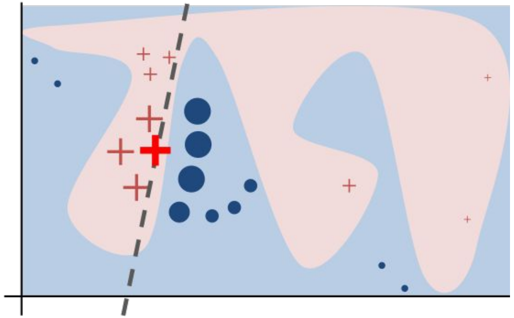
On the Privacy Risks of Algorithmic Recourse

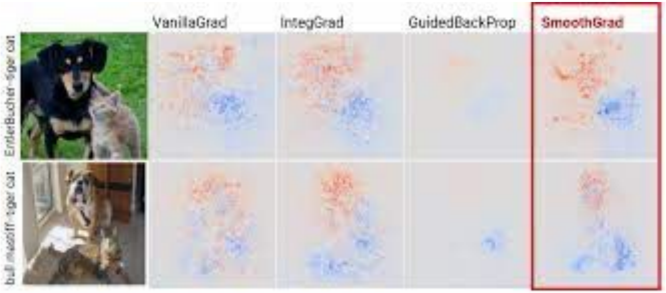
Pawelczyk, Lakkaraju, Neel

Presented by Christina Xiao, Lucas Monteiro Pas, Catherine Huang

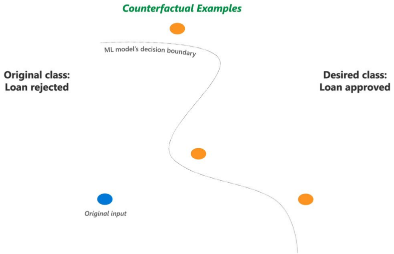
MOTIVATION



Lime

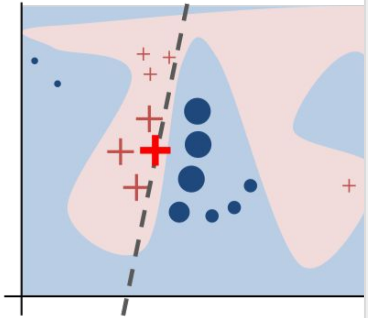


Gradient Based



Counterfactuals

MOTIVATION



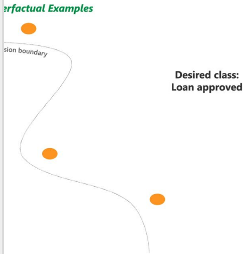
Lime

Can these methods leak:

1 - User’s sensitive information?

2 - Model’s weights?

3 - Training dataset?



Counterfactuals

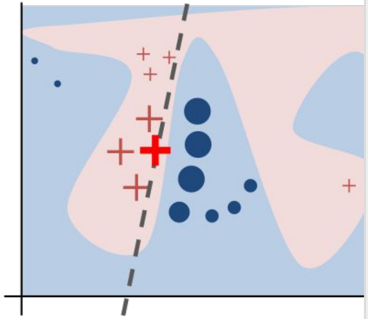
MOTIVATION

Can these methods leak:

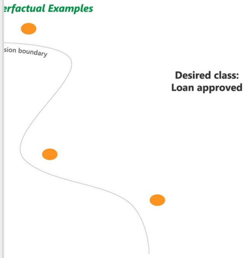
1 - User's sensitive information?

2 - Model's weights?

3 - Training dataset?



Lime



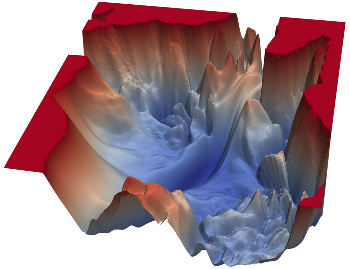
Counterfactuals

Previous works

Membership Inference.

Given access to an instance

+



Loss information

=

Instance is in training

Previous works

Membership Inferen

How to leverage XAI?

