Computing Counterfactuals

A brief history of methods

Philip Naumann

October 21, 2020

Table of Contents

1. Wachter et al., 2017
2. Tolomei et al., 2017
3. Mothilal et al., 2020
4. Dandl et al., 2020
5. Wrap Up

Preliminaries

The following definitions hold for the remaining presentation:

- x_0: denotes the **initial input instance**
- x': denotes a **counterfactual**
- x^*: denotes the **optimal counterfactual**
- y': denotes the **target class**
- $\hat{\mathrm{f}}(\cdot)$: denotes the prediction of a **trained model**
- $\mathrm{d}(\cdot, \cdot)$ or $\delta(\cdot, \cdot)$: denotes a **distance** or **cost** function

Table of Contents

1. Wachter et al., 2017
2. Tolomei et al., 2017
3. Mothilal et al., 2020
4. Dandl et al., 2020
5. Wrap Up

Original Paper

Sandra Wachter et al. (2017). "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR". In: *SSRN Journal*

Loss function

$$L(x_0, x', y', \lambda) = \lambda \cdot (\hat{f}(x') - y')^2 + d(x_0, x') \tag{1}$$

- Balance weight
- **Prediction loss**
- **Distance function** (i.e. smallest tweaking)

Optimization problem

Solve

$$x^* = \arg\min_{x'} \max_{\lambda} L(x_0, x', y', \lambda) \tag{2}$$

with e.g. ADAM, Nelder-Mead, an EA ...

Algorithm

From: Christoph Molnar (2019). *Interpretable Machine Learning - A Guide for Making Black Box Models Explainable*.

1. Select an instance x_0 to be explained, the desired outcome y', a tolerance ε and a (low) initial value for λ
2. Sample a random instance as initial counterfactual x'_0
3. Optimize the loss L with the initially sampled counterfactual as starting point
4. While $|\hat{f}(x') - y'| > \varepsilon$:
    1. Increase λ
    2. Optimize the loss L with the current best counterfactual x'_{t-1} as starting point
    3. Return the counterfactual that minimizes the loss x'_t
5. Repeat steps 2-4 and return the list of counterfactuals (i.e. choose a new x'_0) or the one that minimizes the loss

Advantages & Disadvantages

Advantages:

- Can find multiple CFs by restarting with a different seed position
- Simple objective and efficiently computed if there is access to model gradients

Disadvantages:

- Doesn't include plausibility
- No support for further constraints
- When computing multiple CFs, the additional information per new CF is not maximized with respect to the already found ones

Table of Contents

1. Wachter et al., 2017
2. Tolomei et al., 2017
3. Mothilal et al., 2020
4. Dandl et al., 2020
5. Wrap Up

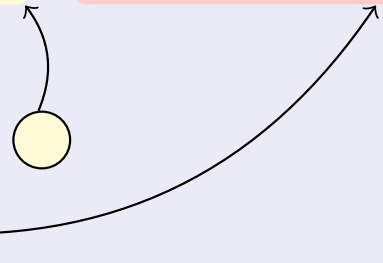
Actionable Feature Tweaking

Gabriele Tolomei et al. (2017). "Interpretable Predictions of Tree-Based Ensembles via Actionable Feature Tweaking". In: KDD '17

Optimization problem

Required: *access* to a trained **Random Forest** $f(\cdot)$

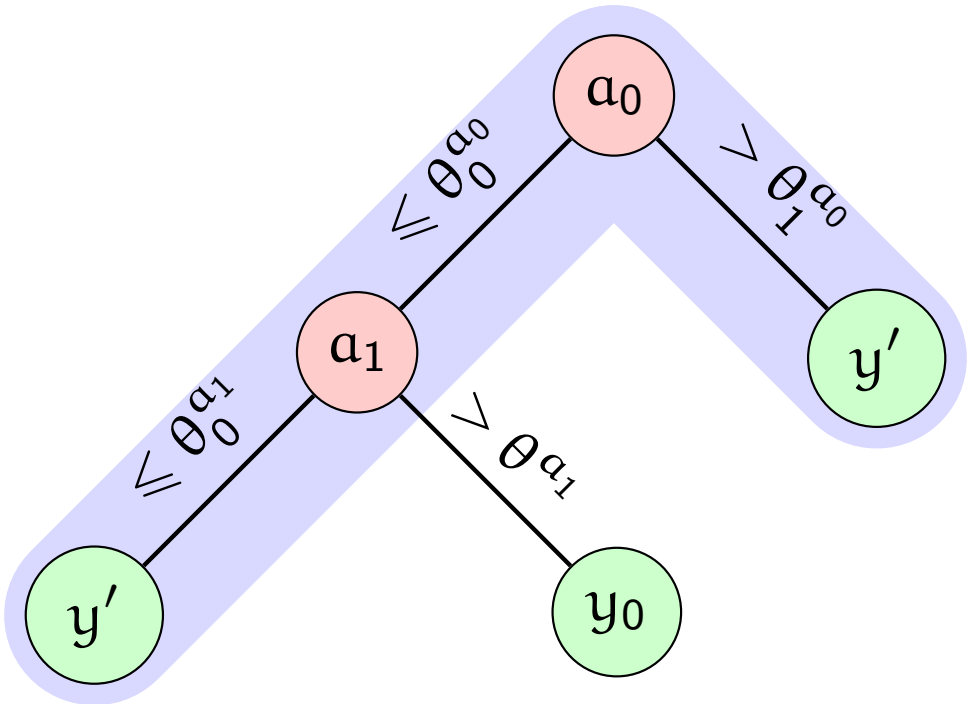
$$x^* = \arg\min_{x'} \{ \delta(x_0, x') \mid \hat{f}(x_0) = -1 \wedge \hat{f}(x') = +1 \} \tag{3}$$



