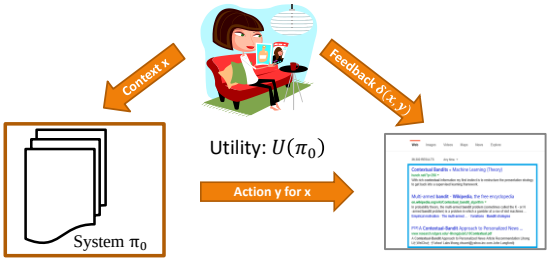
Learning to Act and Causality

CS4780/5780 – Machine Learning
Fall 2019

Nika Haghtalab & Thorsten Joachims
Cornell University

Reading:
G. Imbens, D. Rubin, Causal Inference for Statistics …, 2015. Chapter 1.

Interactive System Schematic

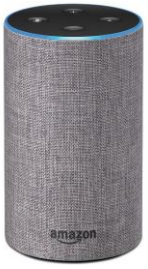


News Recommender

- Context x:
  - User
- Action y:
  - Portfolio of newsarticles
- Feedback $\delta(x,y)$:
  - Reading time in minutes

Music Voice Assistant

- Context x:
  - User and speech
- Action y:
  - Track that is played
- Feedback $\delta(x,y)$:
  - Listened to the end



Search Engine

- Context x:
  - Query
- Action y:
  - Ranking
- Feedback $\delta(x,y)$:
  - Click / no-click

Log Data from Interactive Systems

- Data

context, π_0 action, reward / loss

$$S = \big((x_1, y_1, \delta_1), \ldots, (x_n, y_n, \delta_n)\big)$$

→ Partial Information (aka "Contextual Bandit") Feedback

- Properties
  - Contexts x_i drawn i.i.d. from unknown $P(X)$
  - Actions y_i selected by existing system π_0: $X \to Y$
  - Feedback δ_i from unknown function $\delta: X \times Y \to \Re$

[Zadrozny et al., 2003] [Langford

Goal

Use interaction log data

$$S = \big((x_1, y_1, \delta_1), \ldots, (x_n, y_n, \delta_n)\big)$$

- for evaluation of system π
  - Offline estimate of online performace of some system π.
  - System π can be different from π_0 that generated log.
- for learning new system π

Evaluation: Outline

- Offline Evaluating of Online Metrics
  - A/B Testing (on-policy)
    → Counterfactual estimation from logs (off-policy)
- Approach 1: "Model the world"
  - Imputation via reward prediction
- Approach 2: "Model the bias"
  - Counterfactual model and selection bias
  - Inverse propensity scoring (IPS) estimator

Online Performance Metrics

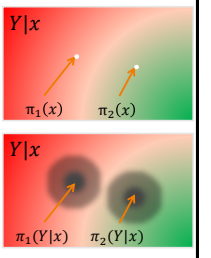
Example metrics
- CTR
- Revenue
- Time-to-success
- Interleaving
- Etc.

→ Correct choice depends on application and is not the focus of this lecture.

This lecture:
Metric encoded as $\delta(x, y)$ [click/payoff/time for (x,y) pair]

System

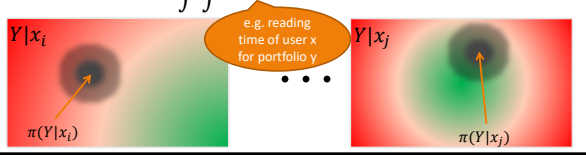
- Definition [Deterministic Policy]:
  Function
  $$y = \pi(x)$$
  that picks action y for context x.
- Definition [Stochastic Policy]:
  Distribution
  $$\pi(y|x)$$
  that samples action y given context x



