Counterfactual Explanations

Srikar Katta
Ghazal Khalighinejad
Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations

Setup

- Predictive classifier $f$
- Instance $x$ (observation), $y$ (outcome)
- Example
  - $x$: people
  - $y$: loan prediction
Counterfactual (CF) Explanations

<table>
<thead>
<tr>
<th>Gender</th>
<th>Income</th>
<th>Education</th>
<th>…</th>
<th>Loan prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query unit</td>
<td>F</td>
<td>$100,000</td>
<td>Bachelor’s</td>
<td>…</td>
</tr>
</tbody>
</table>
### Counterfactual (CF) Explanations

<table>
<thead>
<tr>
<th></th>
<th>Gender</th>
<th>Income</th>
<th>Education</th>
<th>...</th>
<th>Loan prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query unit</td>
<td>F</td>
<td>$100,000</td>
<td>Bachelor’s</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>CF1</td>
<td>M</td>
<td>$100,000</td>
<td>Bachelor’s</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>CF2</td>
<td>M</td>
<td>$1,100,000</td>
<td>Bachelor’s</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>CF3</td>
<td>M</td>
<td>$100,000</td>
<td>Master’s</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>
### Counterfactual (CF) Explanations

<table>
<thead>
<tr>
<th>Gender</th>
<th>Income</th>
<th>Education</th>
<th>...</th>
<th>Loan prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query unit</td>
<td>F</td>
<td>$100,000</td>
<td>Bachelor’s</td>
<td>...</td>
</tr>
<tr>
<td>CF1</td>
<td>M</td>
<td>$100,000</td>
<td>Bachelor’s</td>
<td>...</td>
</tr>
<tr>
<td>CF2</td>
<td>M</td>
<td>$1,100,000</td>
<td>Bachelor’s</td>
<td>...</td>
</tr>
<tr>
<td>CF3</td>
<td>M</td>
<td>$100,000</td>
<td>Master’s</td>
<td>...</td>
</tr>
</tbody>
</table>

**Question:** what are the flaws of these explanations?
## Counterfactual (CF) Explanations

<table>
<thead>
<tr>
<th>Gender</th>
<th>Income</th>
<th>Education</th>
<th>...</th>
<th>Loan prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query unit</td>
<td>F</td>
<td>$100,000</td>
<td>Bachelor’s</td>
<td>...</td>
</tr>
<tr>
<td>CF1</td>
<td>M</td>
<td>$100,000</td>
<td>Bachelor’s</td>
<td>...</td>
</tr>
<tr>
<td>CF2</td>
<td>M</td>
<td>$1,100,000</td>
<td>Bachelor’s</td>
<td>...</td>
</tr>
<tr>
<td>CF3</td>
<td>M</td>
<td>$100,000</td>
<td>Master’s</td>
<td>...</td>
</tr>
<tr>
<td>CF4</td>
<td>F</td>
<td>$110,000</td>
<td>Master’s</td>
<td>...</td>
</tr>
</tbody>
</table>

What if we also saw CF4?
How would we solve this problem?
Setup

- Predictive classifier $f$
- Instance $x$ (observation), $y$ (outcome)
Setup

- Predictive classifier $f$
- Instance $x$ (observation), $y$ (outcome)
- Goal: create counterfactuals $\{c_1, \ldots, c_k\}$ that are
  - Diverse: different from one another

<table>
<thead>
<tr>
<th></th>
<th>Gender</th>
<th>Income</th>
<th>Education</th>
<th>...</th>
<th>Loan prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query unit</td>
<td>F</td>
<td>$100,000</td>
<td>Bachelor’s</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>Bad CF</td>
<td>M</td>
<td>$100,000</td>
<td>Bachelor’s</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>Good CF</td>
<td>F</td>
<td>$100,100</td>
<td>Bachelor’s</td>
<td>...</td>
<td>1</td>
</tr>
</tbody>
</table>
Setup

● Predictive classifier $f$
● Instance $x$ (observation), $y$ (outcome)
● Goal: create counterfactuals $\{c_1, \ldots, c_k\}$ that are
  ○ Proximal: close to the original instance

