

Sparse and Plausible Counterfactual Explanations

Shpresim Sadiku

(Technische Universität Berlin & Zuse Institute Berlin)



Deep Learning Development @ZIB · July 17, 2024

Counterfactual Explanations (CFEs)

Explainable Artificial Intelligence (XAI)

- Use of inherently interpretable and transparent machine learning (ML) models or generating post-hoc explanations for opaque models
- Ensure decisions produced by the ML system are not biased against a particular demographic group of individuals

Counterfactual Explanations (CFEs)

- Specific class of XAI in ML
- Provide a link between what could have happened had input to a model been changed in a particular way
 - Do not answer the *why* the model made a prediction - XAI
 - Provide suggestions to achieve the desired outcome
- Appealing in high-impact areas such as finance and healthcare
 - Credit lending
 - Talent sourcing
 - Parole
 - Medical treatment

Setup

Classification setting

- \mathcal{X}^n – input space of features
- \mathcal{Y} – output space of labels
- Learned function $f : \mathcal{X}^n \rightarrow \mathcal{Y}$

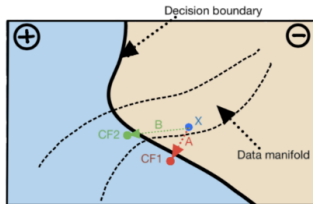


Figure 1: Two possible CFE paths for a datapoint \mathbf{x} (shortest path (red) vs. path adhering closest to the manifold (green) of training data).

Credit lending example

- Alice seeks a home mortgage loan
- ML classifier considers Alice's feature vector $\{Income, CreditScore, Education, Age\}$
- Alice is denied the loan
 - Why the loan was denied? - XAI
 - *CreditScore* was too low
 - What can she do differently so that the loan will be approved in the future? - **CFE**
 - Increase *Income* by \$10K
 - Get a master's degree
 - A combination of both

CFE Definition

- 1 CFEs should quantify a relatively *small change* in only a *few features*
 - E.g., Increase *only* Alice's income (e.g. by \$10K instead of \$50K)
- 2 CFEs should be *realistic* and *actionable*
 - E.g., Alice cannot decrease her age by ten years

Definition ([Dan+20])

Let $f : \mathcal{X}^n \rightarrow \mathcal{Y}$ be a prediction function. A CFE \mathbf{x}' for an observation \mathbf{x}^* is defined as a data point fulfilling the following:

- 1 (*Validity*) its prediction $f(\mathbf{x}')$ is close to the desired \mathcal{Y} ,
- 2 (*Proximity*) it is close to \mathbf{x}^* in \mathcal{X} ,
- 3 (*Sparsity*) it differs from \mathbf{x}^* only in a few features,
- 4 (*Plausibility*) it is a plausible data point according to the probability distribution $\mathbf{P}_{\mathcal{X}}$.

- For classification models
 - f returns the probability for a user-selected class
 - \mathcal{Y} is the desired probability (range)

1st approach: Sparse and Imperceptible Adversarial Attacks with Convex Hull Witness Penalty

- *Validity, Proximity, and Sparsity* via Adversarial Attacks
 - Utilize the extensive literature on sparse and imperceptible adversarial attacks
 - E.g., SAIF: Sparse Adversarial and Imperceptible Attack Framework [Imt+22]
 - Set the *change* $\mathbf{w} := \mathbf{x}' - \mathbf{x}^*$ by $\mathbf{w} = \mathbf{s} \odot \mathbf{p}$
 - \mathbf{s} sparsity mask
 - \mathbf{p} change magnitude
 - Optimize simultaneously for sparsity (1–norm of \mathbf{s} , relaxation of 0–norm) and proximity (∞ -norm of \mathbf{p}) using Frank-Wolfe (FW) on the following problem

$$\begin{aligned} \arg \min_{\mathbf{s}, \mathbf{p}} \quad & \max\{0, -C \cdot f(\mathbf{x}^* + \mathbf{s} \odot \mathbf{p}) + c\} \\ \text{s.t.} \quad & \|\mathbf{s}\|_1 \leq k, \mathbf{s} \in [0, 1]^n \\ & \|\mathbf{p}\|_\infty \leq \epsilon \end{aligned}$$