System Performance

Definition [Utility of Policy]:
The expected reward / utility $U(\pi)$ of policy π is

$$U(\pi) = \int\int \delta(x, y)\pi(y|x)P(x)\,dx\,dy$$



Online Evaluation: A/B Testing

Given $S = \big((x_1, y_1, \delta_1), \ldots, (x_n, y_n, \delta_n)\big)$ collected under π_0,

$$\widehat{U}(\pi_0) = \frac{1}{n}\sum_{i=1}^{n} \delta_i$$

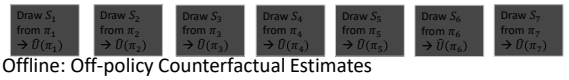
→ A/B Testing
Deploy π_1: Draw $x \sim P(X)$, predict $y \sim \pi_1(Y|x)$, get $\delta(x, y)$
Deploy π_2: Draw $x \sim P(X)$, predict $y \sim \pi_2(Y|x)$, get $\delta(x, y)$
⋮
Deploy $\pi_{|H|}$: Draw $x \sim P(X)$, predict $y \sim \pi_{|H|}(Y|x)$, get $\delta(x, y)$

Pros and Cons of A/B Testing

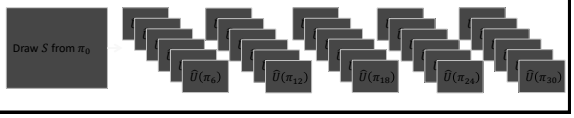
- Pro
  - User centric measure
  - No need for manual ratings
  - No user/expert mismatch
- Cons
  - Requires interactive experimental control
  - Risk of fielding a bad or buggy π_i
  - Number of A/B Tests limited
  - Long turnaround time

Evaluating Online Metrics Offline

- Online: On-policy A/B Test



- Offline: Off-policy Counterfactual Estimates

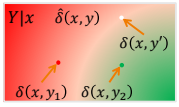


Evaluation: Outline

- Offline Evaluating of Online Metrics
  - A/B Testing (on-policy)
    → Counterfactual estimation from logs (off-policy)
- Approach 1: "Model the world"
  - Imputation via reward prediction
- Approach 2: "Model the bias"
  - Counterfactual model and selection bias
  - Inverse propensity scoring (IPS) estimator

Approach 1: Reward Predictor

- Idea:
  - Use $S = ((x_1, y_1, \delta_1), \ldots, (x_n, y_n, \delta_n))$ from π_0 to estimate reward predictor $\hat{\delta}(x, y)$

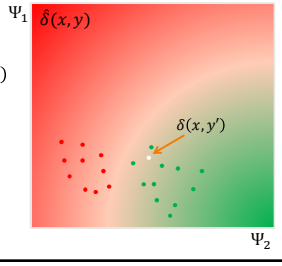


- Deterministic π: Simulated A/B Testing with predicted $\hat{\delta}(x, y)$
  - For actions $y_i' = \pi(x_i)$ from new policy π, generate predicted log $S' = \left(\left(x_1, y_1', \hat{\delta}(x_1, y_1')\right), \ldots, \left(x_n, y_n', \hat{\delta}(x_n, y_n')\right)\right)$
  - Estimate performace of π via $\hat{U}_{rp}(\pi) = \frac{1}{n}\sum_{i=1}^{n} \hat{\delta}(x_i, y_i')$
- Stochastic π: $\hat{U}_{rp}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\sum_{y} \hat{\delta}(x_i, y)\,\pi(y|x_i)$

Regression for Reward Prediction

Learn $\hat{\delta}: x \times y \to \Re$

1. Represent via features $\Psi(x, y)$
2. Learn regression based on $\Psi(x, y)$ from S collected under π_0
3. Predict $\hat{\delta}(x, y')$ for $y' = \pi(x)$ of new policy π



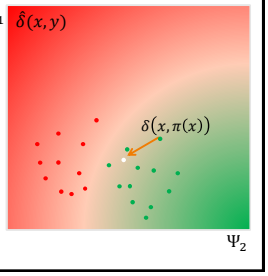
Problems of Reward Predictor

- Modeling bias
  - choice of features and model
- Selection bias
  - π_0's actions are over-represented

→

$$\hat{U}_{rp}(\pi) = \frac{1}{n}\sum_{i} \hat{\delta}(x_i, \pi(x_i))$$

Can be unreliable and biased



Evaluation: Outline

- Offline Evaluating of Online Metrics
  - A/B Testing (on-policy)
    → Counterfactual estimation from logs (off-policy)
- Approach 1: "Model the world"
  - Imputation via reward prediction
- Approach 2: "Model the bias"
  - Counterfactual model and selection bias
  - Inverse propensity scoring (IPS) estimator

