ML Bootcamp - Causality Class

Eustache Diemert

Outline

Causality as an AI frontier

Tools & Methods for Causality

Coffee Break

Optimization of Causal Effects

Lunch Break

Causality as an AI frontier

Causality as an AI frontier

- Examples of Causal Failures
- A Famous Failure ...
- ... and some Easy Fixes

Tools & Methods for Causality

Coffee Break

Optimization of Causal Effects

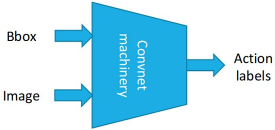
Lunch Break

Causal Failures

Struggles of modern AI

Deep Learning is solving AI, right ?

Example: detection of the action *"giving a phone call"*



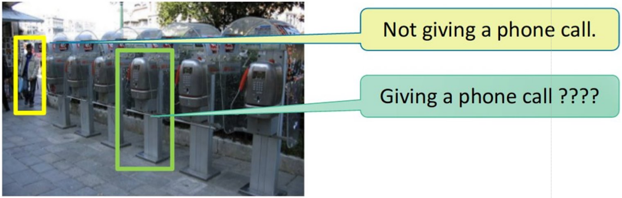
(Oquab et al., CVPR 2014)
~70% correct (SOTA in 2014)

Causal Failures

Struggles of modern AI

Deep Learning is solving AI, right ?

Example: detection of the action "*giving a phone call*"



Example from Léon Bottou '15

Causal Failures

Struggles of modern AI

Deep Learning is solving AI, right ?

Example: detection of the action *"giving a phone call"*

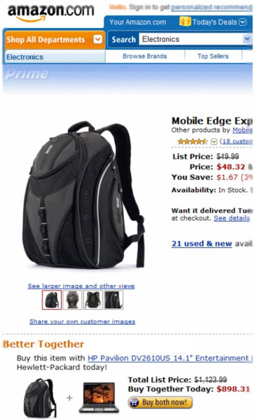
"Phone Call" action
causes
"Person" + "Phone" in image
(but not vice versa !)

Example from Léon Bottou '15

Causal Failures

Struggles of modern AI (2)

With loads of data we can solve AI problems, right ?



Example from Bernard Schoelkopf '17

Struggles of modern AI (2)

With loads of data we can solve AI problems, right ?

"Buying a laptop"
causes
"Buying a 15 in. backpack"
(but not vice versa !)

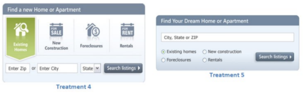
Example from Bernard Schoelkopf '17

12

A solution

A/B test: Experiment where

- subjects are *randomly* assigned to groups
- groups are treated *differently*



Why do A/B Tests ?

Goals

- identify an effect: *is A influencing B ?*
- evaluate the impact of a change: *by how much ?*

Can't we rely on *expertise* ?

Can't we rely on *available data* ?

Kidney Stone Treatment: First Impression

- Two treatment variants, *prescribed by physician*:
  - $Treatment_1$: open surgical procedure
  - $Treatment_2$: ultrasonic probes and small puncture procedure

	$Treatment_1$	$Treatment_2$
Total	273/350= 78%	289/350= **83%**

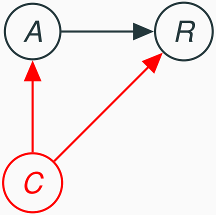
Table 1: Recovery rate

Quick Poll

Is an A/B test[1] necessary to conclude ?

[1] in medecine and social science the proper name is Randomized Control Trial

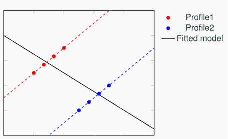
- Simpson's paradox
- *Confounder*: stone size



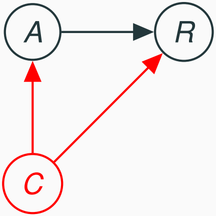
	$Treatment_1$	$Treatment_2$
Total	273/350= 78%	289/350= **83%**
Small stone	81/87= **93%**	234/270= 87%
Big stone	192/263= **73%**	55/80= 69%

Table 2: Recovery rate

- Confounding bias pulls the observed causal relation away from the true association
- Here it inverted the true causal effect



- Simpson's paradox
- *Confounder*: stone size



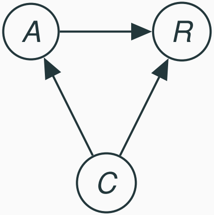
	$Treatment_1$	$Treatment_2$
Total	273/350= 78%	289/350= **83%**
Small stone	81/87= **93%**	234/270= 87%
Big stone	192/263= **73%**	55/80= 69%

