Counterfactual reasoning and learning systems

Bottou, Peters et al.

Presented for R255: Causal Inference by Denizhan Akar and Andrei Alexandru

Part I: counterfactual reasoning

Denizhan Akar

Causality… affects

-

Part II: counterfactual analysis

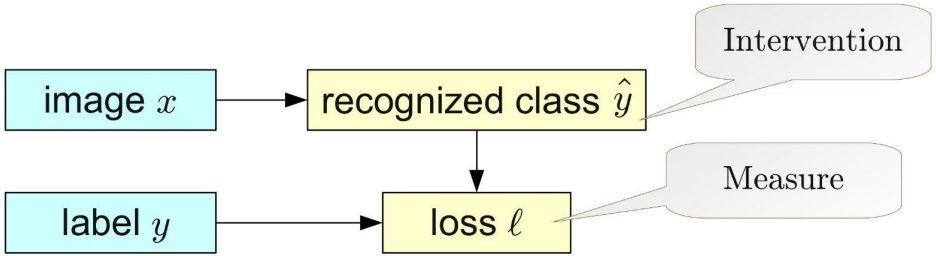
Andrei Alexandru

Enter causal inference

- We've set the stage: we want to find some ad layout *L* which maximises click yield *Y* without running A/B testing*
  - Reminder: A/B testing means that half of all users see layout A, other half layout B
- We want to use **counterfactual reasoning** to understand how changes to the layout would impact click yield
  - Using data we've already gathered!
- Will keep technical notes to a minimum, but the full derivations are not hard to follow along if you're interested!

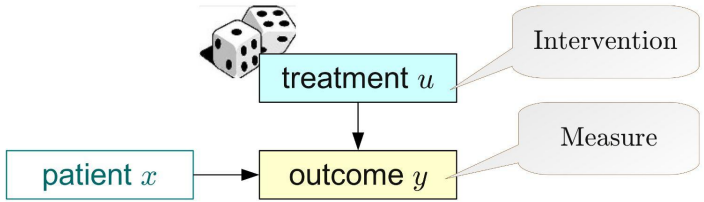
Counterfactual reasoning

- “Counterfactual” is shorthand for “what would have happened if…?”
- Example 1: training a neural network with backprop can be seen through a counterfactual lens
  - We “play” the data
  - We intervene on the parameters
  - We “replay” the data
  - Rinse and repeat



Counterfactual reasoning

- “Counterfactual” is shorthand for “what would have happened if…?”
- Example 2: reweighting randomised trials
  - Say you split patients equally to two treatments, A and B
  - Overall effectiveness of the experiment is $Y = (Y_A + Y_B)/2$
  - What if instead we want to find the effectiveness of some experiment where we apply treatment A with probability *p*, otherwise treatment B?
  - Answer: reweight the original trials: $Y^* \approx pY_A + (1-p)Y_B$



Back to our ad model…

- We start from a structural equation model as above
- This generates a Markov factorisation – ω is just shorthand for the joint distribution of all variables

$$P(\boldsymbol{\omega}) = P(u,v)P(x|u)P(a|x,v)P(b|x,v)P(q|x,a) \\ \times P(s|a,q,b)P(c|a,q,b)P(y|s,u)P(z|y,c)$$

- We model an intervention as changing one factor in the Markov factorisation, in this case changing the scoring function

$$P^*(\boldsymbol{\omega}) = P(u,v)P(x|u)P(a|x,v)P(b|x,v)P^*(q|x,a) \\ \times P(s|a,q,b)P(c|a,q,b)P(y|s,u)P(z|y,c)$$

- Given this alternative factorisation, we want to estimate some desired quantity
- In our case, a good choice is **click yield** – the number of ad clicks per page

$$Y^* = \int_\omega y P^*(\boldsymbol{\omega}) d\boldsymbol{\omega} = \int_\omega y \frac{P^*(q|x,a)}{P(q|x,a)} P(\boldsymbol{\omega}) d\boldsymbol{\omega} \approx \frac{1}{n} \sum_{i=1}^{n} y_i \frac{P^*(q_i|x_i,a_i)}{P(q_i|x_i,a_i)}$$