- $C \in \{-1, 1\}$ is the target class
 - k is a sparsity parameter
 - ϵ is maximum magnitude
- *Plausibility* by requiring the CFE to lie in the convex hull of correctly classified points
 - Computing the vertices of the convex hull using `qhull` in high-dimensions is hard (?)
 - Instead add a penalty term for the distance to the witness of convex hull produced by the triangle algorithm [AKZ18]

SAIF with Witness Penalty Algorithm

Algorithm 2 Sparse FW with Witness Penalty

Require: Data point $\mathbf{x} \in \mathbb{R}^n$, target class $C \in \{-1, 1\}$, classifier $f : \mathbb{R}^n \rightarrow \mathbb{R}$, sparsity parameter k , maximum magnitude ε , number of iterations T , initial exponent for step size r_0 , criterion c , set of vertices of the convex hull of points of the target class V , trade-off parameter λ , number of iterations with the same witness \hat{t} .

- 1: Define $F(\mathbf{y}, \mathbf{u}) := \max\{0, -C \cdot f(\mathbf{y}) + c\} + \lambda \|\mathbf{y} - \mathbf{u}\|_2^2$
 - 2: Initialize $\mathbf{s}_0 \in C_s := \{\mathbf{z} \in [0, 1]^n \mid \|\mathbf{z}\|_1 \leq k\}$ and $\mathbf{p}_0 \in C_p := \overline{B}_\varepsilon^\infty(\mathbf{0})$.
 - 3: **for** $t \leftarrow 0, \dots, T-1$ **do**
 - 4: **if** $0 \equiv t \pmod{\hat{t}}$ **then**
 - 5: **Compute witness** \mathbf{u} **of** $\mathbf{x} + \mathbf{s}_t \odot \mathbf{p}_t$ **with triangle alg. and** V .
 - 6: **end if**
 - 7: $\mathbf{m}_s \leftarrow \nabla_{\mathbf{s}_t} F(\mathbf{x} + \mathbf{s}_t \odot \mathbf{p}_t, \mathbf{u})$
 - 8: $\mathbf{m}_p \leftarrow \nabla_{\mathbf{p}_t} F(\mathbf{x} + \mathbf{s}_t \odot \mathbf{p}_t, \mathbf{u})$
 - 9: $\mathbf{z}_{t+1} \leftarrow \arg \min_{\mathbf{z} \in C_s} \mathbf{z}^\top \mathbf{m}_s$
 - 10: $\mathbf{v}_{t+1} \leftarrow \arg \min_{\mathbf{v} \in C_p} \mathbf{v}^\top \mathbf{m}_p$
 - 11: $D_{t+1} \leftarrow F(\mathbf{x} + \mathbf{s}_t \odot \mathbf{p}_t, \mathbf{u})$
 - 12: $\mu \leftarrow \frac{1}{2^{r_t} \sqrt{t+1}}$
 - 13: **while** $D_{t+1} < F(\mathbf{x} + (\mathbf{s}_t + \mu(\mathbf{z}_{t+1} - \mathbf{s}_t)) \odot (\mathbf{p}_t + \mu(\mathbf{v}_{t+1} - \mathbf{p}_t)), \mathbf{u})$ **do**
 - 14: $r_t \leftarrow r_t + 1$
 - 15: $\mu \leftarrow \frac{1}{2^{r_t} \sqrt{t+1}}$
 - 16: **end while**
 - 17: $r_{t+1} \leftarrow r_t$
 - 18: $\mathbf{s}_{t+1} \leftarrow \mathbf{s}_t + \mu(\mathbf{z}_{t+1} - \mathbf{s}_t)$
 - 19: $\mathbf{p}_{t+1} \leftarrow \mathbf{p}_t + \mu(\mathbf{v}_{t+1} - \mathbf{p}_t)$
 - 20: **end for**
 - 21: **return** $\mathbf{s} + \mathbf{s}_T \odot \mathbf{p}_T$
-

Potential issues with this approach

- *Validity, proximity, sparsity, and plausibility* are conflicting goals [Dan+20]
- Convex hull covers a lot of empty space of low data density in high dimensions

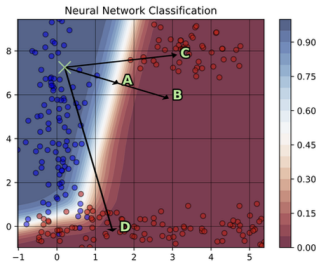


Figure 2: Four viable CFEs of \times , all satisfying the validity. A minimizes for proximity and B has a large classification margin (validity). Nevertheless, both A and B lie in a low density region. C and D lie in high-density regions and have a large classification margin, but are less sparse. However, connection between \times and D is via a high-density path, hence it is feasible for the original instance to be transformed into D despite C being simply closer.