Table 3: Recovery rate

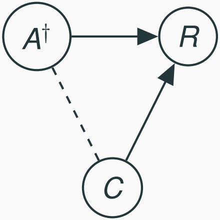
- **Fix #1: put potential confounders in the model**

Kidney Stone Treatment: Fix #2

- not possible to put *all* confounders in the model
- **Fix #2: break confounders to avoid the effects of confounders**



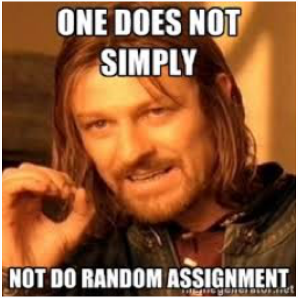
intervention $\Rightarrow$



- *assign* treatment at random to avoid selection bias

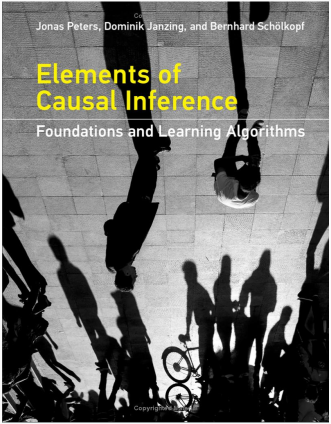
Causality 101 Summary

- Beware of observables
- Randomize to evaluate the causal impact of a change



Tools & Methods for Causality

This section is heavily lifted from Peters, Janzik and Schoelkopf book and presentation



Notation

- A, B event
- X, Y, Z random variable
- x value of a random variable
- Pr probability measure
- P_X probability distribution of X
- p density
- p_X or $p(X)$ density of P_X
- $p(x)$ density of P_X evaluated at the point x
- always assume the existence of a joint density, w.r.t. a product measure

Common Cause Principle

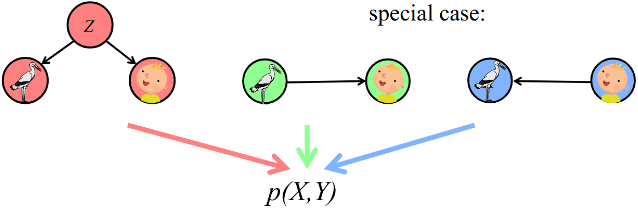
(Reichenbach)

by permission of the University of Pittsburgh.

(i) if X and Y are statistically dependent, then there exists Z *causally* influencing both;

(ii) Z screens X and Y from each other (given Z, X und Y become independent)

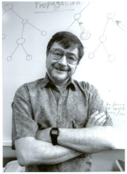
special case:



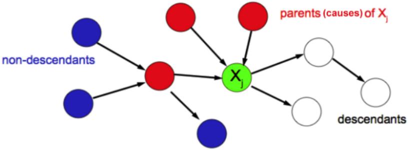
$\sum_z p(x|z)p(y|z)p(z)$ $\qquad$ $p(x)p(y|x)$ $\qquad$ $p(x|y)p(y)$

Definition of a Structural Causal Model

(Pearl et al.)



- directed acyclic graph G with vertices $X_1, \ldots, X_n$
  (following arrows does not lead to loops)
- Semantics: vertices = observables, arrows = direct causation
- $X_i := f_i(\mathrm{PA}_i, U_i)$, with independent RVs $U_1, \ldots, U_n$ that possess a joint density
- U_i stands for "unexplained" (alternatively "noise" or "exogenous variable")
- this is also called a *(nonlinear) structural equation model*



Pearl's do-notation

- Motivation: goal of causality is to infer the effect of interventions
- distribution of Y given that X is set to x:

$$p(Y|do\,X = x) \quad \text{or} \quad p(Y|do\,x)$$

- don't confuse it with $P(Y|x)$
- can be computed from p and G

Difference between seeing and doing

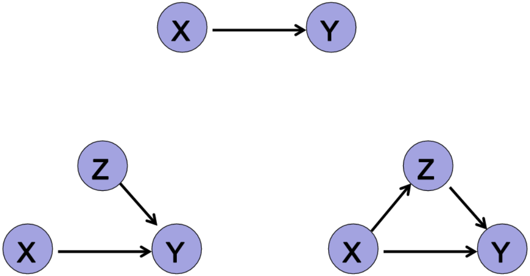
$$p(y|x)$$

probability that someone gets 100 years old given that we know that he/she drinks 10 cups of coffee per day

$$p(y|do\,x)$$

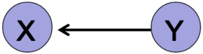
probability that some randomly chosen person gets 100 years old after he/she has been forced to drink 10 cups of coffee per day

