Explainable AI: Probabilistic Methods for Counterfactual Explanations

Patrick Altmeyer (p.altmeyer@tudelft.nl)
Dr. Cynthia Liem (c.c.s.liem@tudelft.nl)

Abstract

Counterfactual Explanations (CE) are a promising approach to explainable artificial intelligence (XAI). They explain how inputs into a model need to change for it to produce different outputs. To ensure that the generated explanations are realistic it is important to understand which input-output pairs are likely to occur. Using probabilistic methods this can be done in one of two ways:

1. Use a generative model to learn the data manifold.
2. Restrict the class of classifiers to Bayesian classifiers.

I present both approaches here and introduce a framework for generating counterfactuals in Julia. In ongoing work I am investigating the dynamics of counterfactual explanations.

Contributions

- ✓ CounterfactualExplanations.jl – a Julia package for generating counterfactual explanations. To be presented as main talk at JuliaCon 2022 and published in proceedings.
- ✓ LaplaceRedux.jl – a Julia package for Bayesian deep learning through Laplace Approximation. To be presented as lightening talk at JuliaCon 2022.
- ✓ Algorithmic Recourse Dynamics: Algorithmic Recourse (AR) relates to the process of providing individuals with actionable and realistic counterfactual explanations. Together with a group of students I have been investigating the dynamics of AR.

Ongoing work

- ❑ CounterfactualExplanations.jl is under active development. I have recently added support for latent space search (through the data manifold) as well as native support for models trained in Python and R.
- ❑ LaplaceRedux.jl still has very limited functionality. I want to improve it and use for my work on CE in the Bayesian context.
- ❑ The work on AR dynamics will be extended to the context of Bayesian classifiers. Preliminary evidence suggests that probabilistic methods help to mitigate undesirable endogenous dynamics to some extent.

BACKGROUND

From 🐱 to 🐶: Suppose we have trained a black-box classifier to discriminate cats from dogs (**Figure 1**). To understand why some individual cat was not classified as a dog, we can move her from her factual state 🐱 to a counterfactual state 🐶. Why has the cat not been classified as a dog? Because she is too short and her tail is too long.

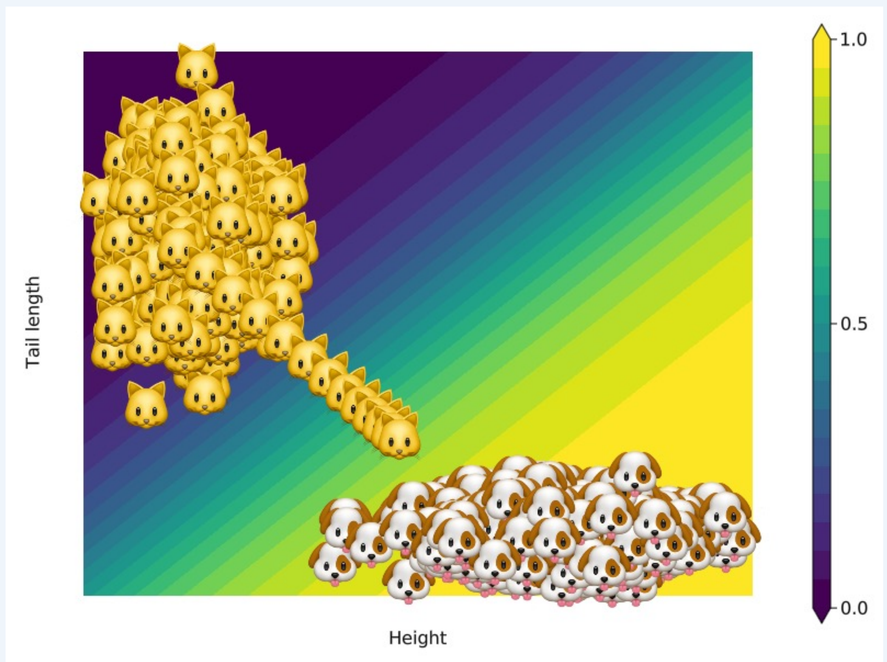

Figure 1: Generating a counterfactual for 🐱 following Wachter et al. (2018)[1]. The contour shows the predictions of a simple multi-layer perceptron (MLP).

COMPETING PROBABILISTIC APPROACHES

1) Leveraging Predictive Uncertainty

Schut et al. (2021) have noted that simply minimizing predictive uncertainty of well-specified Bayesian classifiers typically yields satisfactory counterfactuals: minimal predictive uncertainty corresponds to minimal epistemic uncertainty (realistic counterfactual) and minimal aleatoric uncertainty (unambiguous counterfactuals).