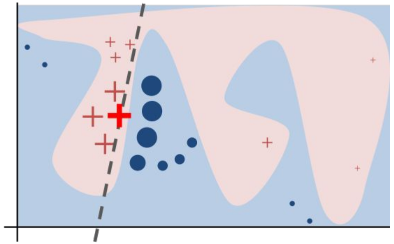
Given access to an ins ... nce is in training

Previous works

How can feature attribution impact membership inference? *Shokri, 2021.*

Given access to an instance

+



Feature attribution

=

Instance is in training

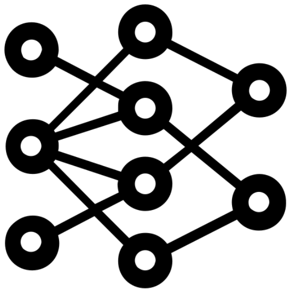
Reza Shokri, Martin Strobel, and Yair Zick. On the privacy risks of model explanations. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES), page 231–241, 2021.

Previous works

Model Extraction.

+

=



Given access to predictions

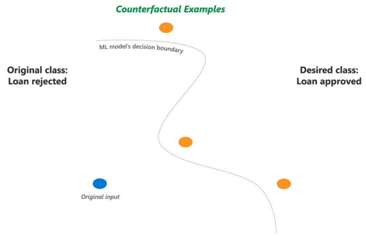
Reconstruct the model

Previous works

How can counterfactual explanations impact model extraction? *Aïvodji, 2020.*

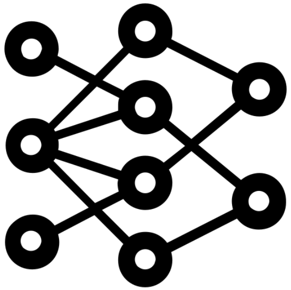
Given access to predictions

+



Counterfactual explanations

=



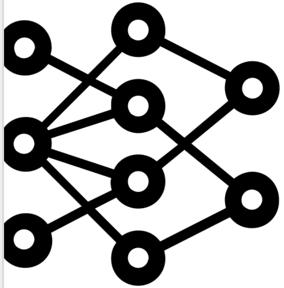
Reconstruct the model

Ulrich Aïvodji, Alexandre Bolot, and Sébastien Gambs. Model extraction from counterfactual explanations. arXiv preprint arXiv:2009.01884, 2020.

PREVIOUS WORKS

How can counterfac[...] extraction? *Aïvodji, 202*

The authors assume that the adversary can query the models multiple times!

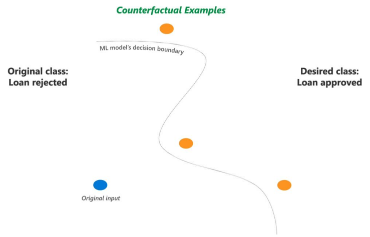


Given access to predic[...] construct the model

Ulrich Aïvodji, Alexandre Bolot, and Sébastien Gambs. Model extraction from counterfactual explanations. arXiv preprint arXiv:2009.01884, 2020.

Contributions

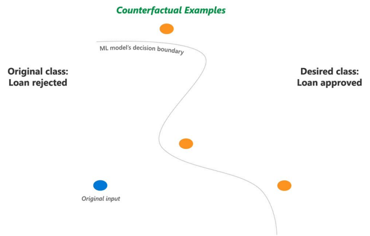
Given access to an instance



Counterfactual explanations

Contributions

Membership Inference.



Given access to an instance

Counterfactual explanations

Instance is in training

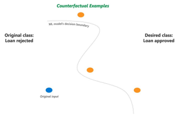
Contributions

Given access to an ins

The adversary can query the models a unique time!

nce is in training

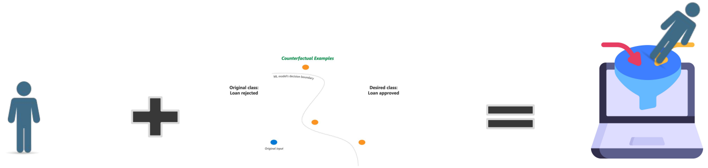
CONTRIBUTIONS



1. Define a **new class of attacks** called counterfactual distance-based attacks

2. Provide two examples of attacks in the class

CONTRIBUTIONS



c (“Instance”, “Counterfactual”)

=

Instance is in training

1. Define a **new class of attacks** called counterfactual distance-based attacks

2. Provide two examples of attacks in the class

Preliminaries: Algorithmic Recourse

$$x' = \arg\min_{x' \in \mathcal{A}^p} \ell(f_{\boldsymbol{\theta}}(x'), 1) + \lambda \cdot c(x, x')$$

Wachter et al.