Examples for $p(.|do\,x) = p(.|x)$

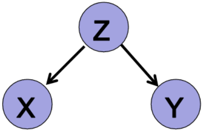


Examples for $p(.|do\,x) \neq p(.|x)$

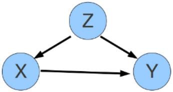
- $p(Y|do\,x) = P(Y) \neq P(Y|x)$



- $p(Y|do\,x) = P(Y) \neq P(Y|x)$



Example: controlling for confounding



$X \not\perp\!\!\!\perp Y$ partly due to the confounder Z and partly due to $X \to Y$

- causal factorization

$$p(X, Y, Z) = p(Z)p(X|Z)p(Y|X, Z)$$

- replace $P(X|Z)$ with δ_{Xx}

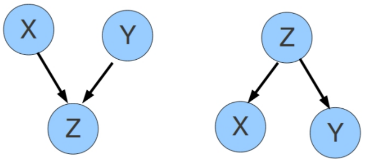
$$p(Y, Z|do\, x) = p(Z)\ \delta_{Xx}\ p(Y|X, Z)$$

- marginalize

$$p(Y|do\, x) = \sum_z p(z)p(Y|x, z) \neq \sum_z p(z|x)p(Y|x, z) = p(Y|x)\,.$$

Asymmetry under inverting arrows

(Reichenbach 1956)



$X \perp\!\!\!\perp Y$ $\quad\quad$ $X \not\!\perp\!\!\!\perp Y$

$X \not\!\perp\!\!\!\perp Y \,|\, Z$ $\quad\quad$ $X \perp\!\!\!\perp Y \,|\, Z$

Algorithmic construction of causal hypotheses

IC algorithm by Verma & Pearl (1990) to reconstruct DAG from p

idea:

1. Construct skeleton
2. Find v-structures
3. direct further edges that follow from
   - graph is acyclic
   - all v-structures have been found in 2)

Causal Inference Method

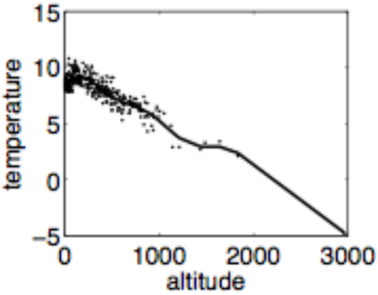
Prefer the causal direction that can better be fit with an additive noise model.

Implementation:

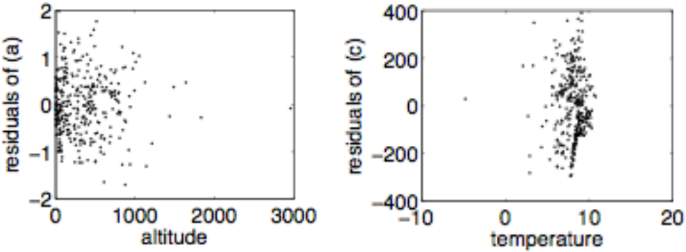
- Compute a function f as non-linear regression of X on Y
- Compute the residual
$$N_Y := Y - f(X)$$
- check whether N_Y and X are statistically independent (uncorrelated is not enough)

Experiments

Relation between altitude (cause) and average temperature (effect) of places in Germany



Causal Discovery (simplified)



Our independence tests detect strong dependence.
Hence the method prefers the correct direction

altitude → temperature

Brain Research

Question:
X and Y are
correlated... but
why ?

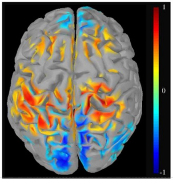
Success Story

Causal inference in brain research

Grosse-Wentrup, Janzing, Siegel, Schölkopf, NeuroImage 2016

Let X, Y be some brain state features and S some randomized experimental condition (i.e., a parentless node!) Assume

$$
\begin{aligned}
S &\not\perp\!\!\!\perp X \\
S &\not\perp\!\!\!\perp Y \\
S &\perp\!\!\!\perp Y \mid X
\end{aligned}
$$



then Markov condition and faithfulness imply

$$S \to X \to Y$$

applied to:
X: γ-power in the parietal cortex
Y: γ-power in the medial prefrontal cortex
S instruction to up- or down-regulate X
(conditional independence verified via regression)

Question:
X and Y are
correlated... but
why ?

