Black-Box Decision Making Systems

Joshi et al. (2019)

Patrick Altmeyer

19 April, 2021

Introduction

Motivation

- Joshi et al. (2019) argue that individuals that are subject to some automated decision making system should be able to improve their outcomes (and know how to!)
- Why? Algorithms are rarely held accountable for their decisions - you cannot appeal to them.
- Blindly relying on algorithms has detrimental consequences for individuals (O'neil 2016)
  - Example of fifth-grade teacher Sarah Wysocki, who had received excellent reviews from peers, supervisors and students, but was fired after a novel teacher evaluation algorithm had rendered her redundant.

"The human victims of WMDs [. . .] are held to a far higher standard of evidence than the algorithms themselves" (O'neil 2016)

Contributions

- *Individual recourse*: given an unfavourable outcome of decision-making system, can an individual take actions in order to improve the outcome?
- Joshi et al. (2019) propose an algorithm that returns the smallest set of changes δ that will lead to a label switch.
- Three key contributions largely extending the work of Ustun, Spangher, and Liu (2019):

1. Framework avoids suggesting unrealistic set of changes by imposing threshold likelihood on sample distribution $p(X)$.
   - Ustun, Spangher, and Liu (2019) only avoids immutable variables like age, sex, gender.
2. It is applicable to broader class of models (black-box classification and causal)
   - Approach proposed by Ustun, Spangher, and Liu (2019) is restricted to linear classifiers.
3. It can be used to detect poorly defined proxies and biases.

Methodology

Optimization - high-level context

Let $y \in \{-1, 1\}$ a binary outcome variable and $X \in \mathbb{R}^d$ a feature matrix containing individuals' attributes. Suppose $y^* = -1$ - the negative outcome - for some individual characterized by attributes X^*. Then we want to find X' closest to X^* such that the classifier assigns the positive outcome (primary constraint) and the likelihood of the sample $p(X')$ exceeds some threshold γ (secondary constraint).

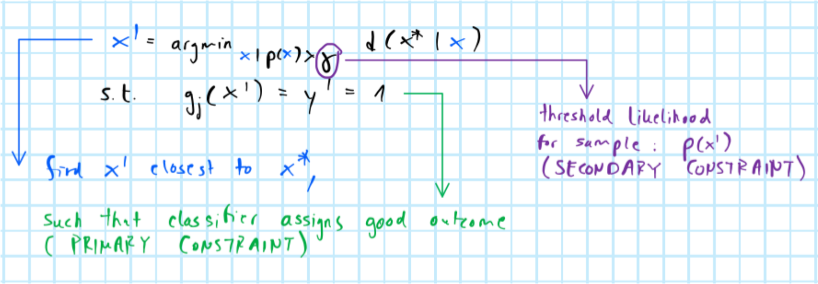


Figure 1: Optimization - high-level context.

Optimization in latent space

- The high-level optimization involves the data distribution $p(X)$ which needs to be characterized.
  - Joshi et al. (2019) use two approaches - VAE and GAN - to generate distribution $p(Z)$ in latent space.
- Then optimization is run in latent space.
  - Note: the primary constraint is now captured by a regularization hyperparameter.

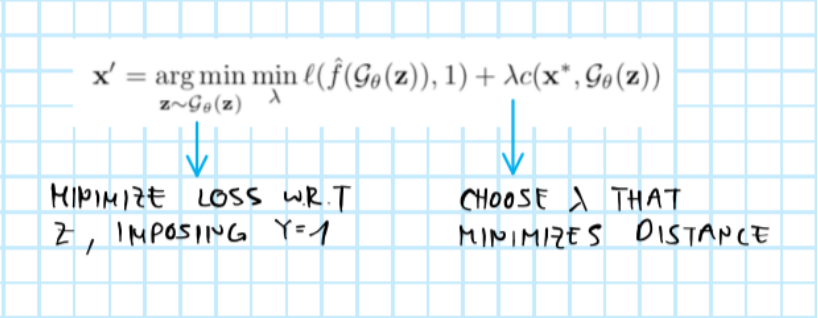


Figure 2: Optimization in latent space.