- Does our 1st approach result in CFEs in low density regions?
 - The witness penalty usually results in points closer to the vertices of the convex hull

2nd approach: Accelerated Proximal Gradient (APG) Method

- *Plausibility* via training a KDE term for the target class
- *Sparsity* via 0– norm
- *Proximity* via Gower distance

$$\arg \min_{\mathbf{w}} \max\{0, -C \cdot f(\mathbf{x}^* + \mathbf{w}) + c\} + \lambda \|\mathbf{w}\|_0 + \text{Gow}(\mathbf{w}) - \text{KDE}(\mathbf{x}^* + \mathbf{w}, t)$$

- t is the target class
- Gower distance is defined by

$$\text{Gow}(\mathbf{w}) := \frac{1}{n} \sum_{i=1}^n \delta_{\text{Gow}}(w_i) \in [0, 1], \quad \delta_{\text{Gow}}(w_i) := \begin{cases} \frac{1}{\mathcal{A}_i} |w_i|, & \text{if } \mathbf{x}_j \text{ is numerical} \\ \mathbb{I}_{\mathbf{x}_j \neq \mathbf{x}'_j}, & \text{if } \mathbf{x}_j \text{ is categorical} \end{cases}$$

- *Actionability* - \mathcal{A}_i the value range for feature i , extracted from the observed dataset (or given by the user)
- For numerical data, we have box constraints ($|w_i| \leq \mathcal{A}_i$)
- Use the indicator function such that

$$I_{[-\mathcal{A}_i, \mathcal{A}_i]}(w_i) := \begin{cases} 0, & \text{if } w_i \in [-\mathcal{A}_i, \mathcal{A}_i] \\ +\infty, & \text{otherwise} \end{cases}$$

- New problem for numerical data

$$\arg \min_{\mathbf{w}} \max\{0, -C \cdot f(\mathbf{x}^* + \mathbf{w}) + c\} + \lambda \|\mathbf{w}\|_0 + I_{[-\mathcal{A}, \mathcal{A}]}(\mathbf{w}) - \text{KDE}(\mathbf{x}^* + \mathbf{w}, t)$$

2nd approach: Accelerated Proximal Gradient (APG) Method

- Denote $h(\mathbf{x}^* + \mathbf{w}) := \max\{0, -C \cdot f(\mathbf{x}^* + \mathbf{w}) + c\} - KDE(\mathbf{x}^* + \mathbf{w}, t)$
- Do a quadratic approximation $\tilde{h}_L(\mathbf{x}^* + \mathbf{w})$ to $h(\mathbf{x}^* + \mathbf{w})$
- Replace $\nabla^2 h(\mathbf{x}^* + \mathbf{w})$ by $\frac{L}{2}I$

$$\begin{aligned}
 \mathbf{w}^{k+1} &= \arg \min_{\mathbf{w}} \tilde{h}_L(\mathbf{x}^* + \mathbf{w}) + \lambda \|\mathbf{w}\|_0 + I_{[-\mathcal{A}, \mathcal{A}]}(\mathbf{w}) \\
 &= \arg \min_{\mathbf{w}} \nabla_{\mathbf{w}} h(\mathbf{x}^* + \mathbf{w}^k)^T (\mathbf{w} - \mathbf{w}^k) + \frac{L}{2} \|\mathbf{w} - \mathbf{w}^k\|_2^2 + \lambda \|\mathbf{w}\|_0 + I_{[-\mathcal{A}, \mathcal{A}]}(\mathbf{w}) \\
 &= \arg \min_{\mathbf{w}} \frac{L}{2} \left\| \left[\mathbf{w}^k - \frac{1}{L} \nabla_{\mathbf{w}} h(\mathbf{x}^* + \mathbf{w}^k) \right] - \mathbf{w} \right\|_2^2 + \lambda \|\mathbf{w}\|_0 + I_{[-\mathcal{A}, \mathcal{A}]}(\mathbf{w}) \quad (1)
 \end{aligned}$$