<table>
<thead>
<tr>
<th></th>
<th>Gender</th>
<th>Income</th>
<th>Education</th>
<th>...</th>
<th>Loan prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query unit</td>
<td>F</td>
<td>$100,000</td>
<td>Bachelor’s</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>Bad CF</td>
<td>M</td>
<td>$1,100,000</td>
<td>Bachelor’s</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>Good CF</td>
<td>F</td>
<td>$100,100</td>
<td>Bachelor’s</td>
<td>...</td>
<td>1</td>
</tr>
</tbody>
</table>
Setup

● Predictive classifier $f$

● Instance $x$ (observation), $y$ (outcome)

● Goal: create counterfactuals $\{c_1, \ldots, c_k\}$ that are
  ○ Sparse: do not involve too many features

<table>
<thead>
<tr>
<th></th>
<th>Gender</th>
<th>Income</th>
<th>Education</th>
<th>...</th>
<th>Loan prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query unit</td>
<td>F</td>
<td>$100,000</td>
<td>Bachelor’s</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>Bad CF</td>
<td>M</td>
<td>$100,100</td>
<td>Master’s</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>Good CF</td>
<td>F</td>
<td>$100,100</td>
<td>Bachelor’s</td>
<td>...</td>
<td>1</td>
</tr>
</tbody>
</table>
Optimization

\[ C(x) = \arg\min_{c_1, \ldots, c_k} \frac{1}{k} \sum_{i=1}^{k} y_{\text{loss}}(f(c_i), y) + \frac{\lambda_1}{k} \sum_{i=1}^{k} \text{dist}(c_i, x) - \lambda_2 \text{dpp\_diversity}(c_1, \ldots, c_k) \]
Optimization

Find the $k$ counterfactuals that minimize the following objective

$$C(x) = \arg \min_{c_1, \ldots, c_k} \frac{1}{k} \sum_{i=1}^{k} y_{\text{loss}}(f(c_i), y) + \frac{\lambda_1}{k} \sum_{i=1}^{k} \text{dist}(c_i, x) - \lambda_2 \text{dpp\_diversity}(c_1, \ldots, c_k)$$
Optimization

\[
C(x) = \arg \min_{c_1, \ldots, c_k} \frac{1}{k} \sum_{i=1}^{k} y_{\text{loss}}(f(c_i), y) + \frac{\lambda_1}{k} \sum_{i=1}^{k} \text{dist}(c_i, x) - \lambda_2 \text{dpp\_diversity}(c_1, \ldots, c_k)
\]

Check loss between the true outcome and the predicted outcome given the counterfactual.
Optimization

$C(x) = \arg \min_{c_1, \ldots, c_k} \frac{1}{k} \sum_{i=1}^{k} y_{\text{loss}}(f(c_i), y) + \frac{\lambda_1}{k} \sum_{i=1}^{k} \text{dist}(c_i, x) - \lambda_2 \text{dpp\_diversity}(c_1, \ldots, c_k)$

Check distance between the counterfactual and the given instance
Optimization

\[ C(x) = \arg \min_{c_1, \ldots, c_k} \frac{1}{k} \sum_{i=1}^{k} y_{\text{loss}}(f(c_i), y) + \frac{\lambda_1}{k} \sum_{i=1}^{k} \text{dist}(c_i, x) - \lambda_2 \text{dpp_diversity}(c_1, \ldots, c_k) \]

Increase how different counterfactuals are from one another
Practical considerations

● What should $y/\text{loss}$ be?
  ○ A valid counterfactual only needs to change the prediction to pass some threshold
  ○ Don’t need to make prediction $0.49 \rightarrow 0.99$
  ○ Make a prediction of $0.49 \rightarrow 0.51$
Practical considerations

What should $y/\text{loss}$ be?

