Counterfactual Explanations and Algorithmic Recourse

Patrick Altmeyer (p.altmeyer@tudelft.nl)
Dr. Cynthia Liem
TU Delft — Delft University of Technology

Explaining black box models through counterfactuals

All too often human operators rely blindly on decisions made by black-box algorithms. **Counterfactual Explanations (CE)** can help programmers make sense of the systems they build: they **explain how inputs into a system need to change for it to produce a different output**. CEs that involve realistic and actionable changes can be used for the purpose of individual recourse: **Algorithmic Recourse (AR) offers individuals subject to algorithms a way to turn a negative decision into positive one.**

Our work so far:

- ✓ Built a scalable library in Julia: CounterfactualExplanations.jl (to be submitted to upcoming JuliaCon 2022).
- ✓ Have run experiments investigating the dynamics of AR: individuals who received recourse form a distinct cluster in target class leaving them potentially vulnerable to discrimination through the system.
- ✓ Proposed a related research project to bachelor's students and aim to submit work-in-progress to AIES '22 student track.

Where to go from here

Open questions (your thoughts are more than welcome!)

- ❑ How much of an issue is this really? Can we think of real-world examples where scope for discrimination may lead to undesirable outcomes?
- ❑ How does the magnitude of domain and model shifts vary across different approaches to generating AR? (student project)
- ❑ Can we assess what factors mitigate endogenous shifts when generating recourse?

From basic principles …

A light-hearted motivating example

Suppose we have fitted some black box classifier to divide cats and dogs based on two features: height and tail length. One individual cat – let's call her Kitty 🐱 – is friends with a lot of cool dogs and wants to remain part of that group. The counterfactual path in Figure 1 shows how 🐱 needs to change her appearance in order to be allocated to the group of dogs by the system.

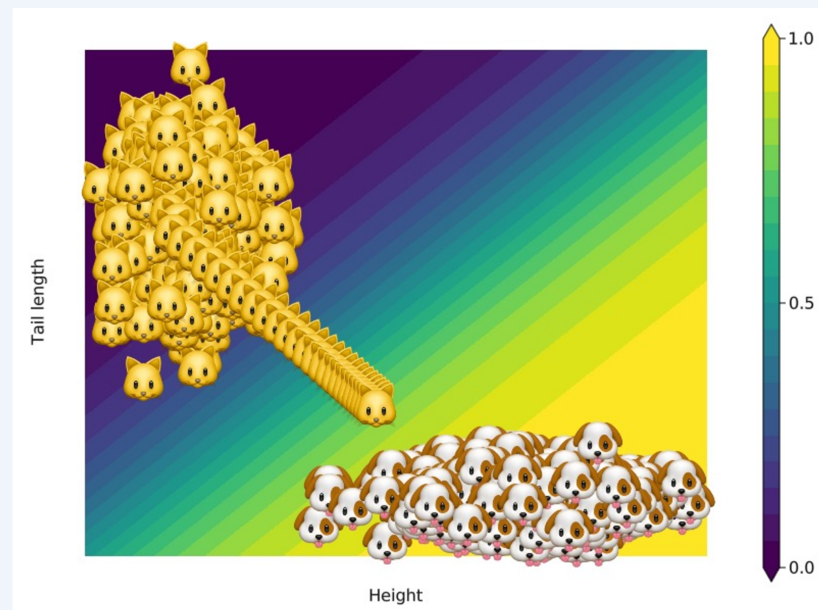

Figure 1: Generating recourse for 🐱 following Wachter et al. (2018)[1]. Contour shows the predictions of a simple MLP.

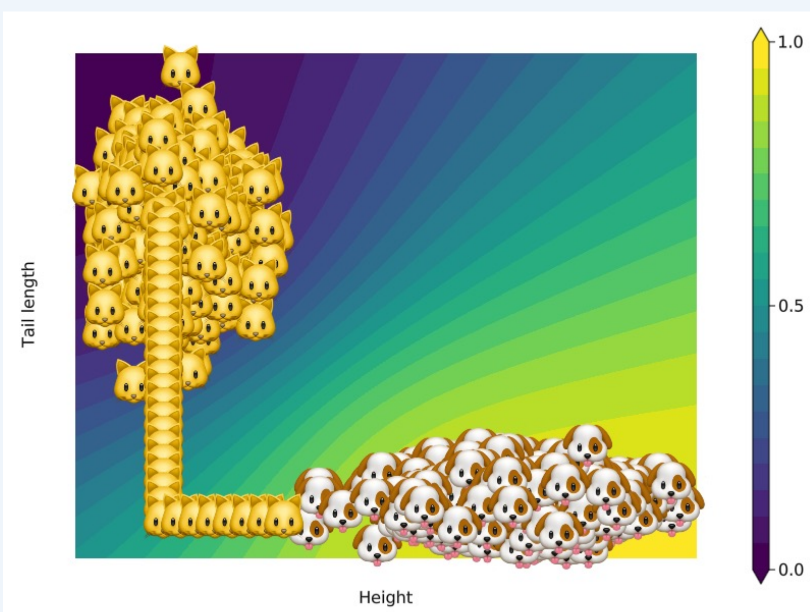

Figure 2: Generating recourse for 🐱 following Schut et al. (2021)[2]. Contour shows the predictions of a deep ensemble.