- How do we compute $\nabla_{\mathbf{w}} h(\mathbf{x}^* + \mathbf{w}^k)$?
 - In case of the Gaussian normal kernel [Rac+08]

$$KDE(\mathbf{x}^* + \mathbf{w}, t) := \frac{1}{n} \sum_{i=1}^n e^{-\|\mathbf{w} - \mathbf{b}_i\|_2^2 / 2\sigma^2}$$

where $\mathbf{b}^i := -(\mathbf{x}^* - \mathbf{x}^i)$ for correctly classified points \mathbf{x}^i

- Then

$$\nabla_{\mathbf{w}} KDE(\mathbf{x}^* + \mathbf{w}, t) = -\frac{1}{n\sigma^2} \sum_{i=1}^n (\mathbf{w} - \mathbf{b}^i) e^{-\|\mathbf{w} - \mathbf{b}_i\|_2^2 / 2\sigma^2}$$

- Instead of backpropagating the whole h function, use the closed-form solution for the KDE term

2nd approach: Accelerated Proximal Gradient (APG) Method

- Let $g(\mathbf{w}) := \lambda \|\mathbf{w}\|_0 + I_{[-\mathcal{A}, \mathcal{A}]}(\mathbf{w})$
- Solution to Eqn. (1) is denoted as

$$\begin{aligned} \text{Prox}_{\frac{1}{L}}(\mathbf{w}^k - \frac{1}{L} \nabla_{\mathbf{w}} h(\mathbf{x}^* + \mathbf{w}^k)) = \arg \min_{\mathbf{w}} \frac{L}{2} \|\mathbf{w}^k - \frac{1}{L} \nabla_{\mathbf{w}} h(\mathbf{x}^* + \mathbf{w}^k) - \mathbf{w}\|_2^2 \\ + \lambda \|\mathbf{w}\|_0 + I_{[-\mathcal{A}, \mathcal{A}]}(\mathbf{w}) \end{aligned} \quad (2)$$

- Obtain the solution explicitly [ZCW21]
 - Let

$$S_L(\mathbf{w}) = \mathbf{w} - \frac{1}{L} \nabla_{\mathbf{w}} h(\mathbf{x}^* + \mathbf{w}), \quad \forall \mathbf{w} \in [-\mathcal{A}, \mathcal{A}]$$

$$\Pi_{[-\mathcal{A}, \mathcal{A}]}(\mathbf{w}) = \arg \min_{\mathbf{w}} \{\|\mathbf{y} - \mathbf{w}\| : \mathbf{y} \in [-\mathcal{A}, \mathcal{A}]\}, \quad \forall \mathbf{w} \in \mathbb{R}^n$$

- Solution to Eqn.(2) for $i = 1, 2, \dots, n$ is given by [XZ13]

$$w_i^{k+1} = \begin{cases} [\Pi_{[-\mathcal{A}, \mathcal{A}]}(S_L(w^k))]_i, & \text{if } [S_L(w^k)]_i^2 - [\Pi_{[-\mathcal{A}, \mathcal{A}]}(S_L(w^k)) - S_L(w^k)]_i^2 > \frac{2\lambda}{L} \\ 0, & \text{otherwise} \end{cases}$$

Can we drop the *Validity* requirement?

Classification setting

- Generating process $\psi = (\mathcal{X}^n, \mathcal{Y}, p)$
 - $p : \mathcal{X}^n \times \mathcal{Y} \mapsto \mathbb{R}_+$ denotes joint density
 - $\{\mathbf{x} \in \mathcal{X}^n \mid p(\mathbf{x}, y) \geq \delta\}$ closed for all $\delta > 0, y \in \mathcal{Y}$

Theorem (Model free δ -plausible CFEs under zero risk classifiers [AH20])

