Counterfactual Model for Online Systems

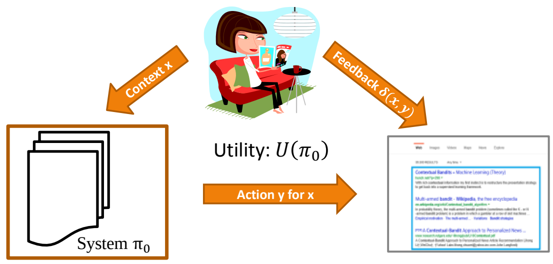
CS 7792 - Fall 2016

Thorsten Joachims

Department of Computer Science & Department of Information Science
Cornell University

Imbens, Rubin, Causal Inference for Statistical Social Science, 2015. Chapters 1,3,12.

Interactive System Schematic



News Recommender

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

Ad Placement

- Context x:
  - User and page
- Action y:
  - Ad that is placed
- Feedback $\delta(x,y)$:
  - Click / no-click

Search Engine

- Context x:
  - Query
- Action y:
  - Ranking
- Feedback $\delta(x,y)$:
  - win/loss against baseline in interleaving

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 & Li], [Bottou, et al., 2014]

Goal

- Use interaction log data
$$S = \big((x_1, y_1, \delta_1), \dots, (x_n, y_n, \delta_n)\big)$$
for evaluation of system π:
  - Estimate online measures of some system π offline.
  - System π can be different from π_0 that generated log.

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

Online Performance Metrics

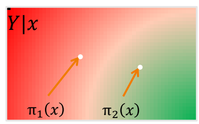
Example metrics
- CTR
- Revenue
- Time-to-success
- Interleaving
- Etc.

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

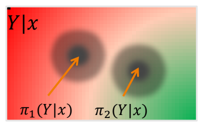
This lecture:
Metric encoded as $\delta(x, y)$ [click/payoff/time for (x,y) pair]

System

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



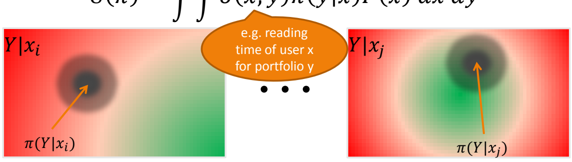
- Definition [Stochastic Policy]:
Distribution
$$\pi(y|x)$$
that samples action y given context x



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 $S = \big((x_1, y_1, \delta_1), \dots, (x_n, y_n, \delta_n)\big)$ collected under π_0,
$$\hat{U}(\pi_0) = \frac{1}{n}\sum_{i=1}^{n} \delta_i$$

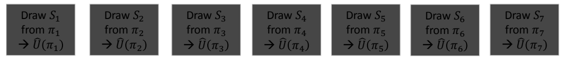
→ A/B Testing
Deploy π_1: Draw $x \sim P(X)$, predict $y \sim \pi_1(Y|x)$, get $\delta(x, y)$
Deploy π_2: Draw $x \sim P(X)$, predict $y \sim \pi_2(Y|x)$, get $\delta(x, y)$
⋮
Deploy $\pi_{|H|}$: Draw $x \sim P(X)$, predict $y \sim \pi_{|H|}(Y|x)$, get $\delta(x, y)$

Pros and Cons of A/B Testing

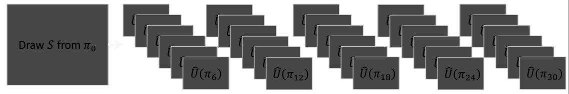
- Pro
  - User centric measure
  - No need for manual ratings
  - No user/expert mismatch
- Cons
  - Requires interactive experimental control
  - Risk of fielding a bad or buggy π_i
  - Number of A/B Tests limited
  - Long turnaround time

Evaluating Online Metrics Offline

- Online: On-policy A/B Test



- Offline: Off-policy Counterfactual Estimates

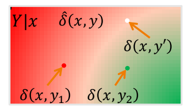


Evaluation: Outline

- 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

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 $\hat{\delta}(x, y)$

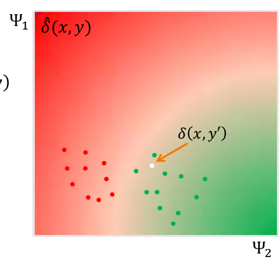


- Deterministic π: Simulated A/B Testing with predicted $\hat{\delta}(x, y)$
  - For actions $y_i' = \pi(x_i)$ from new policy π, generate predicted log $S' = ((x_1, y_1', \hat{\delta}(x_1, y_1')), \ldots, (x_n, y_n', \hat{\delta}(x_n, y_n')))$
  - Estimate performace of π via $\hat{U}_{rp}(\pi) = \frac{1}{n}\sum_{i=1}^{n}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