- What's going on here?
  - P*(ω) is being substituted with P(ω) multiplied by some ratio
  - We sample from P(ω) because we know it (the instantiation to actual values of q_i, x_i etc.)
  - y is the number of clicks
- If this looks familiar, it's because it's **importance sampling**

Importance sampling

- More generally, we can estimate the counterfactual expectation of any quantity *l(ω):*

$$Y^* = \int_\omega l(\boldsymbol{\omega}) P^*(\boldsymbol{\omega}) d\boldsymbol{\omega} = \int_\omega l(\boldsymbol{\omega}) \frac{P^*(\boldsymbol{\omega})}{P(\boldsymbol{\omega})} P(\boldsymbol{\omega}) d\boldsymbol{\omega} \approx \frac{1}{n} \sum_{i=1}^{n} \ell(\omega_i) w_i$$

- With weights:

$$w_i = w(\boldsymbol{\omega}_i) = \frac{P^*(\boldsymbol{\omega}_i)}{P(\boldsymbol{\omega}_i)} = \frac{\text{factors in } P^*(\boldsymbol{\omega}_i) \text{ but not in } P(\boldsymbol{\omega}_i)}{\text{factors in } P(\boldsymbol{\omega}_i) \text{ but not in } P^*(\boldsymbol{\omega}_i)}$$

- Great, we're done!

Not so fast

- Our sampling weights depend on factors which are stochastic in nature, there is noise in their output
- We want the numerator, P*, to be non-zero whenever P is non-zero
- Which means that the counterfactual factors need to be stochastic themselves, i.e. our experiment needs to be **randomised**
  - This means I can't just cherry-pick some values, run it once, get some result and call it a day
  - I must randomise it and run it multiple times to get something decent out
- Because of this randomness, a single estimate of our quantity of interest is no longer enough; we need to know how **confident** we are in that estimate

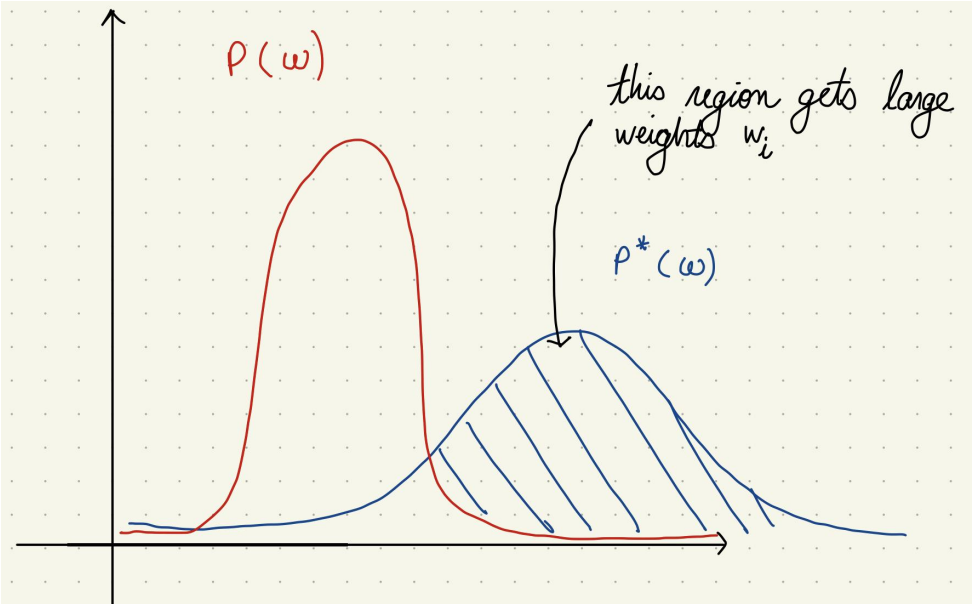
Confidence intervals

- There's a clever way of getting confidence intervals for any distribution given it has a finite variance: you use the central limit theorem
- Reminder: CLT says that given a sequence of i.i.d random variables $\{X_1, \ldots, X_n\}$ the random variable $\sqrt{n}(\bar{X}_n - \mu)$ converges in distribution to a normal $N(0, \sigma^2)$ as the length of the sequence goes to infinity[1]
- **Problem**: in importance sampling, the two distributions need to overlap fairly well for us to get an unbiased estimator

[1] Click here for a refresher

What's the issue?