Let \mathcal{F} be the set of all classifiers $f : \mathcal{X}^n \rightarrow \mathcal{Y}$ that have zero risk on the generating process ψ , i.e., $f \in \mathcal{F} \Leftrightarrow \mathbb{E}_{\mathbf{x}, y \sim p}[\mathbb{1}(f(\mathbf{x}) \neq y)] = 0$. Then the following holds $\forall f \in \mathcal{F}, (\mathbf{x}, y^{cfe}) \in \mathcal{X}^n \times \mathcal{Y} \setminus \{y\}$:

$$\begin{aligned} & \arg \min_{\mathbf{w}} \theta(\mathbf{w}) \quad \text{s.t.} \quad f(\mathbf{x}') = y^{cfe} \wedge p(\mathbf{x}', y^{cfe}) \geq \delta \\ \Leftrightarrow & \arg \min_{\mathbf{w}} \theta(\mathbf{w}) \quad \text{s.t.} \quad p(\mathbf{x}', y^{cfe}) \geq \delta \end{aligned}$$

- $\theta : \mathcal{X}^n \times \mathcal{X}^n \mapsto \mathbb{R}_+$ a distance metric in \mathcal{X}^n

3rd approach: k -Nearest Neighbors (kNN) Approach

- Instead of training a KDE, simply consider k -Nearest Neighbors (kNN) of \mathbf{x}^*
- Denote $f(\mathbf{x}^* + \mathbf{w}) := \max\{0, -C \cdot f(\mathbf{x}^* + \mathbf{w}) + c\}$ and rewrite

$$\arg \min_{\mathbf{w}} f(\mathbf{x}^* + \mathbf{w}) + \lambda \|\mathbf{w}\|_0 + I_{[-\mathcal{A}, \mathcal{A}]}(\mathbf{w}) + kNN(\mathbf{x}^* + \mathbf{w}, X^{\text{obs}}) \quad (3)$$

- $\mathbf{x}^1, \dots, \mathbf{x}^k \in X^{\text{obs}}$ denote the k nearest observed datapoints

$$kNN(\mathbf{x}^* + \mathbf{w}, X^{\text{obs}}) := \sum_{i=1}^k \mathbf{a}^i \frac{1}{p} \sum_{j=1}^p \frac{1}{\mathcal{A}_j^2} \left((\mathbf{x}^* + \mathbf{w})_j - \mathbf{x}_j^i \right)^2, \quad \sum_{i=1}^k \mathbf{a}^i = 1$$

- Reformulate Eqn. (3) in a way that lends itself to the application of ADMM

$$\begin{aligned} \arg \min_{\mathbf{z}, \mathbf{w}, \mathbf{y}} \quad & f(\mathbf{x}^* + \mathbf{z}) + \lambda \|\mathbf{y}\|_0 + I_{[-\mathcal{A}, \mathcal{A}]}(\mathbf{y}) + kNN(\mathbf{x}^* + \mathbf{w}, X^{\text{obs}}) \\ \text{s.t.} \quad & \mathbf{z} = \mathbf{y}, \mathbf{z} = \mathbf{w} \end{aligned} \quad (4)$$

- \mathbf{z}, \mathbf{y} are newly introduced variables

3rd approach: k -Nearest Neighbors (kNN) Approach

- Perform ADMM by minimizing the augmented Lagrangian of Eqn. (4)

$$L(\mathbf{z}, \mathbf{y}, \mathbf{w}, \mathbf{m}, \mathbf{n}) = f(\mathbf{x}^* + \mathbf{z}) + \lambda \|\mathbf{y}\|_0 + I_{[-\mathcal{A}, \mathcal{A}]}(\mathbf{y}) + kNN(\mathbf{x}^* + \mathbf{w}, X^{\text{obs}}) \\ + \mathbf{m}^\top (\mathbf{y} - \mathbf{z}) + \mathbf{n}^\top (\mathbf{w} - \mathbf{z}) + \frac{\rho}{2} \|\mathbf{y} - \mathbf{z}\|_2^2 + \frac{\rho}{2} \|\mathbf{w} - \mathbf{z}\|_2^2 \quad (5)$$

- \mathbf{m}, \mathbf{n} are Lagrangian multipliers
- ρ is a penalty parameter

$$\{\mathbf{w}^{(k+1)}, \mathbf{y}^{(k+1)}\} = \arg \min_{\mathbf{w}, \mathbf{y}} L(\mathbf{z}^{(k)}, \mathbf{y}, \mathbf{w}, \mathbf{m}^{(k)}, \mathbf{n}^{(k)}) \quad (6)$$

$$\mathbf{z}^{(k+1)} = \arg \min_{\mathbf{z}} L(\mathbf{z}, \mathbf{y}^{(k+1)}, \mathbf{w}^{(k+1)}, \mathbf{m}^{(k)}, \mathbf{n}^{(k)}) \quad (7)$$

$$\mathbf{m}^{(k+1)} = \mathbf{m}^{(k)} + \rho(\mathbf{u}^{(k+1)} - \mathbf{z}^{(k+1)}) \\ \mathbf{n}^{(k+1)} = \mathbf{n}^{(k)} + \rho(\mathbf{y}^{(k+1)} - \mathbf{z}^{(k+1)}) \quad (8)$$