- Cost function (i.e. distance)
- Prediction **constraint**

How to compute it?

1. Find all paths in the forest f that lead to $y' = +1$
2. Each feature threshold θ in the path is then *tweaked* by a fixed ε, thus creating (multiple) x'
3. From this set of CFs, choose the x' that minimizes $\delta(\cdot, \cdot)$ and still validates $\hat{f}(x') = +1$ after the tweaking (!)



$$x'_0 = [\theta_0^{a_0} - \varepsilon,\ \theta_0^{a_1} - \varepsilon]$$
$$\delta(x_0, x'_0) = 1.5$$

$$x'_1 = [\theta_1^{a_0} + \varepsilon]$$
$$\delta(x_0, x'_1) = 0.5$$

$$x^* = x'_1$$

Advantages & Disadvantages

Advantages:

- Counterfactuals are based on the embedded ground truth data
- Deterministic and exact (always produces the same CF)

Disadvantages:

- Only works for Random Forests
- Mostly independent of x_0 as the tweaking is only based on the thresholds
- Assumes that each feature has at most one associated threshold value per path (otherwise tweaks may get "overwritten")
- Can only compute exactly one counterfactual
- ε is static and the same for each feature
- No way to (easily) add further constraints

Table of Contents

1. Wachter et al., 2017
2. Tolomei et al., 2017
3. Mothilal et al., 2020
4. Dandl et al., 2020
5. Wrap Up

Towards Diversity and Actionability

Ramaravind K. Mothilal et al. (2020). "Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations". In: FAT* '20

Optimization problem

- Weights

$$\underset{x'_1,\ldots,x'_k}{\arg\min} \frac{1}{k}\sum_{i=1}^{k} \text{loss}(\hat{f}(x'_i), y') + \frac{\lambda_1}{k}\sum_{i=1}^{k} d(x_0, x'_i) - \lambda_2 \, \text{dpp}(x'_1, \ldots, x'_k) \tag{4}$$

- Prediction loss
- Distance function (i.e. distance)
- **Diversity** objective (*Determinental Point Process*)
- Furthermore, they allow to set **range and value constraints**

Determinental Point Process (DPP)

Mothilal et al., 2020

Use a DPP to determine the diversity within the set of counterfatuals:

DPP

$$\mathbf{K}_{i,j} = \frac{1}{1 + d(x'_i, x'_j)} \tag{5}$$

$$dpp(x'_1, \ldots, x'_k) = det(\mathbf{K})$$

- $\mathbf{K}$ is an affinity matrix between all k counterfactuals
- Measures how different the set of counterfactuals are from another (based on feature distances)

Advantages & Disadvantages

Advantages:

- Able to find a *diverse* set of counterfactuals
- Allows to set further constraints to improve the *actionability*

Disadvantages:

- Only works for neural networks as the optimization relies on their gradients
- Can only find a fixed number of CFs (even though there might be more)
- *Settings explicit weights between objectives*