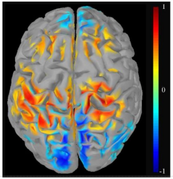
Success Story

Causal inference in brain research

Grosse-Wentrup, Janzing, Siegel, Schölkopf, NeuroImage 2016

Let X, Y be some brain state features and S some randomized experimental condition (i.e., a parentless node!) Assume

$$S \not\perp\!\!\!\perp X$$
$$S \not\perp\!\!\!\perp Y$$
$$S \perp\!\!\!\perp Y \mid X$$



th... and faithfulness imply

intervention

$$S \to X \to Y$$

applied

X: γ-p... r in the parietal cortex

Y: γ-p... r in the medial prefrontal cortex

S instruction to up- or down-regulate X

(conditional independence verified via regression)

Brain Research

Success Story

Causal inference in brain research

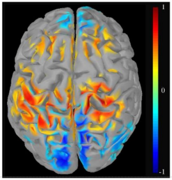
Question:
X and Y are
correlated... but
why ?

(conditional)
Independence
tests

ölkopf, NeuroImage 2016

rain state features and S some randomized experimental
condi ... (i.e., a p ... s node!) Assume

$$S \not\perp\!\!\!\perp X$$
$$S \not\perp\!\!\!\perp Y$$
$$S \perp\!\!\!\perp Y \mid X$$



intervention

and faithfulness imply

$$S \to X \to Y$$

applied
X: γ-p ... r in the parietal cortex
Y: γ-p ... r in the medial prefrontal cortex
S instruction to up- or down-regulate X
(conditional independence verified via regression)

Brain Research

Success Story

Causal inference in brain research

(conditional) Independence tests

Robust conclusion about the underlying causal mechanism

Question: X and Y are correlated... but why ?

...ölkopf, NeuroImage 2016

...rain state features and S some random ... node!) Assume

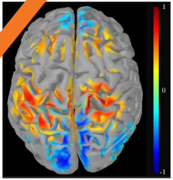
$$S \not\perp\!\!\!\perp X$$
$$S \not\perp\!\!\!\perp Y$$
$$S \perp\!\!\!\perp Y \mid X$$

...and faithfulness imp...

intervention

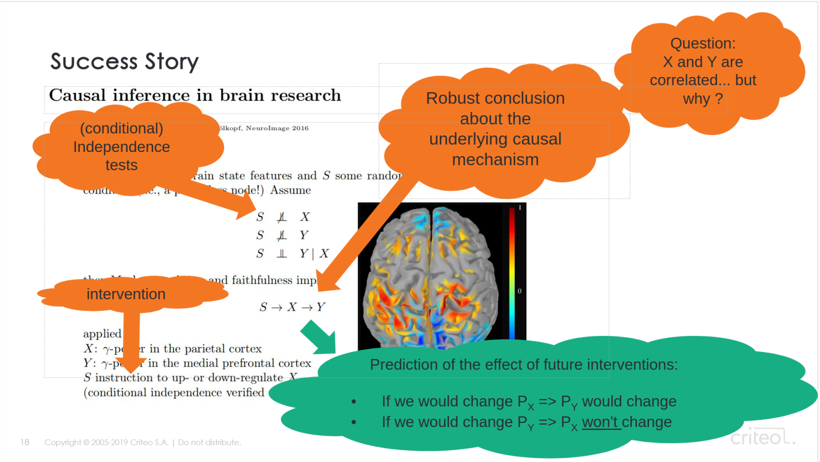
$$S \to X \to Y$$

applied
X: γ-p...r in the parietal cortex
Y: γ-p...r in the medial prefrontal cortex
S instruction to up- or down-regulate X
(conditional independence verified via regression)



17 Copyright © 2005-2019 Criteo S.A. | Do not distribute.

Brain Research



Interventions (e.g. A/B tests) are powerful causal tools, yet can be costly

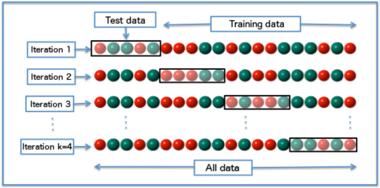
- candidate change must be production-ready (tests, interaction with other tests, certifications . . .)
- a poor candidate can make you loose money real fast (or die . . .)

Need to find a good candidate *before* A/B test (and roll-out !)

Good Practice

Validate model on historical data

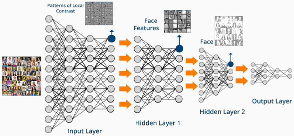
- technical metrics: RMSE, LLH, F1, . . .
- variance estimation: cross-validation

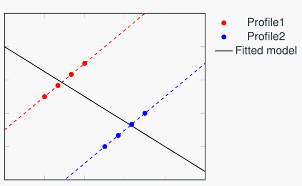


But technical metrics may differ from application/business metrics. E.g. better click prediction doesn't necessarily translates into additional revenue. . .