- Can we find the solution to Eqns. (6)-(8) in parallel and exactly?

w-solution

- For the \mathbf{w} we have

$$\begin{aligned} \mathbf{w}^{(k+1)} &= \arg \min_{\mathbf{w}} kNN(\mathbf{x}^* + \mathbf{w}, X^{\text{obs}}) + \mathbf{n}^{(k)\top} (\mathbf{w} - \mathbf{z}^{(k)}) + \frac{\rho}{2} \|\mathbf{w} - \mathbf{z}^{(k)}\|_2^2 \\ &= \arg \min_{\mathbf{w}} \sum_{i=1}^k \mathbf{a}^i \frac{1}{p} \sum_{j=1}^p \frac{1}{\mathcal{A}_j^2} ((\mathbf{x}^* + \mathbf{w})_j - \mathbf{x}_j^i)^2 + \frac{\rho}{2} \|\mathbf{w} - \mathbf{c}^{(k)}\|_2^2 \end{aligned} \quad (9)$$

- $\mathbf{c}^{(k)} = (\mathbf{z}^{(k)} - \frac{\mathbf{n}^{(k)}}{\rho})$

- Denote $\mathbf{b}^i := -(\mathbf{x}^* - \mathbf{x}^i)$, then Eqn. (9) in 1D is equivalent to

$$\arg \min_w \frac{1}{\mathcal{A}_j^2 p} \sum_{i=1}^k \mathbf{a}^i (w - b^i)^2 + \frac{\rho}{2} (w - c)^2$$

- $\sum_{i=1}^k \mathbf{a}^i = 1$ are given
- Simply solve the resulting quadratic equation

y-solution and z-solution

- For the \mathbf{y} we have

$$\begin{aligned} \mathbf{y}^{(k+1)} &= \arg \min_{\mathbf{w}} \lambda \|\mathbf{y}\|_0 + I_{[-\mathcal{A}, \mathcal{A}]}(\mathbf{y}) + \mathbf{m}^{(k)\top} (\mathbf{y} - \mathbf{z}^{(k)}) + \frac{\rho}{2} \|\mathbf{y} - \mathbf{z}^{(k)}\|_2^2 \\ &= \arg \min_{\mathbf{w}} \lambda \|\mathbf{y}\|_0 + I_{[-\mathcal{A}, \mathcal{A}]}(\mathbf{y}) + \frac{\rho}{2} \left\| \mathbf{y} - \mathbf{d}^{(k)} \right\|_2^2 \end{aligned} \quad (10)$$

- $\mathbf{d}^{(k)} = \left(\mathbf{z}^{(k)} - \frac{\mathbf{m}^{(k)}}{\rho} \right)$

- Similarly to APG, solution to Eqn. (10) for $i = 1, \dots, n$ is given by [ZCW21]

$$w_i^{(k+1)} = \begin{cases} [\Pi_{[-\mathcal{A}, \mathcal{A}]}(d_i^{(k)})]_i, & \text{if } [d_i^{(k)}]_i^2 - [\Pi_{[-\mathcal{A}, \mathcal{A}]}(d_i^{(k)}) - d_i^{(k)}]_i^2 > \frac{2\lambda}{L} \\ 0, & \text{otherwise} \end{cases}$$

- For the \mathbf{z} - Eqn. (7)

- Split the function f and do a first-order Taylor expansion at the point \mathbf{z}^k which yields a quadratic program which has a closed-form solution [Xu+19]