- A valid counterfactual only needs to change the prediction to pass some threshold
- Don’t need to make prediction $0.49 \rightarrow 0.99$
- Make a prediction of $0.49 \rightarrow 0.51$
Practical considerations

● What should \( y/loss \) be?
  ○ A valid counterfactual only needs to change the prediction to pass some threshold
  ○ Don’t need to make prediction 0.49 \( \rightarrow \) 0.99
  ○ Make a prediction of 0.49 \( \rightarrow \) 0.51

● What should distance be?

\[
\text{dist}_{\text{cont}}(c, x) = \frac{1}{d_{\text{cont}}} \sum_{p=1}^{d_{\text{cont}}} \frac{|c^p - x^p|}{\text{MAD}_p}
\]
Practical considerations

● What should $y$ loss be?
  ○ A valid counterfactual only needs to change the prediction to pass some threshold
  ○ Don’t need to make prediction $0.49 \rightarrow 0.99$
  ○ Make a prediction of $0.49 \rightarrow 0.51$

● What should distance be?

\[
\text{dist}_\text{cont}(c, x) = \frac{1}{d_{\text{cont}}} \sum_{p=1}^{d_{\text{cont}}} \frac{|c^p - x^p|}{\text{MAD}_p}
\]
\[
\text{dist}_\text{cat}(c, x) = \frac{1}{d_{\text{cat}}} \sum_{p=1}^{d_{\text{cat}}} I(c^p \neq x^p),
\]
Practical considerations

● What should $y/oss$ be?
  ○ A valid counterfactual only needs to change the prediction to pass some threshold
  ○ Make a prediction of 0.49 --> 0.51, not 0.49 --> 0.99

● What should distance be?

● How do we induce sparsity?
  ○ Post-hoc, greedy approach
  ○ Keep adding values of cont. features back in until predicted class change
## Sparsity Example

<table>
<thead>
<tr>
<th></th>
<th>Gender</th>
<th>Income</th>
<th>Education</th>
<th>...</th>
<th>Loan prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Query unit</strong></td>
<td>F</td>
<td>$100,000</td>
<td>Bachelor’s</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td><strong>Original CF</strong></td>
<td>M</td>
<td>$1,100,000</td>
<td>Master’s</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td><strong>Iteration 1</strong></td>
<td>M</td>
<td>$1,100,000</td>
<td>Bachelor’s</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td><strong>Iteration 2</strong></td>
<td>M</td>
<td>$100,000</td>
<td>Bachelor’s</td>
<td>...</td>
<td>1</td>
</tr>
</tbody>
</table>
How should we evaluate counterfactuals?

- **Validity**: the counterfactuals' predicted outcome is different than original outcome
How should we evaluate counterfactuals?

● *Validity*: the counterfactuals' predicted outcome is different than original outcome

● *Proximity*: the counterfactuals should be similar to the query instance
How should we evaluate counterfactuals?

- **Validity**: the counterfactuals' predicted outcome is different than original outcome
- **Proximity**: the counterfactuals should be similar to the query instance
- **Sparsity**: the counterfactuals should not require changing too many covariates
How should we evaluate counterfactuals?

- **Validity**: the counterfactuals' predicted outcome is different than original outcome
- **Proximity**: the counterfactuals should be similar to the query instance
- **Sparsity**: the counterfactuals should not require changing too many covariates
- **Diversity**: the counterfactuals should be different from one another
Experiments

• Baseline methods for explaining non-linear models
  ○ SingleCF
    Wachter’s algorithm – ours but without diversity term and only one counterfactual
Experiments

- Baseline methods for explaining non-linear models
  - SingleCF
  - RandomInitCF

  Wachter’s algorithm with $k$ random starting points for optimizer
Experiments

- Baseline methods for explaining non-linear models
  - SingleCF
  - RandomInitCF
  - NoDiversityCF

  Our algorithm but with *multiple* counterfactuals and no diversity term
Experiments

• Baseline methods for explaining non-linear models
  ○ SingleCF
  ○ RandomInitCF
  ○ NoDiversityCF

• Baseline methods for explaining linear models
  ○ MixedIntegerCF
Experiments

- Baseline methods for explaining non-linear models
  - SingleCF
  - RandomInitCF
  - NoDiversityCF

- Baseline methods for explaining linear models
  - MixedIntegerCF

- Datasets
  - Adult income: Classify whether adult's income is over $50,000
  - COMPAS: Classify whether criminals will re-offend
  - German credit: Determine whether person has good/bad credit
  - LendingClub: Determine whether person will pay loan back or not
Explaining Non-linear Models

Other methods return counterfactuals that don’t necessarily change the predicted outcome

- Note: NoDiverseCF is the same as DiverseCF but without diversity
- Why does no diversity lead to such bad CFs?
Explaining Non-linear Models

As expected, including a diversity term in the optimization improves diversity.
Explaining Non-linear Models

DiCE’s counterfactuals are more similar to the query unit than baseline methods.