REVISE Algorithm

1. Register X in latent space: $Z \leftarrow \mathcal{F}(X)$.
2. Use generative model to generate distribution of latent variable Z. Note: can now impose threshold likelihood on sample through $P(Z)$.
3. Optimize in latent space and obtain solution Z'.
4. Decode: $X' \leftarrow \mathcal{G}(Z')$.

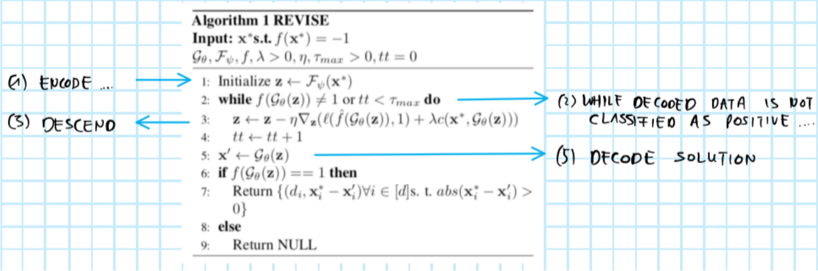


Figure 3: The final algorithm.

Structural causal models - and their pitfalls

- As before, let $y \in \{-1, 1\}$ a binary outcome variable. Let $t \in \{0, 1\}$ denote a treatment indicator.
- Structural causal models (SCM) are generally concerned with estimating the average treatment affect of t:

$$\alpha_{ATE} = \mathbb{E}\left[Y_{1i} - Y_{0i}\right]$$

- Estimation hinges on the assumption that treatment is randomly assigned: $(Y_{1i}, Y_{0i}) \perp\!\!\!\perp t$
  - Otherwise estimation is subject to selection bias.

Pitfalls

1. In the presence of confounders X we may establish conditional independence $(Y_{1i}, Y_{0i}) \perp\!\!\!\perp t|X$ but only of X is observed.
2. Estimated effect is still an average: we never observe both Y_{1i} and Y_{0i}. Individual outcome may still be unfavourable even if $t = 1$ and $\alpha_{ATE} > 0$.

Individual recourse for causal outcomes (1)

- Joshi et al. (2019) draw a parallel between the latent variables Z learned from the generative model and hidden confounders in an SCM (pitfall 1):

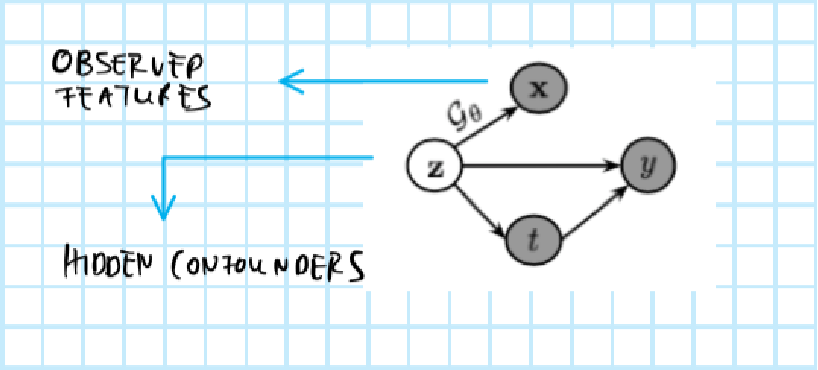


Figure 4: From the latent data manifold to hidden confounders.