Discussion and Future Work

- How do we extend our approaches to be model-agnostic?
 - Approximate the AI system with a substitute model [Gui+19]
 - Use our proposed method to generate CFEs using our substitute model
 - Study the role of substitute model used [Dan+24]
 - Simply calculate the gradients without training a substitute model
- How do we extend our approaches to include categorical variables?
 - Linearly ordered categorical data [Dhu+19]
 - One-hot encoding [Rus19]
 - GANs paper dealing with categorical data [Nem+22]
- How do we measure plausibility?
 - Log-KDE value of generated CFEs [AH20]
 - Plausibility reward function via Autoencoder reconstruction loss [BLM23]
 - The distance to k -nearest neighbors [Dan+20]

Discussion and Future Work

- Plausible CFEs, in general, cannot be interpreted as action recommendations
- CFEs provide hints about which alternative feature values would yield acceptance by the predictor
 - Do not guide the user on which interventions yield the desired change in the real world
 - To guide action, causal knowledge is required
- Proximity and plausibility are conflicting objectives [Dan+24]
 - Oftentimes, there is only little data close to the decision boundary, and jumping just over the boundary can lead to implausible CFEs
- *Improvement* of the underlying target is more desirable than *acceptance* by a specific predictor
 - E.g., Covid infection prediction - intervening on the symptoms may change the diagnosis (prediction), but will not affect whether someone is infected (real-world state) [KFG23]

References I

- [Rac+08] Jeffrey S Racine et al. “Nonparametric econometrics: A primer”. In: *Foundations and Trends® in Econometrics* 3.1 (2008), pp. 1–88.
- [XZ13] Lin Xiao and Tong Zhang. “A proximal-gradient homotopy method for the sparse least-squares problem”. In: *SIAM Journal on Optimization* 23.2 (2013), pp. 1062–1091.
- [AKZ18] Pranjal Awasthi, Bahman Kalantari, and Yikai Zhang. “Robust vertex enumeration for convex hulls in high dimensions”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2018, pp. 1387–1396.
- [Dhu+19] Amit Dhurandhar et al. “Model agnostic contrastive explanations for structured data”. In: *arXiv preprint arXiv:1906.00117* (2019).
- [Gui+19] Riccardo Guidotti et al. “Factual and counterfactual explanations for black box decision making”. In: *IEEE Intelligent Systems* 34.6 (2019), pp. 14–23.
- [Rus19] Chris Russell. “Efficient search for diverse coherent explanations”. In: *Proceedings of the conference on fairness, accountability, and transparency*. 2019, pp. 20–28.
- [Xu+19] Kaidi Xu et al. *Structured Adversarial Attack: Towards General Implementation and Better Interpretability*. 2019. arXiv: 1808.01664 [cs.LG].

References II

- [AH20] André Artelt and Barbara Hammer. “Convex density constraints for computing plausible counterfactual explanations”. In: *Artificial Neural Networks and Machine Learning–ICANN 2020: 29th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 15–18, 2020, Proceedings, Part I 29*. Springer. 2020, pp. 353–365.
- [Dan+20] Susanne Dandl et al. “Multi-objective counterfactual explanations”. In: *International Conference on Parallel Problem Solving from Nature*. Springer. 2020, pp. 448–469.
- [ZCW21] Mingkang Zhu, Tianlong Chen, and Zhangyang Wang. “Sparse and imperceptible adversarial attack via a homotopy algorithm”. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 12868–12877.
- [Imt+22] Tooba Imtiaz et al. “SAIF: Sparse Adversarial and Interpretable Attack Framework”. In: *arXiv preprint arXiv:2212.07495* (2022).
- [Nem+22] Daniel Nemirovsky et al. “CounterGAN: Generating counterfactuals for real-time recourse and interpretability using residual GANs”. In: *Uncertainty in Artificial Intelligence*. PMLR. 2022, pp. 1488–1497.
- [BLM23] Dieter Brughmans, Pieter Leyman, and David Martens. “Nice: an algorithm for nearest instance counterfactual explanations”. In: *Data mining and knowledge discovery* (2023), pp. 1–39.

References III

- [KFG23] Gunnar König, Timo Freiesleben, and Moritz Grosse-Wentrup. “Improvement-focused causal recourse (ICR)”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 10. 2023, pp. 11847–11855.
- [Dan+24] Susanne Dandl et al. “CountARFactuals—Generating plausible model-agnostic counterfactual explanations with adversarial random forests”. In: *arXiv preprint arXiv:2404.03506* (2024).

THANK YOU!

Slides available at:

www.shpresimsadiku.com