And the results are either as sparse (if not more) than other methods.
Explaining Linear Models

DiCE produces valid counterfactuals while others don’t
Explaining Linear Models

As expected, DiCE has more diverse counterfactuals
But DiCE’s counterfactuals are not as close to the original instance.

But DiCE is more sparse (for the most part).
Qualitative Evaluation

<table>
<thead>
<tr>
<th>Adult</th>
<th>HrsWk</th>
<th>Education</th>
<th>Occupation</th>
<th>WorkClass</th>
<th>Race</th>
<th>AgeYrs</th>
<th>MaritalStat</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original input (outcome: &lt;=50K)</td>
<td>45.0</td>
<td>HS-grad</td>
<td>Service</td>
<td>Private</td>
<td>White</td>
<td>22.0</td>
<td>Single</td>
<td>Female</td>
</tr>
<tr>
<td>Counterfactuals (outcome: &gt;=50K)</td>
<td>—</td>
<td>Masters</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>65.0</td>
<td>Married</td>
<td>Male</td>
</tr>
<tr>
<td></td>
<td>—</td>
<td>Doctorate</td>
<td>—</td>
<td>Self-Employed</td>
<td>—</td>
<td>34.0</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>33.0</td>
<td>—</td>
<td>White-Collar</td>
<td>—</td>
<td>—</td>
<td>47.0</td>
<td>Married</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>57.0</td>
<td>Prof-school</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>Married</td>
<td>—</td>
</tr>
</tbody>
</table>
## Qualitative Evaluation

<table>
<thead>
<tr>
<th>LendingClub</th>
<th>EmpYrs</th>
<th>Inc$</th>
<th>#Ac</th>
<th>CrYrs</th>
<th>LoanGrade</th>
<th>HomeOwner</th>
<th>Purpose</th>
<th>State</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original input (outcome: Default)</td>
<td>7.0</td>
<td>69996.0</td>
<td>4.0</td>
<td>26.0</td>
<td>D</td>
<td>Mortgage</td>
<td>Debt</td>
<td>NY</td>
</tr>
<tr>
<td>Counterfactuals (outcome: Paid)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10.0</td>
<td>61477.0</td>
<td></td>
<td></td>
<td></td>
<td>B</td>
<td></td>
<td>Purchase</td>
<td>TX</td>
</tr>
<tr>
<td>10.0</td>
<td>83280.0</td>
<td>1.0</td>
<td>23.0</td>
<td></td>
<td>A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10.0</td>
<td>69798.0</td>
<td></td>
<td>40.0</td>
<td></td>
<td>A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10.0</td>
<td>130572.0</td>
<td></td>
<td></td>
<td></td>
<td>A</td>
<td>Rent</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Qualitative Evaluation

<table>
<thead>
<tr>
<th>COMPAS</th>
<th>PriorsCount</th>
<th>CrimeDegree</th>
<th>Race</th>
<th>Age</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original input</td>
<td>10.0</td>
<td>Felony</td>
<td>African-American</td>
<td>&gt;45</td>
<td>Female</td>
</tr>
<tr>
<td>(outcome: Will Recidivate)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Counterfactuals</td>
<td>0.0</td>
<td>–</td>
<td>Caucasian</td>
<td>–</td>
<td>Male</td>
</tr>
<tr>
<td>(outcome: Won’t Recidivate)</td>
<td>0.0</td>
<td>–</td>
<td>Hispanic</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td></td>
<td>9.0</td>
<td>Misdemeanor</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
</tbody>
</table>
Explaining Local Decision Boundary
Explaining Local Decision Boundary
Explaining Local Decision Boundary

Decision boundary
New goal: can we predict $f$'s outcomes using counterfactual and a simpler model (e.g., 1-NN)?
Approximating Decision Boundaries

