}}{P(A|\boldsymbol{X})} \right\}$$

  feasible for some classes Π if n isn't too large but generally not tractable

Linking IPWE with cost-sensitive classification

- A tale of brilliance and the despair of being too late

$$\begin{aligned}
\arg\max_{\pi} \widehat{V}_n^{\mathrm{IPWE}}(\pi) &= \arg\max_{\pi} \mathbb{P}_n \left\{ \frac{Y 1_{A\pi(\boldsymbol{X})\geq 0}}{\widehat{P}_n(A|\boldsymbol{X})} \right\} \\
&= \arg\min_{\pi} \mathbb{P}_n \left\{ \frac{Y 1_{A\pi(\boldsymbol{X})< 0}}{\widehat{P}_n(A|\boldsymbol{X})} \right\} \\
&= \arg\min_{\pi} \mathbb{P}_n \widehat{W}_n 1_{YA\pi(\boldsymbol{X})<0},
\end{aligned}$$

where $\widehat{W}_n = |Y|/\widehat{P}_n(A\boldsymbol{X})$

- Egad! This looks like a weighted classification problem!

Derivation of weighted classification

Convex surrogates and optimal decisions

- Consider decision rules $\pi(\boldsymbol{x}) = \operatorname{sign}\{f(\boldsymbol{x})\}$ where $f : \mathbb{R}^p \to \mathbb{R}$ is a (generally smooth) fn, e.g., $f(\boldsymbol{x}) = \boldsymbol{x}^\mathsf{T}\beta$
- Let $\phi : \mathbb{R} \to \mathbb{R}_+$ be one of our convex surrogates from classification, e.g., hinge loss, logistic loss, exp loss, etc.
- Let $\mathcal{F}$ be a class of functions from $\mathbb{R}^p$ into $\mathbb{R}$, the outcome weighted estimator (OWL) is given by $\widehat{\pi}_n(\boldsymbol{x}) = \operatorname{sign}\left\{\widehat{f}_n(\boldsymbol{x})\right\}$ where

$$\widehat{f}_n = \arg\min_{f \in \mathcal{F}} \mathbb{P}_n \widehat{W}_n \phi\{YAf(\boldsymbol{X})\}$$

OWL example

- OWL of linear decision rule is given by

$$\widehat{\beta}_n = \arg\min_{\beta} \mathbb{P}_n \widehat{W}_n \phi\left(YA\boldsymbol{X}^\intercal\beta < 0\right) + \lambda||\beta||^2$$

  so that $\widehat{\pi}_n^{\mathrm{OWL}}(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{x}^\intercal\widehat{\beta}_n)$

- Same theory applies as in classification!

Fact: IPWE is terrible

- IPWE is highly unstable b/c small propensities inflate variance and only a fraction of data is used, e.g., in a randomized trial only 1/2 the data appear (on average) in the weighted sum[1]

- Better: AIPWE which is given by

$$\widehat{V}_n^{\mathrm{AIPWE}}(\pi) = \mathbb{P}_n \left[\frac{Y 1_{A=\pi(\boldsymbol{X})}}{\widehat{P}_n(A|\boldsymbol{X})} - \frac{1_{A=\pi(\boldsymbol{X})} - \widehat{P}_N(A|\boldsymbol{X})}{\widehat{P}_n(A|\boldsymbol{X})} \widehat{Q}_n \left\{ \boldsymbol{X}, \pi(\boldsymbol{X}) \right\} \right]$$

[1] All the data are used in the estimation of the propensity score.

Why AIPWE work? Double robustness.

Why AIPWE work? Double robustness. cont'd

Roadmap

- One-stage decision problems
- **K-stage decision problems**
- Contextual bandits
- Markov Decision Problems
- Freedom!

Sequential decision problems

- Multi-stage treatment strategies
- Planning a transition to carbon-neutrality
- Navigation
- Optimizing vaccine distribution over the next three months
- ...

Setup

- Observe $\{\boldsymbol{X}_{1,i}, A_{1,i}, \boldsymbol{X}_{2,i}, A_{2,i}, \ldots, \boldsymbol{X}_{T,i}, A_{T,i}, Y_i\}_{i=1}^n$ comprising n i.i.d. trajectories drawn from unknown distn P
  - $\boldsymbol{X}_t \in \mathcal{X} \subseteq \mathbb{R}^p$ measurements at time t
  - $A_t \in \mathcal{A} = \{-1, 1\}$ action/decision/txt at time t
  - $Y \in \mathbb{R}$ outcome/utility coded so that higher is better
- Define $\boldsymbol{H}_1 = \boldsymbol{X}_1$ and $H_t = (H_{t-1}, A_{t-1}, \boldsymbol{X}_t)$ to be history, i.e., info available to decision maker before decision at time t

Multi-stage policy

- Policy $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_T)$

Thank you.

eric.laber@duke.edu

laber-labs.com