Idea #1: Prediction aka Direct Method

After all we got nowadays lots of powerful learning models . . .





Why can't we predict performance from historical data ?

Policy $\pi : X \rightarrow A$

$\mathbb{E}_{\pi'}[R(\pi'(X))] \approx \mathbb{E}_{\pi}[\hat{R}(\pi'(X))]$?

- dangerous when π influences the next X
- even fanciest model not immune to confounders

Idea #1.5 : Replaying actions

In digital advertising, auctions are run to win ad placements. Ideas is to replay past, won auctions.

Utility metric:

$$U(p') = \sum_{i \in \mathscr{D}} (v_i - c_i)\mathbb{I}[p'_i v_i > c_i]$$

- p / p' : reference/candidate value prediction model
- r: value of placement if auction won
- c: best opponent bid = cost of winning auction

Noise injection:

$$c \sim \Gamma(a, b); \mathbb{E}[\Gamma(a, b)] = c$$

Still, no access to lost auctions ...

Idea #2: Counter-factual Reasoning

- Goal: evaluate multiple possible policies and what would happen if we were to apply them.
- Idea: randomize the production policy to estimate *all* possible variants

When you can intervene (i.e. randomize), counterfactual reasoning is the most powerful tool to optimize a system.

Importance Sampling

- For each display draw a random variable for color selection
- Choose one of two colors:
  - Red with probability 0.80
  - Green with probability 0.20

	Proba(red)	Proba(green)
Production	80 %	20 %
Test	20 %	80 %

	Clicks(red)	Clicks(green)
Production	1000	300
Test	?	?

Importance Sampling

	Proba(red)	Proba(green)
Production	80 %	20 %
Test	20 %	80 %

	Clicks(red)	Clicks(green)
Production	1000	300
Test	250	1200

$$Clicks_{Test} = \frac{1}{4} \times Clicks_{Prod}(Red) + 4 \times Clicks_{Prod}(Green)$$

with $\frac{1}{4} = \mathbb{P}_{Test}(Red)/\mathbb{P}_{Prod}(Red)$

$4 = \mathbb{P}_{Test}(Green)/\mathbb{P}_{Prod}(Green)$

Importance Sampling

$$Clicks_{Test} = \frac{\mathbb{P}_{Test}(Red)}{\mathbb{P}_{Prod}(Red)} \times Clicks_{Prod}(Red) + \frac{\mathbb{P}_{Test}(Green)}{\mathbb{P}_{Prod}(Green)} \times Clicks_{Prod}(Green)$$

$$\mathbb{E}_{Test}[R] \simeq \sum_{i \in samples} w_i \times R_i$$

Where:

- $w_i := \frac{\mathbb{P}_{Test}(A=a_i)}{\mathbb{P}_{Prod}(A=a_i)}$
- a_i the action sampled (under Prod distribution) on line i
- R_i reward observed on line i

Coffee Break

Coffee Break

Questions/Discussion welcome during the break :-)

Optimization of Causal Effects

Causality as an AI frontier

Tools & Methods for Causality

Coffee Break

Optimization of Causal Effects

- Learning under changing distribution: Counterfactual Learning
- Application: Recommender Systems

Lunch Break

Predicting vs Acting

Predicting

- annotating categories
- predicting a click/sale probability
- recognize items in images

Acting

- placing a bid
- recommending products
- changing our margin

What could go wrong ?

engine begins to buy some new traffic / FBL reaction / Advertiser adjust CPC / Publisher adapts reserve prices...

Interpolation vs extrapolation

criteo.

Recommender System $A \sim \pi(X)$: best-of or personalized

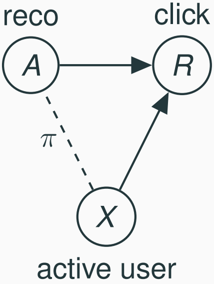
Data Collection: dormant (best-of) – active (personalized)

Objective: learn to personalize on dormant

best of

personalized

- Personalized products only known in the context of active users
- Need to explore reco on dormants