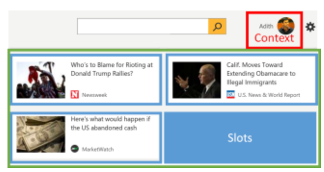
Learn $\hat{\delta}: x \times y \rightarrow \Re$

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



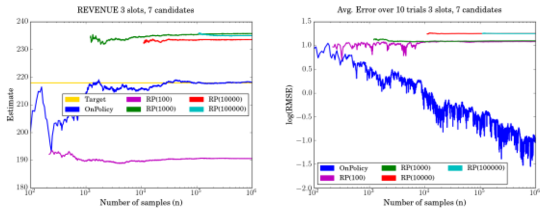
News Recommender: Exp Setup

- Context x: User profile
- Action y: Ranking
  - Pick from 7 candidates to place into 3 slots
- Reward δ: "Revenue"
  - Complicated hidden function



- Logging policy π_0: Non-uniform randomized logging system
  - Placket-Luce "explore around current production ranker"

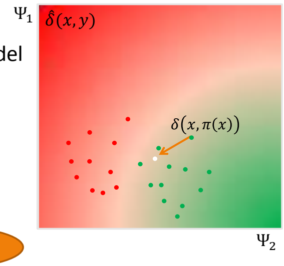
News Recommender: Results



RP is inaccurate even with more training and logged data

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

- 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

Approach "Model the Bias"

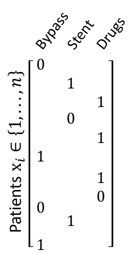
- Idea:

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

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

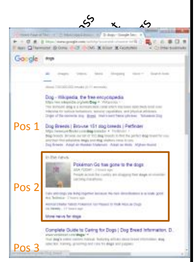
Counterfactual Model

- 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?



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?



Counterfactual 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
0		
	1	
1		

Treatment Effects

- Average Treatment Effect of Treatment y
  - $U(y) = \frac{1}{n} \sum_i \delta(x_i, y)$
- Example
  - $U(bypass) = \frac{4}{11}$
  - $U(stent) = \frac{6}{11}$
  - $U(drugs) = \frac{3}{11}$

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	1
0	1	0
0	1	0
1	1	0
1	1	0

Experimental vs Observational

- 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

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

Recorded in Log

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 Scoring

Definition [IPS Utility Estimator]:
Given $S = \left((x_1, y_1, \delta_1), \ldots, (x_n, y_n, \delta_n)\right)$ 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 $\pi(y_i|x_i) > 0$.

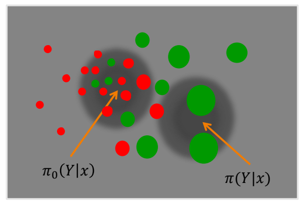
Note:
If $\pi = \pi_0$, then online A/B Test with $\hat{U}_{ips}(\pi_0) = \frac{1}{n}\sum_i \delta_i$
→ Off-policy vs. On-policy estimation.

[Horvitz & Thompson, 1952] [Rubin, 1983] [Zadrozny et al., 2003] [Li et al., 2011]

Illustration of IPS

IPS Estimator:

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



IPS Estimator is Unbiased

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

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

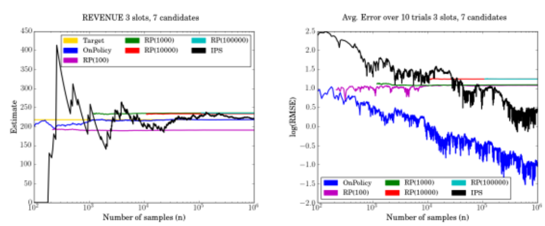
$$= \frac{1}{n}\sum_i \sum_{x_1,y_1} \pi_0(y_1|x_1) P(x_1) \ldots \sum_{x_n,y_n} \pi_0(y_n|x_n) P(x_n) \left[\frac{\pi(y_i|x_i)}{\pi_0(y_i|x_i)} \delta(x_i, y_i)\right]$$

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

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

Probabilistic Assignment

News Recommender: Results



IPS eventually beats RP; variance decays as $O\left(\frac{1}{\sqrt{n}}\right)$

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