Approach "Model the Bias"

- Idea:
  Fix the mismatch between the distribution $\pi_0(Y|x)$ that generated the data and the distribution $\pi(Y|x)$ we aim to evaluate.

$$\mathrm{U}(\pi) = \int\int \delta(x,y)\frac{\pi(y|x)}{\pi_0(y|x)}P(x)\,dx\,dy$$

Potential Outcome Model

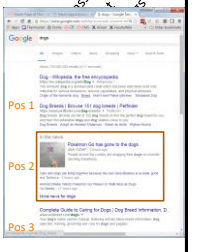
- Example: Treating Heart Attacks
  - Treatments: Y
    - Bypass / Stent / Drugs
  - Chosen treatment for patient x_i: y_i
  - Outcomes: δ_i
    - 5-year survival: 0 / 1
  - Which treatment is best?

Patients $x_i, i \in \{1,\dots,n\}$

Bypass	Stent	Drugs
0	1	1
0	1	1
1	0	1
0	1	0
1		

Counterfactual Model

- Example: Treating Heart Attacks
  Placing Vertical
  - Treatments: Y
    - Bypass / Stent / Drugs
      Pos 1 / Pos 2/ Pos 3
  - Chosen treatment for patient x_i: y_i
  - Outcomes: δ_i
    - 5-year survival: 0 / 1
      Click / no Click on SERP
  - Which treatment is best?



Potential Outcome Model

- Example: Treating Heart Attacks
  - Treatments: Y
    - Bypass / Stent / Drugs
  - Chosen treatment for patient x_i: y_i
  - Outcomes: δ_i
    - 5-year survival: 0 / 1
  - Which treatment is best?
    - Everybody Drugs
    - Everybody Stent
    - Everybody Bypass
      → Drugs 3/4, Stent 2/3, Bypass 2/4 – really?

Patients $x_i, i \in \{1,\dots,n\}$

Bypass	Stent	Drugs
0	1	1
0	1	1
1	0	1
0	1	0
1		

Treatment Effects

- Average Treatment Effect of Treatment y
  - $\mathrm{U}(y) = \frac{1}{n}\sum_i \delta(x_i, y)$
- Example
  - $\mathrm{U}(bypass) = \frac{5}{11}$
  - $\mathrm{U}(stent) = \frac{7}{11}$
  - $\mathrm{U}(drugs) = \frac{3}{11}$

Factual Outcome

Counterfactual Outcomes

Patients

Bypass	Stent	Drugs
0	1	0
1	1	0
0	0	1
0	0	0
0	1	1
1	0	0
1	0	1
0	1	0
0	1	0
1	1	0
1	1	0

Assignment Mechanism

- Probabilistic Treatment Assignment
  - For patient i: $\pi_0(Y_i = y|x_i)$
  - Selection Bias
- Inverse Propensity Score Estimator
  - $\hat{U}_{ips}(y) = \frac{1}{n}\sum_i \frac{\mathbb{I}\{y_i = y\}}{p_i}\delta(x_i, y_i)$
  - Propensity: $p_i = \pi_0(Y_i = y_i|x_i)$
  - Unbiased: $E[\hat{U}(y)] = U(y)$, if $\pi_0(Y_i = y|x_i) > 0$ for all i
- Example
  - $\hat{U}(drugs) = \frac{1}{11}\left(\frac{1}{0.8} + \frac{1}{0.7} + \frac{1}{0.8} + \frac{0}{0.1}\right) = 0.36 < 0.75$

$\pi_0(Y_i = y|x_i)$

0.3	0.6	0.1
0.5	0.4	0.1
0.1	0.1	0.8
0.6	0.3	0.1
0.2	0.5	0.7
0.7	0.2	0.1
0.1	0.1	0.8
0.1	0.8	0.1
0.3	0.3	0.4
0.3	0.6	0.1
0.4	0.4	0.2

Patients

Bypass	Stent	Drugs
0	1	0
1	1	0
0	0	1
0	0	0
0	1	1
1	0	0
1	0	0
0	1	0
0	1	0
1	1	0
1	1	0