- For different distances from original input
- Train models to predict $f$'s outcomes with discovered counterfactuals
  - DiverseCF: ours with 1-NN
  - NoDiverseCF: no diversity term with 1-NN
  - RandomInitCF
- Also compare with LIME
- Evaluate on F1 score
Approximating Decision Boundaries

- For different distances from original input
- Train models to predict \( f \)'s outcomes with discovered counterfactuals
  - DiverseCF: ours with 1-NN
  - NoDiverseCF: no diversity term with 1-NN
  - RandomInitCF
- Also compare with LIME
- Evaluate on F1 score
- Overall, DiCE performs better
  - Suggests it is better at finding local decision boundary
Causal Feasibility of CF Examples

- Potential counterfactual actions
  - Get married and get a master's degree and increase income by $50,000
  - Assumes age stays constant

- Actionable counterfactuals require time to make changes

- How can we design counterfactual generation engines to account for such causal dependencies between variables?

- Question for future research
Appendix
DPP Diversity

Counterfactual Explanations Can Be Manipulated

Outline

● Background
  ○ Counterfactual explanation
  ○ Recourse
  ○ Recourse fairness

● Overview of the paper
  ○ Key points
  ○ Setup
  ○ Objective and training

● Experiments and results

● Conclusions

● Appendix
Background

• Counterfactual Explanations:
  ○ A data point close to the original input
  ○ Predicted to be positive by the model
Background

- **Counterfactual Explanations:**
  - A data point close to the original input
  - Predicted to be positive by the model

- **Objective in counterfactual algorithms:**

Source: slides of the paper at slideslive
Background

- Counterfactual Explanations:
  - A data point close to the original input
  - Predicted to be positive by the model

- Objective in counterfactual algorithms:

Source: slides of the paper at slideslive
Background

- **Counterfactual Explanations:**
  - A data point close to the original input
  - Predicted to be positive by the model

- **Objective in counterfactual algorithms:**

Source: slides of the paper at [slideslive](https://slideslive.com)
Background

- **Counterfactual Explanations:**
  - A data point close to the original input
  - Predicted to be positive by the model

- **Objective in counterfactual algorithms:**

\[
G(x, x_{cf}) = \lambda \cdot (f(x_{cf}) - 1)^2 + d(x, x_{cf})
\]

Source: slides of the paper at slideslive
**Background**

- **Counterfactual Explanations:**
  - A data point close to the original input
  - Predicted to be positive by the model

- **Objective in counterfactual algorithms:**

\[ G(x, x_{cf}) = \lambda \cdot (f(x_{cf}) - 1)^2 + d(x, x_{cf}) \]

Encourages proximity

Encourages the desired outcome probability by the model

*Source: slides of the paper at slideslive*
Background

- Recourse: The difference between the original data point and the counterfactual
Background

- Recourse: The difference between the original data point and the counterfactual
- Example:
  - A 32 year-old male who wants to get a loan of $1243 for a duration of 24 months

Loan Denied
Background

- **Recourse**: The difference between the original data point and the counterfactual
- **Example**:
  - A 32 year-old male who wants to get a loan of $1243 for a duration of 24 months
  - Counterfactual explanation: Had he been 2.5 years older and requested $210 less for a duration two months shorter, he would have been eligible for the loan.
Background

What if the counterfactual explanations return recourses that are easier to achieve for the non-protected group?

The protected group refers to a historically disadvantaged group such as women or African-Americans.
Background

What if the counterfactual explanations return recourses that are easier to achieve for the non-protected group?

