ORIE 6745

Lecture 1

Logistics

- Instructor: Nathan Kallus
- Office hours: by appointment
  - don't hesitate to reach out
- kallus@cornell.edu

- Website: nathankallus.com/6745F17
- Slack channel: 6745f17.slack.com

Who are you?

- Name
- Program
- Research interests
- Why are you here?
- What are you hoping to get out of the class?
- What are you bringing to the class?

Logistics

- Timing: Wednesdays 11:35–12:50 and 13:10–14:25
- Physical vs virtual location will change between Ithaca and NYC

Logistics

- "Grade breakdown":
  - participation and lecture scribing (25%)
    - 1-2 lectures / student
  - homework (25%)
    - 2~3 problem sets
  - final project (50%)
- But: this is a PhD class – whether you learn something is on you
  - (everyone should expect to do well as long as you do an honest try on each of the assignments)

Final project

- Can be about:
  - Critical literature survey
    - Will post some suggested papers/topics
  - A small/precursory research project
    - Will post some suggested directions
    - Can be empirical or theoretical
  - Combination
- Two parts:
  - A short report + an in-class presentation
- Tip: make the project be synergistic with your current or intended research efforts
  - Save time but, more importantly, use this as an opportunity

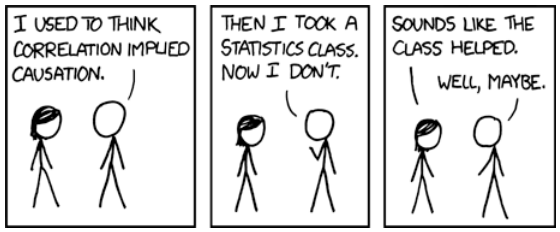
"Causality and Learning for Intelligent Decision Making" ???

- What is this class about?

What is causality?

- Did *A* happen *because* of *B?*
- Perhaps easier to understand as:
  How different would *A* be under B_1 vs B_0, keeping everything else constant?
- Correlation is not causation
  - Ice cream consumption leads to drowning deaths
  - Sleeping with shoes on leads to headaches
  - Gender discrimination in graduate admissions
- The central problem of causal inference:
  - We only know *A* under $B_{realized}$
  - *A* under $B_{not\ realized}$ is an unknown counterfactual

Association and Causation



Association and Causation

- Eating makes you faithful
  - "Will he cheat? How to tell. Ladies, you probably think that it's just in his nature. He can't help it - he HAS to cheat. But here's the sad truth: you're not feeding him enough. If you're worried your guy might cheat, try checking out his waistline. A new study says the size of his belly may reveal whether he'll stray." (match.com)
- "Murder rates affect IQ tests scores" (reuters.com)
  - "If a murder occurred in a child's neighborhood the children's test scores fell"
  - "Curiously, there were enough murders in Hispanic neighborhoods but Latino children seemed unaffected."
- "Watching too much TV can KILL you" (dailymail.co.uk)
- "Lack of sleep may shrink your brain" (CNN.com)
- "To spoon or not to spoon? After-sex affection boosts sexual and relationship satisfaction" (scienceofrelationships.com)
- "Teenage sex "leads to bad moods" in later life" (dailymail.co.uk)

Interventional vs Counterfactual

- Interventional: decision making
  - What is the expected effect of going to college at age 18 on earnings at age 50 on average in the US population?
  - Can we tell Joe age 18 what is the expected effect conditioned on his observable characteristics at age 18?
  - Should we recommend to Joe age 18 to go to college?
- Counterfactual: blame finding
  - How much more would college-uneducated Joe age 50 make if he did go to college when he was 18?
- Counterfactual often important in legal settings but also nearly impossible to identify reliably
- We will almost exclusively focus on interventional: hard enough already and sufficient for decision making

"Causality and Learning for Intelligent Decision Making" ???

- Predictions vs Decisions: passive vs active
  - Decisions have consequences, predictions don't
- Given a day on which the average price for NYC→LA flight was high, predict a best guess for the demand for the flight on that day
  - ⇒ supervised learning – just care about how close the prediction is the observed demand
  - ⇒ black box approach to ML – use any complex black box prediction and just validate
  - But doesn't inform decision making
    - High prices ⇒ high demand?! Set price to infinity?!