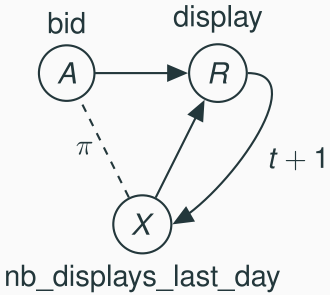


Acting changes input distribution

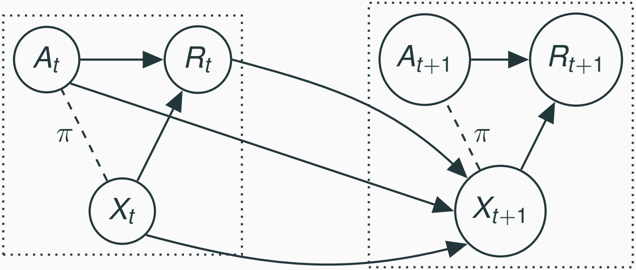
Bidder: bid $\sim \pi$(nb_displays_last_day)

Objective: Change bidder to bid more on low fatigue and less on high fatigue.

- Increasing bids on low fatigue will increase the fatigue
- Several solutions:
  - consider an action to be a sequence of bids
  - move to reinforcement learning setting



Reinforcement Learning

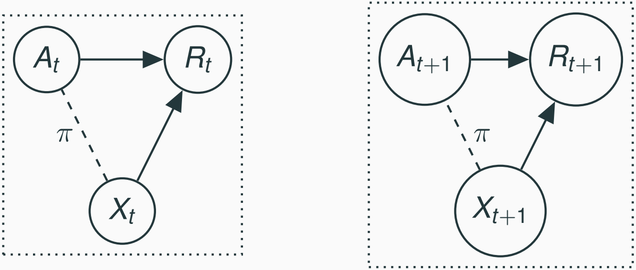


Reward is a bit complex:

$$\mathbb{E}_\pi[R] = \sum_x d^\pi(x) \sum_a \pi(a|x)\mathbb{E}[R|x, a]$$

where $d_\pi(x)$ is the stationary distribution of states under π.

Logged Bandit Feedback



Reward is a bit complex:

$$\mathbb{E}_\pi[R] = \sum_x \mathbb{P}(x) \sum_a \pi(a|x)\mathbb{E}[R|x, a]$$

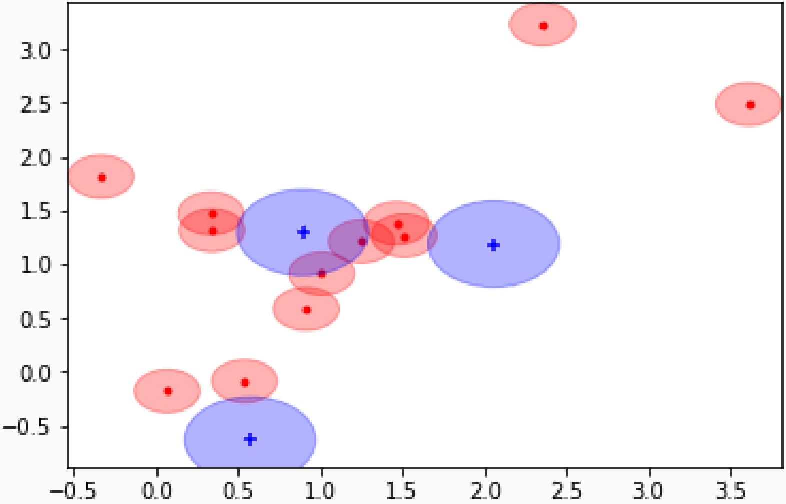
where $\mathbb{P}(x)$ does not depends on π.

Counterfactual Risk Minimization a.k.a. Policy Gradient++

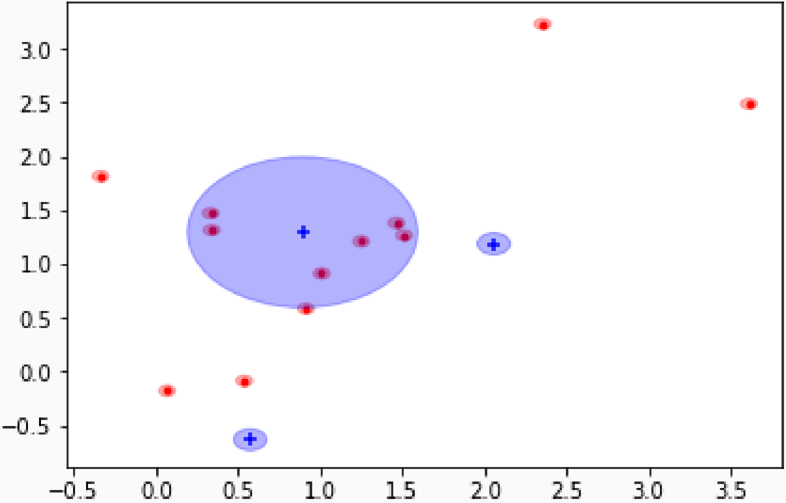
$$R(\pi_\theta, \pi_0) = \frac{1}{N} \sum_i r_i w_i(\theta) + \lambda \sqrt{\frac{V_N(r_i w_i(\theta))}{N}}$$