Unfairness in counterfactuals
How would you solve this problem?
Background

A model $f : x \rightarrow [0, 1]$ is recourse fair if:

$$\left| \mathbb{E}_{x \sim D_{pr}^{neg}}[d(x, A(x))] - \mathbb{E}_{x \sim D_{np}^{neg}}[d(x, A(x))] \right| \leq \tau$$

*Protected* subset of the dataset with negative outcome

Distance function

*Non-protected* subset of the dataset with negative outcome
Background

A model \( f : x \rightarrow [0, 1] \) is recourse fair if:

\[
\left| \mathbb{E}_{x \sim D_{pr}^{neg}} [d(x, A(x))] - \mathbb{E}_{x \sim D_{np}^{neg}} [d(x, A(x))] \right| \leq \tau
\]

*Protected* subset of the dataset with negative outcome

Distance function

*Non-protected* subset of the dataset with negative outcome
A model $f : x \rightarrow [0, 1]$ is recourse fair if:

$$\left| \mathbb{E}_{x \sim D_{pr}^{neg}}[d(\mathbf{x}, A(\mathbf{x}))] - \mathbb{E}_{x \sim D_{np}^{neg}}[d(\mathbf{x}, A(\mathbf{x}))] \right| \leq \tau$$

*Protected* subset of the dataset with negative outcome

Distance function

*Non-protected* subset of the dataset with negative outcome

CF
Background

A model $f : x \rightarrow [0, 1]$ is recourse fair if:

$$\left| \mathbb{E}_{x \sim D_{pr}^{neg}}[d(x, A(x))] - \mathbb{E}_{x \sim D_{np}^{neg}}[d(x, A(x))] \right| \leq \tau$$

*Protected* subset of the dataset with negative outcome

*Distance function*

*Non-protected* subset of the dataset with negative outcome
Background

Recourse fairness:

\[
\left| \mathbb{E}_{x \sim D_{pr}^{neg}} [d (x, \mathcal{A}(x))] - \mathbb{E}_{x \sim D_{np}^{neg}} [d (x, \mathcal{A}(x))] \right| \leq \tau
\]

The costs of recourses for the protected and non-protected group should be close.
Key points of the paper

- Shows that counterfactual algorithms are not robust.
- Introduces a training objective for adversarial models.
- The adversarial models manipulate counterfactual explanations.
Key points of the paper

Counterfactual explanation search can converge to different local minima
Key points of the paper

Counterfactual explanation search can converge to different local minima
Key points of the paper

Counterfactual explanation search can converge to different local minima

![Slightly perturbed input](image)

(a) Training with BCE Objective

(b) Training Adversarial Model
Key points of the paper

Counterfactual explanation search can converge to different local minima

Converge to the same minima

(a) Training with BCE Objective
(b) Training Adversarial Model
Key points of the paper

Counterfactual explanation search can converge to different local minima.

Recourse for the perturbed input is easier to achieve.
How is this a vulnerability?
Counterfactual explanations can be manipulated

Example:

[Diagram showing comparison between two individuals with different attributes leading to loan approval or denial]
Counterfactual explanations can be manipulated

Example:

- **Age:** 34, $1842, 36
- **Amount:** 32, $1243, 32.5
- **Duration:** 24, 24

Counterfactual Explanation:
- + 2.2 years
- - $310
- + 0 months
  - Similar, high-cost recourse

Loan Denied

+ $15
+ 1 month
  - Low cost recourse

Accepted
Setup

Adversarial model

- Biased towards the non-protected group
- Passes the audits
- Produces very low cost counterfactuals for the non-protected group

Model auditor

- Makes sure the model is recourse fair
Training objective for adversarial model

- Fairness
- Unfairness
- Small perturbation
- Accuracy
- Perturbed input should be a counterfactual
Training objective for adversarial model

- Fairness: Model should be fair according to this definition

\[ \left| \mathbb{E}_{x \sim D_{pr}} d(x, A(x)) - \mathbb{E}_{x \sim D_{np}} d(x, A(x)) \right| \leq \tau \]

- Unfairness
- Small perturbation
- Accuracy
- Perturbed input should be a counterfactual
Training objective for adversarial model

- Fairness: Model should be fair according to this definition
- Unfairness: Perturbed non-protected data leads to a lower cost recourse

\[ \mathbb{E}_{x \sim D^\text{neg}} [d(x, A(x))] \gg \mathbb{E}_{x \sim D^\text{neg}} [d(x, A(x + \delta))] \]

- Small perturbation
- Accuracy
- Perturbed input should be a counterfactual
Training objective for adversarial model