Interventional vs Observational

- Interventional Controlled Experiment
  - Assignment Mechanism under our control
  - Propensities $p_i = \pi_0(Y_i = y_i|x_i)$ are known by design
  - Requirement: $\forall y: \pi_0(Y_i = y|x_i) > 0$ (probabilistic)
- Observational Study
  - Assignment Mechanism not under our control
  - Propensities p_i need to be estimated
  - Estimate $\hat{\pi}_0(Y_i|z_i) = \pi_0(Y_i|x_i)$ based on features z_i
  - Requirement: $\hat{\pi}_0(Y_i|z_i) = \hat{\pi}_0(Y_i|\delta_i, z_i)$ (unconfounded)

Conditional Treatment Policies

- Policy (deterministic)
  - Context x_i describing patient
  - Pick treatment y_i based on x_i: $y_i = \pi(x_i)$
  - Example policy:
    - $\pi(A) = drugs, \pi(B) = stent, \pi(C) = bypass$
- Average Treatment Effect
  - $U(\pi) = \frac{1}{n}\sum_i \delta(x_i, \pi(x_i))$
- IPS Estimator
  - $\hat{U}_{ips}(\pi) = \frac{1}{n}\sum_i \frac{\mathbb{I}\{y_i = \pi(x_i)\}}{p_i}\delta(x_i, y_i)$

Patients

Bypass	Stent	Drugs	X
0	1	0	B
1	1	0	C
0	0	1	A
0	0	0	B
0	1	1	A
1	0	0	B
1	0	1	A
0	1	0	C
0	1	0	A
1	1	0	C
1	1	0	B

Stochastic Treatment Policies

- Policy (stochastic)
  - Context x_i describing patient
  - Pick treatment y based on x_i: $\pi(Y|x_i)$
- Note
  - Assignment Mechanism is a stochastic policy as well!
- Average Treatment Effect
  - $U(\pi) = \frac{1}{n}\sum_i \sum_y \delta(x_i, y)\pi(y|x_i)$
- IPS Estimator
  - $\hat{U}(\pi) = \frac{1}{n}\sum_i \frac{\pi(y_i|x_i)}{p_i}\delta(x_i, y_i)$

Patients

Bypass	Stent	Drugs	X
0	1	0	B
1	1	0	C
0	0	1	A
0	0	0	B
0	1	1	A
1	0	0	B
1	0	1	A
0	1	0	C
0	1	0	A
1	1	0	C
1	1	0	B

Counterfactual Model = Logs

Recorded in Log

Context x_i	
Treatment y_i	
Outcome δ_i	
Propensities p_i	
New Policy π	
T-effect $U(\pi)$	Average quality of new policy.

Evaluation: Outline

- Evaluating Online Metrics Offline
  - A/B Testing (on-policy)
    → Counterfactual estimation from logs (off-policy)
- Approach 1: "Model the world"
  - Estimation via reward prediction
- Approach 2: "Model the bias"
  - Counterfactual Model
  - Inverse propensity scoring (IPS) estimator

System Evaluation via Inverse Propensity Score Weighting

Definition [IPS Utility Estimator]:
Given $S = \big((x_1, y_1, \delta_1), \dots, (x_n, y_n, \delta_n)\big)$ collected under π_0,

$$\hat{U}_{ips}(\pi) = \frac{1}{n}\sum_{i=1}^{n} \delta_i \frac{\pi(y_i|x_i)}{\pi_0(y_i|x_i)}$$

Propensity p_i

→ Unbiased estimate of utility for any π, if propensity nonzero whenever $\pi(y_i|x_i) > 0$.