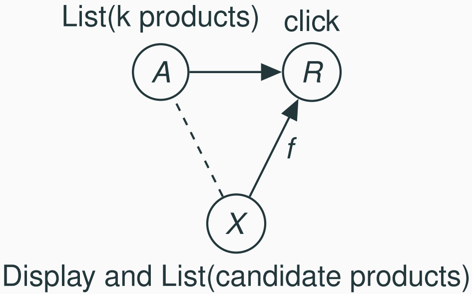
where $w_i(\theta) = \frac{\pi_\theta(a_i|x_i)}{\pi_0(a_i|x_i)}$.

What we would like



What happens





Action: choosing a list of products, from a list of candidates.

Choosing final products from candidates

Sampling the list of products:

- Computing one prediction per candidate product
  - Proba(a Sale matched to this product | it is displayed)
- Product Score := $\textit{Prediction}^{\alpha}$ (Usually α = 4)
- Repeat k times
  - Pick one product with probabilities proportional to scores
  - Add it to the list of displayed products
  - Remove it from the candidates

Probability of one banner – Example

$\mathbb{P}(A = (p_a, p_b))$?

Product	p_a	p_b	p_c	p_d	Total
Prediction	2	3	1	2	

Probability of one banner – Example

$\mathbb{P}(A = (p_a, p_b))$?

Product	p_a	p_b	p_c	p_d	Total
Prediction	2	3	1	2	
Score (alpha = 2)	4	9	1	4	18

Probability of one banner – Example

$\mathbb{P}(A = (p_a, p_b))$?

Product	p_a	p_b	p_c	p_d	Total
Prediction	2	3	1	2	
Score (alpha = 2)	4	9	1	4	18
$\mathbb{P}$(First item)	**4/18**	9/18	1/18	4/18	1

Probability of one banner – Example

$\mathbb{P}(A = (p_a, p_b))$?

Product	p_a	p_b	p_c	p_d	Total
Prediction	2	3	1	2	
Score (alpha = 2)	4	9	1	4	18
$\mathbb{P}$(First item)	**4/18**	9/18	1/18	4/18	1
Score for rank 2		9	1	4	14

Probability of one banner – Example

$\mathbb{P}(A = (p_a, p_b))$?

Product	p_a	p_b	p_c	p_d	Total
Prediction	2	3	1	2	
Score (alpha = 2)	4	9	1	4	18
$\mathbb{P}$(First item)	**4/18**	9/18	1/18	4/18	1
Score for rank 2		9	1	4	14
$\mathbb{P}$(second item)		**9/14**	1/14	4/14	1

Probability of one banner – Example

$\mathbb{P}(A = (p_a, p_b))$?

Product	p_a	p_b	p_c	p_d	Total
Prediction	2	3	1	2	
Score (alpha = 2)	4	9	1	4	18
$\mathbb{P}$(First item)	**4/18**	9/18	1/18	4/18	1
Score for rank 2		9	1	4	14
$\mathbb{P}$(second item)		**9/14**	1/14	4/14	1

Computing banner probability with chain rule:

$$\mathbb{P}(A = (p_a, p_b)) = \mathbb{P}(First = p_a) \times \mathbb{P}(Second = p_b | First = p_a)$$

Here:

$$\mathbb{P}(A = (p_a, p_b)) = 4/18 \times 9/14$$

Countefactual estimator for recommendation

$$\mathbb{E}_{Test}(R) \approx \sum_{i \in displays} r_i \times \frac{\mathbb{P}_{Test}(A = a_i)}{\mathbb{P}_{Prod}(A = a_i)}$$

Countefactual estimator for recommendation

$$\mathbb{E}_{Test}(R) \approx \sum_{i \in displays} r_i \times \frac{\mathbb{P}_{Test}(A = a_i)}{\mathbb{P}_{Prod}(A = a_i)}$$

We log:

- a description of all the candidates
- their predictions from prod model
- the chosen action (which products were displayed)
- the reward r_i

Countefactual estimator for recommendation

$$\mathbb{E}_{Test}(R) \approx \sum_{i \in displays} r_i \times \frac{\mathbb{P}_{Test}(A = a_i)}{\mathbb{P}_{Prod}(A = a_i)}$$

We log:

- a description of all the candidates
- their predictions from prod model
- the chosen action (which products were displayed)
- the reward r_i

We can recompute:

- products scores for prod model and $\mathbb{P}_{Prod}(A = a_i)$
- the predictions of a test model on each candidate
- products scores for test model and $\mathbb{P}_{Test}(A = a_i)$

Practical problems with oabt on Reco

Three difficulties:

- Non recommendable items
- Large importance weights
- Independance beween displays ?