Individual recourse for causal outcomes (2)

- For individuals with unfavourable post-intervention outcome - when treatment is known to have positive average treatment effect - provide individual recourse with respect to hidden confounders (pitfall 2):

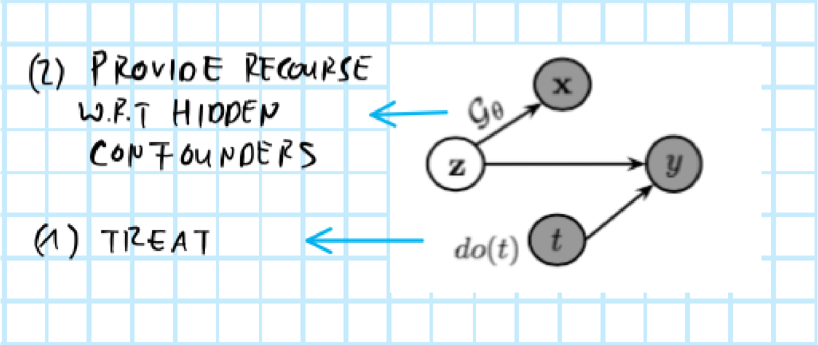


Figure 5: After intervention, provide recourse with respect to hidden confounders.

Main findings

Classification models

- Train both a linear (white-box) and a non-linear (black-box) classifier to predict if an individual can be expected default in the next month or not.
- Ustun, Spangher, and Liu (2019) recommends a large (possibly infeasible) change in "Most Recent Payment Amount."
- REVISE (MLP) (Joshi et al. 2019) recommends small changes to "Max Payment Amount Over Last 6 Months" and "Most Recent Bill Amount" and contrary to the other approaches suggests that the former should be smaller than the latter (contradiction).

Attribute	original	REVISE (Linear)	REVISE (MLP)	Ustun et. al. '18 (Linear)
Max Bill Amount Over Last 6 Months	2240.0	3461.2947	1548.9572	-
Max Payment Amount Over Last 6 Months	110.0	100.3251	17.0988	-
Months With High Spending Over Last 6 Months	6.0	0.0547	1.9147	-
Most Recent Bill Amount	2050.0	1768.1843	2059.7888	-
Most Recent Payment Amount	80.0	28.2974	0.0	6010.0
Total Overdue Counts	1.0	1.7552	0.5058	-
Total Months Overdue	12.0	1.05	0.4	-
Others (Marital Status)	0.0	-	-	1

Contradiction

Unrealistic

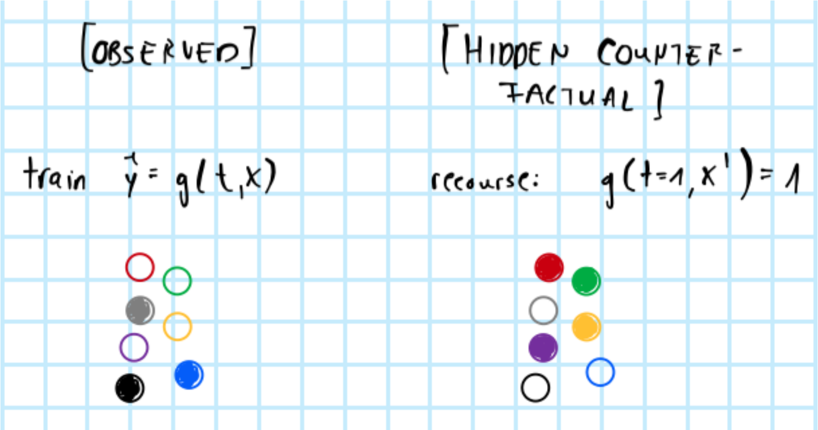
Figure 6: Results

Causal models - Louizos et al. (2017)

- As mentioned above, in practice we only observe either Y_{1i} (outcome under treatment) or Y_{0i} (outcome without treatment).
- Louizos et al. (2017) introduce a benchmark task using data from twin births and treating pair of twins as individual i:
  - Y_{1i} : mortality of heavier twin $t = 1$
  - Y_{0i} : mortality of lighter twin $t = 0$
- Allows them to estimate "true" $\alpha_{ATE} = -2.5$ and use as benchmark

Causal models - applied to individual recourse

- Joshi et al. (2019) use this set up to (1) train classifier g on "observed" data X and (2) provide individual recourse to individuals in the hidden counterfactual $\tilde{X}$:
  - case of interest: what if individual i (heavier twin) had received treatment, but the outcome was still $y = -1$ (child death)?