- The counterfactual distribution and the actual distribution of our model don't always overlap
- When the counterfactual distribution assigns probability mass in regions where the original distribution has none, our weights are very large



What might this look like in practice?

- Imagine that your counterfactual is a change that works so well it gives you a click yield in the millions, when your previous highest value for the same amount of traffic was 1,000.
- It's really *really* unlikely that your original distribution assigns any real probability to this scenario
- But your counterfactual says that's possible; in fact, that's the whole point!

Solution

- Clip the weights such that in those domains that are poorly explored by the original distribution, the resulting weight is 0.

$$\bar{w}(\boldsymbol{\omega}) = \begin{cases} w(\boldsymbol{\omega}) & P^*(\boldsymbol{\omega}) < RP(\boldsymbol{\omega}) \\ 0 & \text{otherwise} \end{cases}$$

- With R being an empirically chosen reweighting ratio
  - This is a limitation: R should in theory be chosen before seeing the data, but the authors select this such that they get consistent results in practice

Solution

- Now, decompose the click yield we want to estimate, Y*, into two terms:

$$Y^* = \int_{\omega \in \Omega_R} l(\boldsymbol{\omega}) P^*(\boldsymbol{\omega}) d\boldsymbol{\omega} + \int_{\omega \in \Omega \setminus \Omega_R} l(\boldsymbol{\omega}) P^*(\boldsymbol{\omega}) d\boldsymbol{\omega} = \bar{Y}^* + (Y^* - \bar{Y}^*)$$

- Where Ω_R is the set of weights ω that satisfy the constraint on the previous slide
- We call Y* bar the clipped expectation, and it's much easier to estimate because clipped weights are bounded by R

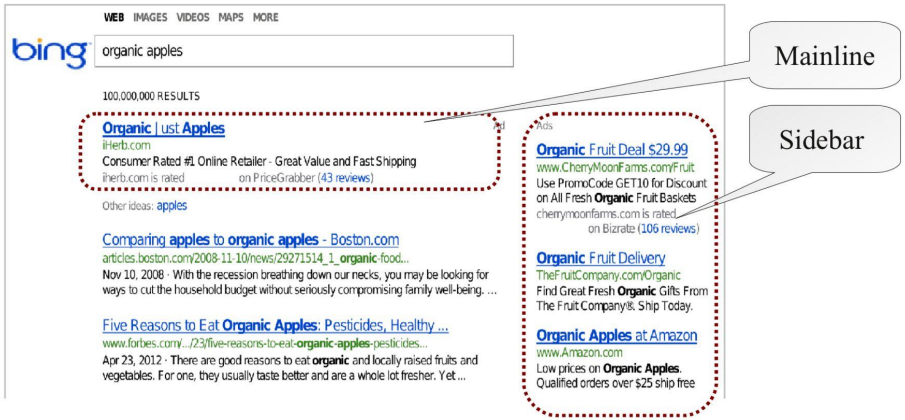
$$\bar{Y}^* = \int_{\omega \in \Omega_R} l(\boldsymbol{\omega}) P^*(\boldsymbol{\omega}) d\boldsymbol{\omega} = \int_{\omega \in \Omega_R} l(\boldsymbol{\omega}) \bar{w}(\boldsymbol{\omega}) P(\boldsymbol{\omega}) d\boldsymbol{\omega} \approx \hat{Y}^* = \frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{\omega}_i) \bar{w}(\boldsymbol{\omega}_i)$$