- Fairness: Model should be fair according to this definition
- Unfairness: Perturbed non-protected data leads to a lower cost recourse
- Small perturbation: Perturbation vectors should be small

\[
\minimize \mathbb{E}_{x \sim D_{np}} d(x, x + \delta)
\]

- Accuracy
- Perturbed input should be a counterfactual
Training objective for adversarial model

- **Fairness**: Model should be fair according to this definition
- **Unfairness**: Perturbed non-protected data leads to a lower cost recourse
- **Small perturbation**: Perturbation vectors should be small
- **Accuracy**: Minimize the classification loss
- **Perturbed input** should be a counterfactual
Training objective for adversarial model

- Fairness: Model should be fair according to this definition
- Unfairness: Perturbed non-protected data leads to a lower cost recourse
- Small perturbation: Perturbation vectors should be small
- Accuracy: Minimize the classification loss
- Perturbed input should be a counterfactual

$$\text{minimize } \mathbb{E}_{x \sim D_{np}} (f(x + \delta) - 1)^2$$
Training the adversarial model

1. First stage:
   - Small perturbations
   - Counterfactuals under perturbations
   - Accuracy
   - Passes the perturbations and model weights to the second stage

2. Second stage:
   - Fairness
   - Unfairness
   - Accuracy
Training the adversarial model

1. First stage:

\[ \delta := \arg \min_{\delta} \min_{\theta} \mathcal{L}(\theta, D) + \mathbb{E}_{x \sim \mathcal{D}_{np}} (f(x + \delta) - 1)^2 + \mathbb{E}_{x \sim \mathcal{D}_{np}} d(x, x + \delta) \]
Training the adversarial model

1. First stage:

\[
\delta := \arg \min_{\delta} \min_{\theta} \mathcal{L}(\theta, D) + \mathbb{E}_{x \sim D^{neg}} (f(x + \delta) - 1)^2 + \mathbb{E}_{x \sim D^{neg}} d(x, x + \delta)
\]

Classification loss
Training the adversarial model

1. First stage:

\[ \delta := \arg \min_{\delta} \min_{\theta} \mathcal{L}(\theta, \mathcal{D}) + \mathbb{E}_{x \sim \mathcal{D}_{np}} (f(x + \delta) - 1)^2 + \mathbb{E}_{x \sim \mathcal{D}_{np}} d(x, x + \delta) \]

Perturbed input to be a counterfactual
Training the adversarial model

1. First stage:

\[
\delta := \arg \min_{\delta} \min_{\theta} L(\theta, D) + \mathbb{E}_{x \sim D_{np}^{neg}} (f(x + \delta) - 1)^2 + \mathbb{E}_{x \sim D_{np}^{neg}} d(x, x + \delta)
\]

Perturbation should be small
Training the adversarial model

1. First stage:

$$\delta := \arg \min_{\delta} \min_{\theta} \mathcal{L}(\theta, \mathcal{D}) + \mathbb{E}_{x \sim \mathcal{D}_{\text{np}}} (f(x + \delta) - 1)^2 + \mathbb{E}_{x \sim \mathcal{D}_{\text{np}}} d(x, x + \delta)$$

2. Second stage:

$$\theta := \arg \min_{\theta} \mathcal{L}(\theta, \mathcal{D}) + \mathbb{E}_{x \sim \mathcal{D}_{\text{np}}} [d(x, A_\theta(x + \delta))] + \left( \mathbb{E}_{x \sim \mathcal{D}_{\text{pr}}} [d(x, A_\theta(x))] - \mathbb{E}_{x \sim \mathcal{D}_{\text{np}}} [d(x, A_\theta(x))] \right)^2$$

s.t.  $$\mathbb{E}_{x \sim \mathcal{D}_{\text{np}}} [d(x, A_\theta(x + \delta))] < \mathbb{E}_{x \sim \mathcal{D}_{\text{pr}}} [d(x, A_\theta(x))]$$
Experiments

• Dataset
  ○ Used two datasets: “German Credit” and “Communities and Crimes”
  ○ Strong incentives to “game the system” in both datasets
Experiments

● Dataset
  ○ Used two datasets: “German Credit” and “Communities and Crimes”
  ○ Strong incentives to “game the system” in both datasets