Table of Contents

1 Wachter et al., 2017

2 Tolomei et al., 2017

3 Mothilal et al., 2020

4 Dandl et al., 2020

5 Wrap Up

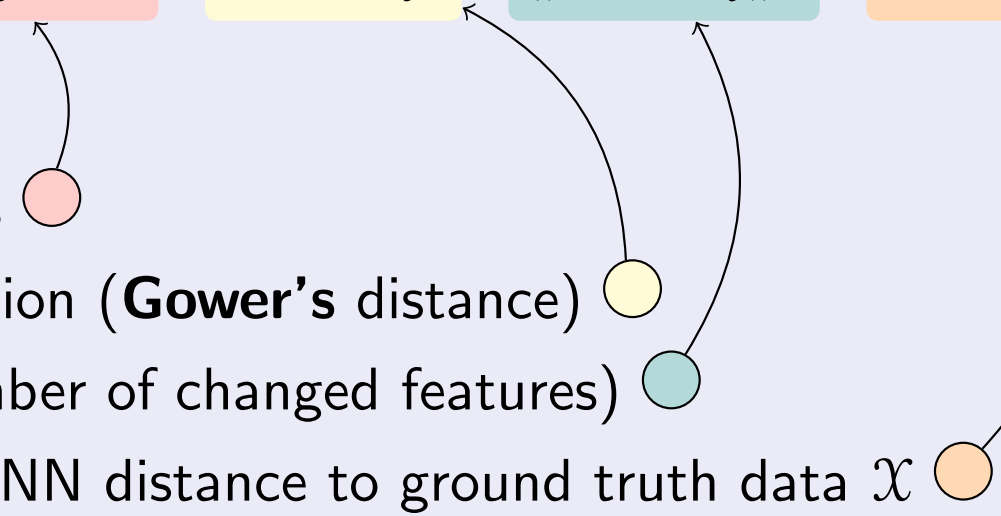
A Multi-objective Perspective

Susanne Dandl et al. (2020). "Multi-Objective Counterfactual Explanations". In: Lecture Notes in Computer Science

- Find a *set* of CFs that fulfill different aspects of the objective space
- No need for explicitly setting objective weights

Optimization problem

$$\underset{x_i'}{\arg\min}\Big[\ \text{loss}(\hat{f}(x_i'), y')\ ,\ \delta(x_0, x_i')\ ,\ \|x_0 - x_i'\|_0\ ,\ \text{kNN-d}(x_i', \mathcal{X})\ \Big] \quad (6)$$



- Prediction loss
- Distance function (**Gower's** distance)
- **Sparsity** (number of changed features)
- **Plausibility**, kNN distance to ground truth data $\mathcal{X}$

NSGA-II

- Popular multi-objective EA using the **pareto** principle
- In particular, **non-dominated sorting** and **crowding distance**