Confidence intervals

- We now have a quantity with finite variance, to which we can apply the central limit theorem to get confidence intervals
- The paper uses two types of confidence intervals, which differ slightly in how the quantity Y* is bounded
- There's an **inner confidence interval** that captures the uncertainty from not exploring the domain G_R that is high in probability in P* but not in P
  - If this is wide, we may have to adjust how we collect data so we get better coverage
- There's an **outer confidence interval** that captures uncertainty from a limited sample size

An experiment: mainline reserve

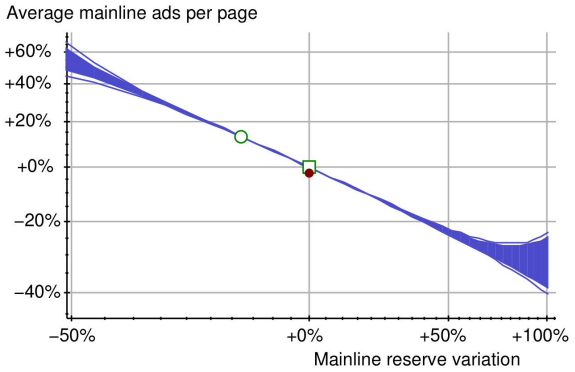
- Mainline reserve := a threshold that the rank-score of an ad needs to clear for it to be included in the mainline – the main search section – rather than in the sidebar
  - Scale this up: fewer ads clear the threshold, fewer ads in the main search section
  - Scale down: more ads clear the threshold and end up in the mainline section

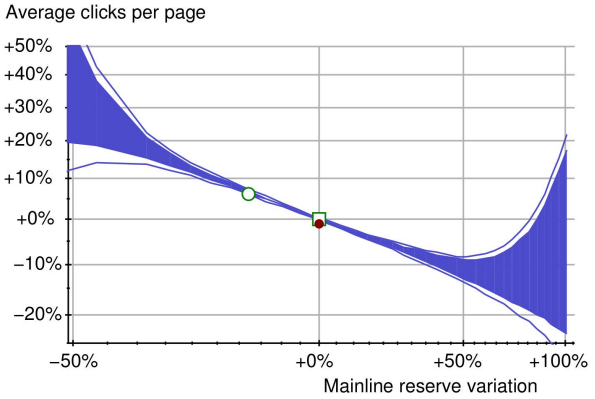


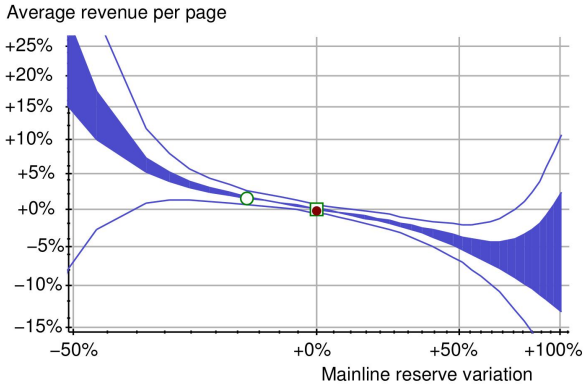
An experiment: mainline reserve

- Experiment: scale the mainline reserve according to some multiplier
  - Where ρ, σ are hyperparameters
- Collect data using ρ = 1, σ = 0.3. (i.e. generate ads, let users search, record click yield)
- Use this to estimate what the click yield would have been given a different ρ*, σ*

An experiment: mainline reserve







Estimated variations of three performance metrics in response to mainline reserve changes. The curves delimit 95% confidence intervals for the metrics we would have observed if we had increased the mainline reserves by the percentage shown on the horizontal axis. The filled areas represent the inner confidence intervals. The hollow squares represent the metrics measured on the experimental data. The hollow circles represent metrics measured on a second experimental bucket with mainline reserves reduced by 18%. The filled circles represent the metrics effectively measured on a control bucket running without randomization.