✅	❌
• Smaller engineering overhead. • Fast and greedy search. • Counterfactual reflects quality of classifier, not some generative model.	• Restricted to Bayesian models.

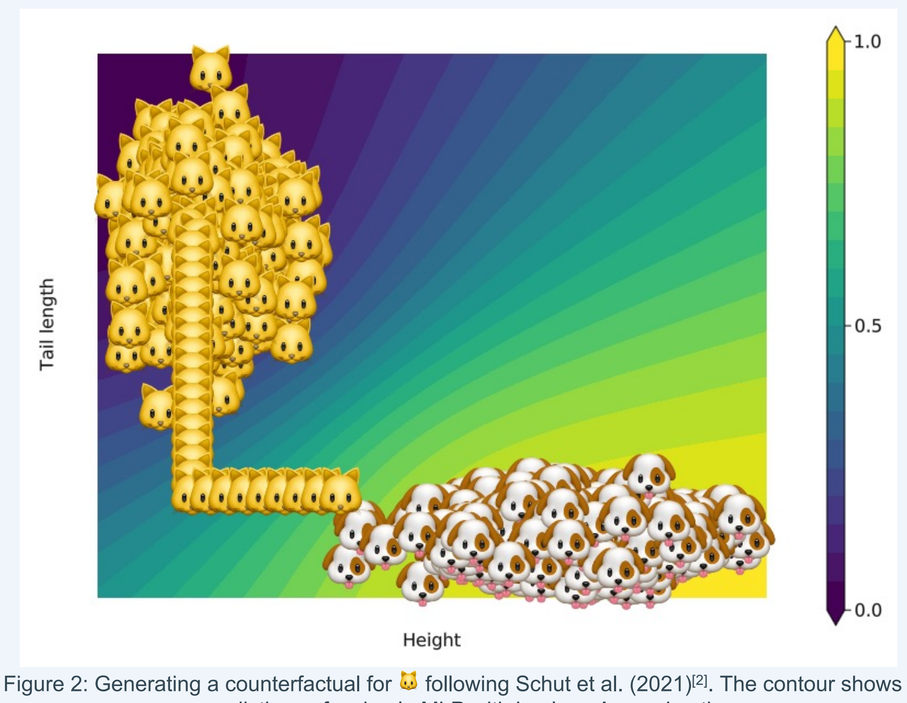

Figure 2: Generating a counterfactual for 🐱 following Schut et al. (2021)[2]. The contour shows the predictions of a simple MLP with Laplace Approximation.

Application to MNIST: results for deep ensemble are more realistic (**Figure 3**).

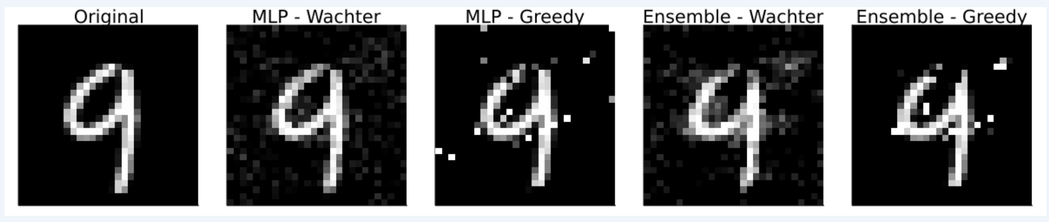

Figure 3: Counterfactual explanations for MNIST data. Turning a nine (9) into a four (4).

2) Traversing Latent Embeddings

Instead of perturbing samples directly, some have proposed to instead traverse a lower-dimensional latent embedding learned through a generative model (**Figure 4**). This helps to produce realistic counterfactuals, because the joint likelihood of input-output pairs is implicitly encoded in the latent embedding.

✅	❌
• Fast search in lower-dimensional latent space.	• Engineering overhead. • Quality of counterfactuals may still be poor if classifier is highly decisive (?).