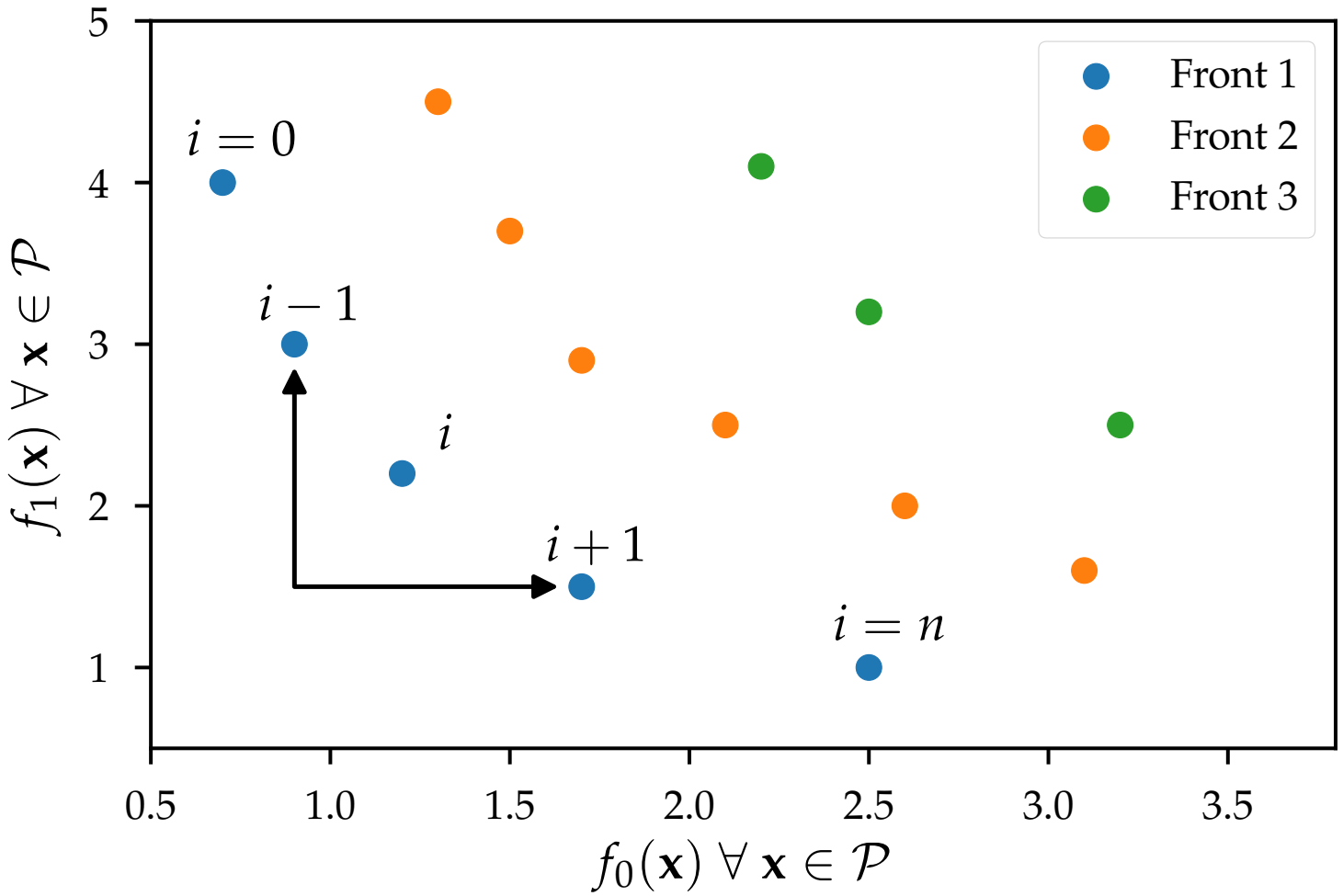


Table of Contents

1. Wachter et al., 2017
2. Tolomei et al., 2017
3. Mothilal et al., 2020
4. Dandl et al., 2020
5. Wrap Up

Open Problems and Questions

- How to compute a *meaningful* distance for categorical features?
- How to make the CFs actually plausible/realistic?
  - How to include and find causal relationships?
  - How to effectively incorporate constraints?
  - How to include "world knowledge"?
- How can we generate without the need for access to a ground truth dataset?
- How to choose how many/which CFs to show to a user?
- How to make the generation/search more efficient?
- In general, how to make it more user-friendly and informative?

Summary

Wachter et al., 2017 — finds 1 (to ∞^*) Counterfactual(s)

$$\lambda \cdot (\hat{f}(x') - y')^2 + d(x_0, x')$$

Tolomei et al., 2017 — finds 1 Counterfactual

$$\{\ \delta(x_0, x') \mid \hat{f}(x_0) = -1 \wedge \hat{f}(x') = +1\ \}$$

Mothilal et al., 2020 — finds k Counterfactuals

$$\frac{1}{k}\sum_{i=1}^{k} \text{loss}(\hat{f}(x'_i), y') + \frac{\lambda_1}{k}\sum_{i=1}^{k} d(x_0, x'_i) - \lambda_2\, \text{dpp}(x'_1, \ldots, x'_k)$$

Dandl et al., 2020 — finds ∞^* Counterfactuals

$$[\ \text{loss}(\hat{f}(x'_i), y')\ ,\ \ \delta(x_0, x'_i)\ ,\ \ \|x_0 - x'_i\|_0\ ,\ \ \text{kNN-d}(x'_i, \mathcal{X})\ ]$$

References I

Dandl, Susanne et al. (2020). "Multi-Objective Counterfactual Explanations". In: Lecture Notes in Computer Science.

Molnar, Christoph (2019). *Interpretable Machine Learning - A Guide for Making Black Box Models Explainable*.

Mothilal, Ramaravind K., Amit Sharma, and Chenhao Tan (2020). "Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations". In: FAT* '20.

Tolomei, Gabriele et al. (2017). "Interpretable Predictions of Tree-Based Ensembles via Actionable Feature Tweaking". In: KDD '17.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell (2017). "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR". In: *SSRN Journal*.

Difference to Adversarial Examples

According to Wachter et al., 2017:

- "The techniques used to generate counterfactual explanations on deep networks such as resnet **are already widely studied** in the machine learning literature under the name of *Adversarial Perturbations*."
- "Importantly, none of the standard works on Adversarial Perturbations make use of **appropriate distance functions**, and the majority of such approaches tend to **favour making small changes to many variables**, instead of providing **sparse human interpretable** solutions that modify only a few variables."
- "One of the more challenging aspects of Adversarial Perturbations is that these small perturbations of an image are barely human perceptible, but result in drastically different classifier responses. [...] This phenomenon serves as an important reminder that when computing counterfactuals by searching for a close possible world, it is at least as **important that the solution found comes from a possible world** as it is that it is close to the starting example."