“Causality and Learning for Intelligent Decision Making” ???

- If we set the price high, how would the demand be affected? Need to find the *effect of the decision*
- What is the best policy for maximal causal effect? When should we price high and when low?
- But how do we even estimate this effect, make inferences about it, or find a good policy?
- This class is about:
  - How do we learn effects and policies
  - How do we use optimization to do so in efficient, optimal, and robust ways
  - How we use ML to do so when data is high dimensional and complex and we just want a good decision

Roadmap: very high level

- Prologue: making decisions when effects are already known
- Part I: estimation and inference on unknown effects
  - Settings: Controlled experiments (A/B testing), noncompliance, IVs, RDDs, unconfoundedness
  - Methods: optimal balance, regression adjustment, propensity scores, matching, generalized optimal matching, domain adaptation, double robustness
- Part II: learning to decide
  - Machine learning for heterogeneous effects
  - Learning online: contextual bandits
  - Learning offline: policy learning from observational data
- Epilogue: other consequences of decisions, fairness

How might we estimate effects?

- If we set the price high, how would the demand be affected? Need to find the *effect of the decision*
- Options:
  - Do a controlled experiment on prices to compare low vs high (next few lectures)
  - Classic: complete randomization OR assume linear
  - Such an experiment may be expensive so what's the best way to design such an experiment so we can learn a lot from a little? Need Opt
  - What if outcomes are almost a black box? Need ML

How might we estimate effects?

- If we set the price high, how would the demand be affected? Need to find the *effect of the decision*
- Options:
  - Use instrumental variable analysis (in 2-3 weeks)
  - Jet fuel prices influence flight prices
  - Jet fuel prices influence demand only via flight prices
  - Classic: assume linear + strong instrument
  - How to use ML+Opt to deal with *weak* instruments?

How might we estimate effects?

- If we set the price high, how would the demand be affected? Need to find the *effect of the decision*
- Options:
  - Use regression discontinuity (in 3-4 weeks)
  - Effect of passing (75% answers right) LSATs on earnings?
  - Whether you get 75+ϵ or 75-ϵ is essentially ***random***
  - I.e. not associated with inherent ability etc
  - Classic: assume linear or use local polynomial regression
  - But surface near the threshold may be complex and we may have high dimensional covariates to control for

How might we estimate effects?

- If we set the price high, how would the demand be affected? Need to find the *effect of the decision*
- Options:
  - Control for *all* confounding variables (in 4-5 weeks)
    - All things that could affect both price and demand: weather, competition, season, …
  - Classic: propensity scores, pairwise matching
  - ML: how to deal with high dimensional covariates
  - Opt: how to make the most of your data

Some other scientific solutions

- Discover causal *mechanisms*
  - Things move when you apply force
  - Drug helps because it bonds to receptor and blocks something else from bonding
- Perfect unit homogeneity, temporal stability, and causal transience
  - E.g.: determine freezing point and boiling point of water molecules contained in a closed box.
- The statistical solution:
  - create *statistical* homogeneity
  - compute *average* effects

Controlled vs Observational

- What's the effect of smoking on lung cancer incidence?
- What's the effect of lead in home on school performance?
- Cannot compare naively
- But no way to construct a controlled experiment
- Observational data has
  - Power: huge sizes
  - Limits: difficult identification

Why policy learning?

- What is the best decision policy? When to price high? When to price low?
- Don't need to learn an effect – just a policy
- Vapnik's maxim:
  - When solving a problem of interest, do not solve a more general problem as an intermediate step.
- A really good policy can be simple while a good outcome model must be complicated

Roadmap: on the horizon

- What is causality?
  - Continuing the discussion
  - Intro to formalism of potential outcomes
- Making decisions when effects are known
  - Intro and review of basic decision theory
  - Statistical learning generalization theory
    - Rademacher averages, Maser, Sauer, VC dim
    - How does it apply to decision making
    - Useful tools for analysis when effects are not known
- Controlled experiments (A/B testing)
  - …