Note:

If $\pi = \pi_0$, then online A/B Test with

→ Off-policy vs. On-policy estimation.

$$\hat{U}_{ips}(\pi_0) = \frac{1}{n}\sum_i \delta_i$$

[Horvitz & Thompson, 1952] [Rubin, 1983] [Zadrozny et al., 2003]

Illustration of IPS

IPS Estimator:

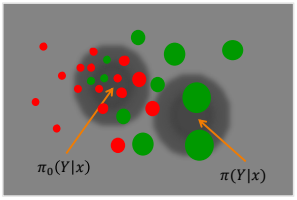
$$\hat{U}_{IPS}(\pi) = \frac{1}{n}\sum_i \frac{\pi(y_i|x_i)}{\pi_0(y_i|x_i)}\delta_i$$

Unbiased:

If

then: $\forall y: \pi(y|x)P(x) > 0 \rightarrow \pi_0(y|x) > 0$

$$\mathrm{E}[\hat{U}_{IPS}(\pi)] = U(\pi)$$



IPS Estimator is Unbiased

$$E[\hat{U}_{IPS}(\pi)] = \frac{1}{n}\sum_{x_1,y_1}\dots\sum_{x_n,y_n}\left[\sum_i \frac{\pi(y_i|x_i)}{\pi_0(y_i|x_i)}\delta(x_i,y_i)\right]\pi_0(y_1|x_1)\dots\pi_0(y_n|x_n)P(x_1)\dots P(x_n)$$

independent

$$= \frac{1}{n}\sum_{x_1,y_1}\pi_0(y_1|x_1)P(x_1)\dots\sum_{x_n,y_n}\pi_0(y_n|x_n)P(x_n)\left[\sum_i \frac{\pi(y_i|x_i)}{\pi_0(y_i|x_i)}\delta(x_i,y_i)\right]$$

$$= \frac{1}{n}\sum_i\sum_{x_1,y_1}\pi_0(y_1|x_1)P(x_1)\dots\sum_{x_n,y_n}\pi_0(y_n|x_n)P(x_n)\left[\frac{\pi(y_i|x_i)}{\pi_0(y_i|x_i)}\delta(x_i,y_i)\right]$$

marginal

$$= \frac{1}{n}\sum_i\sum_{x_i,y_i}\pi_0(y_i|x_i)P(x_i)\left[\frac{\pi(y_i|x_i)}{\pi_0(y_i|x_i)}\delta(x_i,y_i)\right]$$

full support

$$= \frac{1}{n}\sum_i\sum_{x_i,y_i}P(x_i)\pi(y_i|x_i)\delta(x_i,y_i) = \frac{1}{n}\sum_i \mathrm{U}(\pi) = U(\pi)$$

identical x,y

Counterfactual Policy Evaluation

- Controlled Experiment Setting:
  - Log data: $D = \big((x_1, y_1, \delta_1, p_1), \dots, (x_n, y_n, \delta_n, p_n)\big)$
- Observational Setting:
  - Log data: $D = \big((x_1, y_1, \delta_1, z_1), \dots, (x_n, y_n, \delta_n, z_n)\big)$
  - Estimate propensities: $p_i = P(y_i|x_i, z_i)$ based on x_i and other confounders z_i

→ Goal: Estimate average treatment effect of new policy π.

- IPS Estimator

$$\hat{U}(\pi) = \frac{1}{n}\sum_i \delta_i \frac{\pi(y_i|x_i)}{p_i}$$

or many others.

Evaluation: Summary

- Offline Evaluation of Online Metrics
  - A/B Testing (on-policy)
    → Counterfactual estimation from logs (off-policy)
- Approach 1: "Model the world"
  - Estimation via reward prediction
  - Pro: low variance
  - Con: model mismatch can lead to high bias
- Approach 2: "Model the bias"
  - Counterfactual Model
  - Inverse propensity scoring (IPS) estimator
  - Pro: unbiased for known propensities
  - Con: large variance

From Evaluation to Learning

Setting: Batch Learning from Bandit Feedback (BLBF)

- "Model the World" Learning:
  - Learn: $\hat{\delta}: x \times y \rightarrow \Re$
  - Derive Policy:

$$\pi(y|x) = \operatorname{argmin}_{y'}\big[\hat{\delta}(x, y')\big]$$

- "Model the Bias" Learning:
  - Find policy that optimizes IPS training error

$$\pi = \operatorname{argmin}_{\pi'}\left[\sum_i \frac{\pi'(y_i|x_i)}{\pi_0(y_i|x_i)}\delta_i\right]$$