● Manipulated Model
  ○ 4 layer feed-forward neural network
  ○ Tanh activation function
  ○ Adam optimizer and cross entropy loss
How do you expect the accuracy to be impacted in the manipulated model?
## Results

Impact of the manipulated model on accuracy:

<table>
<thead>
<tr>
<th></th>
<th>Comm. &amp; Crime</th>
<th>German Credit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc</td>
<td>$|\delta|_1$</td>
</tr>
<tr>
<td>Unmodified</td>
<td>81.2</td>
<td>-</td>
</tr>
<tr>
<td>Wachter et al.</td>
<td>80.9</td>
<td>0.80</td>
</tr>
<tr>
<td>Sparse Wachter</td>
<td>77.9</td>
<td>0.46</td>
</tr>
<tr>
<td>Prototypes</td>
<td>79.2</td>
<td>0.46</td>
</tr>
<tr>
<td>DiCE</td>
<td>81.1</td>
<td>1.73</td>
</tr>
</tbody>
</table>
Results

- Metrics
  - Effectiveness of manipulation
Results

- Metrics
  - Effectiveness of manipulation

Table 2: **Recourse Costs of Manipulated Models**: Counterfactual algorithms find similar cost recourses for both subgroups, however, give much lower cost recourse if $\delta$ is added before the search.

<table>
<thead>
<tr>
<th></th>
<th>Communities and Crime</th>
<th>German Credit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protected</td>
<td>35.68</td>
<td>54.16</td>
</tr>
<tr>
<td>Non-Protected</td>
<td>35.31</td>
<td>52.05</td>
</tr>
<tr>
<td>Disparity</td>
<td>0.37</td>
<td>2.12</td>
</tr>
<tr>
<td>Non-Protected+(\delta)</td>
<td>1.76</td>
<td>22.59</td>
</tr>
<tr>
<td><em>Cost reduction</em></td>
<td>20.1×</td>
<td>2.3×</td>
</tr>
</tbody>
</table>
Results

- Metrics
  - Effectiveness of manipulation
  - Outlier factor of counterfactuals: How realistic are the counterfactuals returned by the model?

\[
P(A(x)) = \frac{d(A(x), a_0)}{\min_{x \neq a_0 \in D_{pos} \cap \{\forall x \in D_{pos} | f(x) = 1\}} d(a_0, x)}
\]
Results

- **Metrics**
  - Effectiveness of manipulation
  - Outlier factor of counterfactuals: How realistic are the counterfactuals returned by the model?

\[
P(A(x)) = \frac{d(A(x), a_0)}{\min_{x \neq a_0 \in D_{pos} \cap \{ \forall x \in D_{pos} | f(x) = 1 \}} d(a_0, x)}
\]

The local outlier factor of the counterfactuals with respect to the positively classified data (Breunig et al. “LOF: identifying density-based local outliers”)

Will be >1 if the counterfactual is an outlier.
Results

- Metrics
  - Effectiveness of manipulation
  - Outlier factor of counterfactuals: How realistic are the counterfactuals returned by the model?
Conclusions

- The paper shows that counterfactual explanations can be manipulated.
- They train an adversarial model that produces seemingly fair recourses but is in fact biased towards the non-protected group.
- They show that the manipulations are effective and realistic.
Appendix
How to train the adversarial model if the counterfactual algorithm is black box?

Lemma 3.1 Assuming the counterfactual explanation $A_\theta(x)$ follows the form of the objective in Equation 1, $\frac{\partial}{\partial x_{cf}} G(x, A_\theta(x)) = 0$, and $m$ is the number of parameters in the model, we can write the derivative of counterfactual explanation $A$ with respect to model parameters $\theta$ as the Jacobian,

$$\frac{\partial}{\partial \theta} A_\theta(x) = - \left[ \frac{\partial^2 G(x, A_\theta(x))}{dx_{cf}^2} \right]^{-1} \cdot \left[ \frac{\partial}{\partial \theta_1} \frac{\partial}{\partial x_{cf}} G(x, A_\theta(x)) \cdots \frac{\partial}{\partial \theta_m} \frac{\partial}{\partial x_{cf}} G(x, A_\theta(x)) \right]$$