What reco really does:

- Score each candidate
- Repeat
  - Sample one item
  - Check if this item is recommendable
- Until it found k recommendable items

We do not know if non selected items are recommendable.

Why do non recommendable items matter ?

Banner of size 1. $\mathbb{P}(Action = (p_b))$?

Product	p_a , Not recommendable	p_b	Total
Score	1.0	1.0	2.0
$\mathbb{P}$(Selected First)	1/2	1/2	1

Banner of size 1. $\mathbb{P}(Action = (p_b))$?

Product	p_a , Not recommendable	p_b	Total
Score	1.0	1.0	2.0
$\mathbb{P}$(Selected First)	1/2	1/2	1

$$\mathbb{P}(Action = display(p_b)) = 1$$

Why do non recommendable items matter ?

Banner of size 1. $\mathbb{P}(Action = (p_b))$?

Product	p_a , Not recommendable	p_b	Total
Score	1.0	1.0	2.0
$\mathbb{P}$(Selected First)	1/2	1/2	1

$$\mathbb{P}(Action = display(p_b)) = 1$$

But when we select p_b first, we estimate:

$$\mathbb{P}(Action = display(p_b)) = 0.5$$

The problem with large weights

Product	P_a	P_b	P_c	P_d	Total
Score Prod	1	1	1	100	103
Score Test	1000	100	10	1	1111

$$\mathbb{P}_{Prod}(A = (P_a, P_b, P_c)) = \frac{1}{103 \times 102 \times 101} \approx 10^{-6}$$
$$\mathbb{P}_{Test}(A = (P_a, P_b, P_c)) = \frac{1000 \times 100 \times 10}{1111 \times 111 \times 11} \approx 0.7$$

$$IW(A = (P_a, P_b, P_c)) = \mathbb{P}_{Test} / \mathbb{P}_{Prod} \approx 10^{+6}$$

With a probability of 10^{-6}, we get a weight of 10^{+6}

=> Huge variance

Capped estimator:

$$\sum_{i \in Displays} r_i \times Min(w_i, C)$$

Lower variance, but biased. This estimator tells that policies very different from prod would perform very badly.

Naive solution: capping the weights

Capped estimator:

$$\sum_{i \in Displays} r_i \times Min(w_i, C)$$

Lower variance, but biased. This estimator tells that policies very different from prod would perform very badly.

In practice, no value with both:

- acceptable variance
- acceptable bias

Capping and normalizing

Idea:

- $\mathbb{E}(Min(w_i, C)) < 1$
- Trying to compensate this by a global factor, estimating how the weights where reduced on average.

Normalized estimator: =

$$\sum_{i \in Displays} r_i \times Min(w_i, C) \times \frac{N}{\sum_{i \in Displays} Min(w_i, C)}$$

Toy Example: «New source for low context»

- Two kind of displays :
  - Low context ($\mathbb{E}(R)$ low whatever the action)
  - High context ($\mathbb{E}(R)$ high whatever the action)
- On High context Test algo = Prod algo
  - w = 1 on high context
- On low context: Test algo choose actions (almost) never chosen by Prod
  - w = 0 or » C on low context
  - $\sum Min(w_i, C) \approx 0$

$$Normalized = \sum_{i \in Displays} R_i \times Min(w_i, C) \times \frac{N}{\sum_{i \in Displays} Min(w_i, C)}$$

$$\approx (\sum_{highcontext} R_i \times Min(w_i, C)) \times \frac{N}{\sum_{highcontext} Min(w_i, C)}$$

$$\approx \mathbb{E}(R|highcontext) \times N$$

Lunch Break

Questions/Discussion welcome during lunch break :-)

This course contains material from

- Elements of Causal Inference (Peters, Janzing, Schoelkopf - MIT Press - 2017)
- Causality Tutorial (Calauzènes, Gilotte, Diemert, Aslan - Criteo ML Big Days - 2018)