Causal models - results

- Bottom line is that REVISE recommends a reduction of risk factors.
- Confounding changes the qualitative results - caution!

feature name	original	recourse (no confounding)	recourse (confounding)
risk factor Hvdramnios (0=no risk)	1.0	0.0	0.0
risk factor, Incompetent cervix (0=no risk)	1.0	0.0	0.0
total number of births before twins	8.0	-	1.0
Other Medical Risk Factors (0=no risk)	1.0	0.0	0.0
risk factor, Diabetes (0=no risk)	0.0	1.0	-

Figure 7: Results for causal model.

Attribute confounding

- Gender classification from face images through two deep neural networks, f_1 (weakly biased) and f_2 (strongly biased).
  - biased in the sense that in the subset fed to f_2, all black-haired samples are male and all bold samples are female
- Use their REVISE algorithm to change the hair-colour attribute: label switching observed for f_2.

Figure 8: Biased classifier is sensitive to its inherent bias: gender labels are switched as REVISE provides individual recourse.

Critical review

Caveats

- People are actually fairly good at finding the “smallest set of actions” to game the system:
  - The case of Sarah Wysocki: teachers realized that their own faith depended on their students’ exam grades and evidence suggested that many took action by artificially inflating their students’ test scores (cheating).
- If the decision making system uses poorly defined rules and proxies, then individual recourse may still lead to undesirable outcomes for society.
  - As we have seen, Joshi et al. (2019) pick up on this issue: there algorithm can be used to identify poorly defined, biased proxies.
- Individual recourse for causal outcome hinges on the assumption that hidden confounders Z can be estimated from observed confounders X - this may not always hold.

Future avenues for research - Theory

- Algorithm does not provide an order for recourse with respect to individual attributes $X_1, ..., X_d$ - could this be easily extended? Why not just use individual distances as a natural ranking?
- Individual recourse for causal outcome when treatment is non-binary: regression discontinuity $\lim_{z \to z_0^+} P(t_i = 1 | Z_i = z_0) \neq \lim_{z \to z_0^-} P(t_i = 1 | Z_i = z_0)$
- To simplify the optimization, how about just assuming a prior for $P(X)$? May be useful in settings where d is relatively small and one can reasonably assume $X_i \sim \mathcal{N}(\mu, \sigma)$, for example.
- Extend to regression case (continuous outcome): instead of imposing $g(X') = 1$ could we impose something like $g(X') = y^*$ where y^* is some target level? Or simply encode $\tilde{y} = 1$ if $y > y^*$ and $\tilde{y} = -1$ otherwise.

Future avenues for research - Applications

- Use REVISE to identify potential biases in decision support systems, for example: credit approval, candidate hire, etc. Similarly, REVISE can be used to provide individual recourse to customers.
- Public policy: let $y_i = f(X_i)$ be a model for CO^2 emissions y in region i and X_i is a set of observables. Suppose $y_i > y^*$ where y^* is some target level. Use REVISE to provide individual recourse for region i to reduce emissions.

References

Joshi, Shalmali, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. 2019. "Towards Realistic Individual Recourse and Actionable Explanations in Black-Box Decision Making Systems." *arXiv Preprint arXiv:1907.09615*.

Louizos, Christos, Uri Shalit, Joris Mooij, David Sontag, Richard Zemel, and Max Welling. 2017. "Causal Effect Inference with Deep Latent-Variable Models." *arXiv Preprint arXiv:1705.08821*.

O'neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

Ustun, Berk, Alexander Spangher, and Yang Liu. 2019. "Actionable Recourse in Linear Classification." In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 10–19.

Annex

Causal models - immutable variables

- Let $X_{\mathbf{I}}$ denote the set of immutable variables and $X_{\mathbf{M}} = X \backslash X_{\mathbf{I}}$ the remaining mutable variables.
- Joshi et al. (2019) propose a modified *conditional* causal decision making system:

Figure 9: Causal graphs with immutable attributes.