Other things we could do

- This experiment kept $\sigma^* = \sigma$, but we could change it to see what would happen if the reserve fluctuated more widely
  - It would also be interesting to ask the question “What would click yield be if we had shown some people more mainline ads, other people fewer”
- We could try to estimate an exact value of the mainline reserve without randomising
- We could add more dimensions along which to experiment – not just changing the score function
  - The difficulty here is that we’d have to effectively collect more data exploring multiple dimensions

Next…

- The next section in the paper shows ways to use the causal graph that our structural equation model induces to improve this counterfactual analysis
- Better reweighting variables
- Better confidence intervals using invariant predictors

- “Learning” section explores how to fit a model to the counterfactual distribution to predict a variable of interest

Conclusion

- Using causal inference techniques enables you to reason counterfactually – about things that haven't happened
- We can apply this in an advertising context to find good ad layouts that maximise click yield
- In theory we could approximate a counterfactual estimate of the click yield simply by sampling from our existing distribution and reweighting the samples
- In practice, importance sampling has a key limitation: the two distributions must overlap somewhat, otherwise our variance blows up
- Clipping the weights fixes this, and enables you to get an estimate + confidence intervals on the estimate

Extra slides

The first term of this decomposition is the *clipped expectation* $\bar{Y}^*$. Estimating the clipped expectation $\bar{Y}^*$ is much easier than estimating Y^* from (7) because the clipped weights $\bar{w}(\omega)$ are bounded by R.

$$\bar{Y}^* = \int_{\omega \in \Omega_R} \ell(\omega)\, \mathrm{P}^*(\omega) = \int_{\omega} \ell(\omega)\, \bar{w}(\omega)\, \mathrm{P}(\omega) \approx \widehat{Y}^* = \frac{1}{n} \sum_{i=1}^{n} \ell(\omega_i)\, \bar{w}(\omega_i). \tag{10}$$

The second term of Equation (9) can be bounded by leveraging assumption (8). The resulting bound can then be conveniently estimated using only the clipped weights.

$$Y^* - \bar{Y}^* = \int_{\omega \in \Omega \setminus \Omega_R} \ell(\omega)\, \mathrm{P}^*(\omega) \in \left[0,\, M\, \mathrm{P}^*(\Omega \setminus \Omega_R)\right] = \left[0,\, M\,(1 - \bar{W}^*)\right] \quad \text{with}$$

$$\bar{W}^* = \mathrm{P}^*(\Omega_R) = \int_{\omega \in \Omega_R} \mathrm{P}^*(\omega) = \int_{\omega} \bar{w}(\omega)\, \mathrm{P}(\omega) \approx \widehat{W}^* = \frac{1}{n} \sum_{i=1}^{n} \bar{w}(\omega_i). \tag{11}$$

Since the clipped weights are bounded, the estimation errors associated with (10) and (11) are well characterized using either the central limit theorem or using empirical Bernstein bounds (see appendix B for details). Therefore we can derive an *outer confidence interval* of the form

$$\mathbb{P}\left\{ \widehat{Y}^* - \varepsilon_R \leq \bar{Y}^* \leq \widehat{Y}^* + \varepsilon_R \right\} \geq 1 - \delta \tag{12}$$

and an *inner confidence interval* of the form

$$\mathbb{P}\left\{ \bar{Y}^* \leq Y^* \leq \bar{Y}^* + M(1 - \widehat{W}^* + \xi_R) \right\} \geq 1 - \delta. \tag{13}$$

The names *inner* and *outer* are in fact related to our preferred way to visualize these intervals (e.g., Figure 13). Since the bounds on $Y^* - \bar{Y}^*$ can be written as

$$\bar{Y}^* \leq Y^* \leq \bar{Y}^* + M\,(1 - \bar{W}^*), \tag{14}$$

we can derive our final confidence interval,

$$\mathbb{P}\left\{ \widehat{Y}^* - \varepsilon_R \leq Y^* \leq \widehat{Y}^* + M(1 - \widehat{W}^* + \xi_R) + \varepsilon_R \right\} \geq 1 - 2\delta. \tag{15}$$