… to realistic recourse.

CE by implicitly minimizing predictive uncertainty

As 🐱 crosses the decision boundary in Figure 1 she fools the system, but we can still clearly distinguish her from the rest of her dog friends. Her counterfactual self is **ambiguous** and **unrealistic**. Consider instead the counterfactual path generated in Figure 2 which uses a Bayesian approach: for the same confidence threshold 🐱 ends up in a much denser area.

Applied to MNIST data the Bayesian approach arguably generates the most realistic counterfactuals, albeit with mixed success (Figures 3-6).

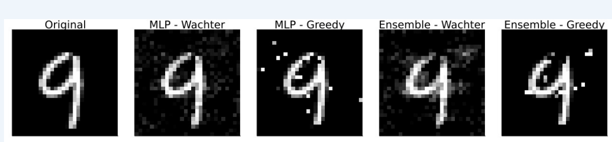

Figure 3: Counterfactual explanations for MNIST data – turning a 9 into a 4.

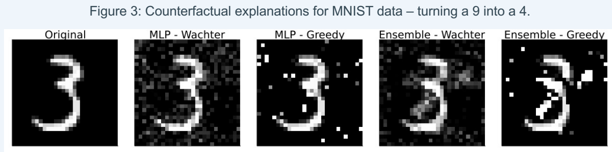

Figure 4: Counterfactual explanations for MNIST data – turning a 3 into an 8.

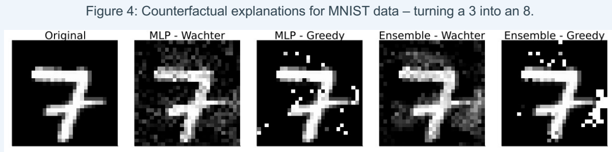

Figure 5: Counterfactual explanations for MNIST data – turning a 7 into a 2.

Figure 6: Counterfactual explanations for MNIST data – turning a 1 into a 7.

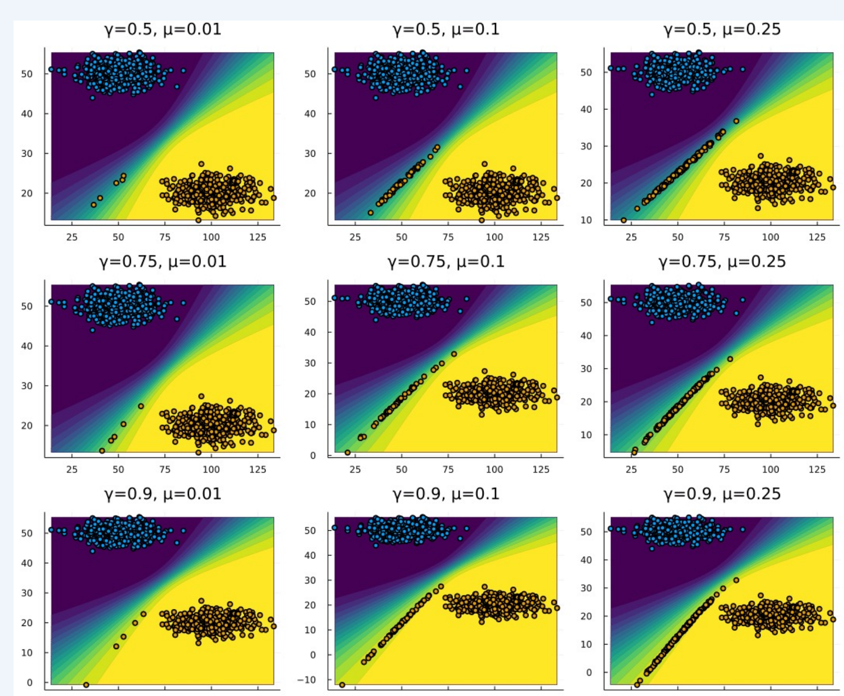

Figure 7: Algorithmic recourse leads to domain and model shifts.

But wait a minute …

Beyond the static setting

In practice decision-making systems are regularly updated. Recent work has investigated the robustness of AR[3]: can we be sure that 🐱 can stay with her dog friends after model updates? In our work we go a step further and ask ourselves:

- ❑ Does 🐱 herself trigger model shifts through her move across the decision boundary?
- ❑ Does that have consequences for other cats or dogs that want to implement recourse?
- ❑ More generally, **what are the dynamics of algorithmic recourse?**

Preliminary experiments show:

- ➢ Individuals like 🐱 form a distinct cluster in the target class leaving them potentially vulnerable to further discrimination through the system (Figure 7).

References

[1] Wachter et al. (2018). "Counterfactual explanations without opening the black box: automated decisions and the GDPR.". In: Harvard Journal of Law & Technology (31)

[2] Schut et al. (2021). "Generating interpretable counterfactual explanations by implicit minimisation of epistemic and aleoteric uncertainty.". In: Proceedings of Machine Learning Research (130)

[3]. Upadhyay et al. (2021). "Towards Robust and Reliable Algorithmic Recourse.". In: Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021).