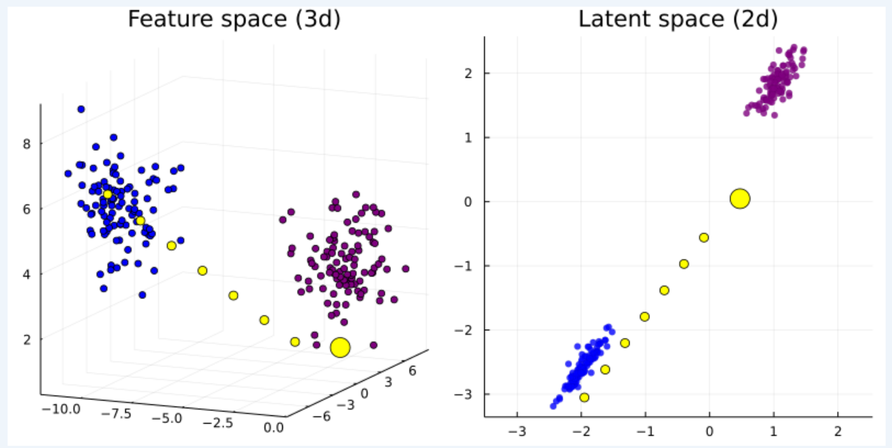

Figure 4: Counterfactual (yellow) generated through latent space search (right panel) following Joshi et al. (2019)[3]. The corresponding counterfactual path in the feature space is shown in the left panel.

Application to MNIST: Provided the generative model is expressive enough, search in the latent space can yield very realistic counterfactuals (**Figure 5**). But things can go wrong: the counterfactual highlighted in red (**Figure 6**) was produced using a less expressive VAE. It is classified as a seven (7) by the classifier despite looking like a nine (9). Conversely, the counterfactual highlighted in blue is classified as a nine (9) despite looking like a seven (7). This is likely due to a combination of an overspecified classifier and an underspecified VAE.

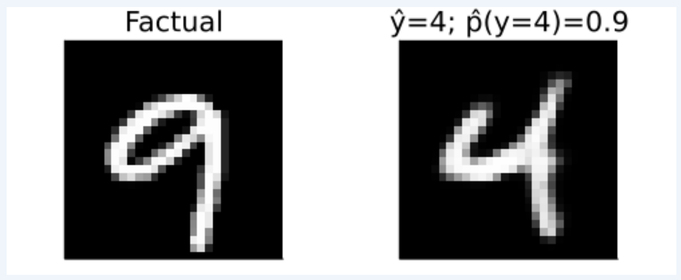

Figure 5: Turning a nine (9) into a four (4) using generic search in the feature space. It appears that the VAE is well-specified in this case.

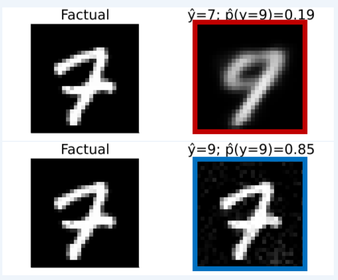

Figure 6: Turning a seven (7) into a nine (9) using generic search in the latent space (red) and feature space (blue).

ALGORITHMIC RECOURSE DYNAMICS

In ongoing work, I am comparing different counterfactual generators in a dynamic setting (**Figure 7**): the implementation of AR for a subset of individuals leads to a domain shift (b), which in turn triggers a model shift (c). As this is repeated, the decision boundary moves towards the negative class (d). Such dynamics are undesirable: in the context of loan applications, for example, one ends up with a group of borrowers that has potentially much higher average default risk.

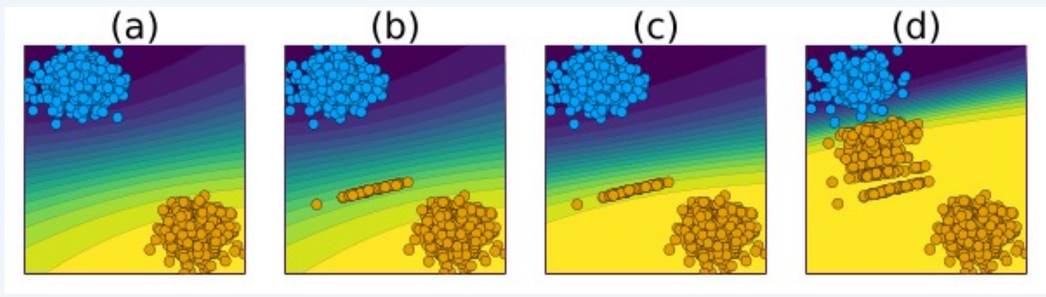

Figure 7: Repeated implementation of AR and subsequent classifier updates lead to undesirable endogenous dynamics.

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 aleatoric uncertainty.”. In: Proceedings of Machine Learning Research (130)

[3] Joshi et al. (2013). “Towards realistic individual recourse and actionable explanations in black-box decision making systems.”. In: arXiv preprint arXiv:1907.09615.