Preliminaries: MI Attacks

Preliminaries: MI Attacks

Thresholding on Loss (Yeom et al.)

loss of model with params θ on instance $z = (x, y)$

threshold

$$M_{\text{Loss}}(x) = \begin{cases} \texttt{MEMBER} & \text{if } \ell(\boldsymbol{\theta}, z) \leq \tau_L \\ \texttt{NON-MEMBER} & \text{if } \ell(\boldsymbol{\theta}, z) > \tau_L . \end{cases}$$

very powerful and simple, but only feasible when can access instance's y; model's ℓ, θ; underlying data distribution $\mathscr{D}$ (to practically get τ)

Preliminaries: MI Attacks

Thresholding on Loss (Yeom et al.)

loss of model with params θ
on instance $z = (x, y)$

threshold

$$M_{\text{Loss}}(x) = \begin{cases} \texttt{MEMBER} & \text{if } \ell(\boldsymbol{\theta}, z) \leq \tau_L \\ \texttt{NON-MEMBER} & \text{if } \ell(\boldsymbol{\theta}, z) > \tau_L. \end{cases}$$

very powerful and simple, but only feasible when can access instance's *y*; model's ℓ, θ; underlying data distribution $\mathscr{D}$ (to practically get τ)

Loss Likelihood Ratio Attack (Carlini et al.)

Given: sample access to underlying data distribution $\mathscr{D}$

1. Adversary trains shadow models
2. Computes confidence in each model f_θ when z **in/out** train set
3. Fits normal distributions to these in/out confidences
4. Computes approximate likelihood ratio Λ
5. Predicts MEMBER when Λ > τ

Setting: Recourse-based MI Game

owner $\mathscr{O}$ and **adversary** $\mathscr{A}$

Setting: Recourse-based MI Game

owner $\mathcal{O}$:

1. Draws training set D_t from underlying data distribution $\mathscr{D}$
2. Trains model f_θ
3. Labels every point z in D_t with binary label $f_\theta(z)$
4. Flips coin to determine where to sample x from
   a. If heads, conditional distribution $\mathscr{D} \mid f_\theta(z) = 0$
   b. If tails, subset of D_t with label 0
5. Generates recourse x' for x
6. Sends (x', x)

all data labelled 0 (unfavorable outcome), but combination of training data or not

Setting: Recourse-based MI Game

owner $\mathscr{O}$:

1. Draws training set D_t from underlying data distribution $\mathscr{D}$
2. Trains model f_θ
3. Labels every point z in D_t with binary label $f_\theta(z)$
4. Flips coin to determine where to sample x from
   a. If heads, conditional distribution $\mathscr{D} \mid f_\theta(z) = 0$
   b. If tails, subset of D_t with label 0
5. Generates recourse x' for x
6. Sends (x', x)

all data labelled 0 (unfavorable outcome), but combination of training data or not

adversary $\mathscr{A}$:

1. Can also query $\mathscr{D}$, knows implementation details of $\mathscr{O}$
2. Guesses whether x is MEMBER (in D_t) or NON-

Intuition

Loss-based attacks are good at determining MEMBER, because model typically overfits to training points

↓

This may be because, during training, decision boundary is pushed away from training points (Shroki et al.)

↓

Points in training set should be further from boundary than points in test set

↓

Counterfactual distance-based attacks!

Attack 1: Thresholding on CFD

counterfactual distance between *x*, *x'*, from whichever recourse method

threshold

$$M_{\text{Distance}}(x) = \begin{cases} \textit{MEMBER} & \text{if } c(x, x') \geq \tau_D \\ \textit{NON-MEMBER} & \text{if } c(x, x') < \tau_D. \end{cases}$$

(recall: $M_{\text{Loss}}(x) = \begin{cases} \textit{MEMBER} & \text{if } \ell(\boldsymbol{\theta}, z) \leq \tau_L \\ \textit{NON-MEMBER} & \text{if } \ell(\boldsymbol{\theta}, z) > \tau_L. \end{cases}$)

assume that $\mathscr{A}$ knows a priori optimal threshold τ_α that maximizes TPR given FPR α; in practice, will plot TPR v. FPR over all τ_D

ATTACK 2: CFD LRT

again, similar to preliminaries, but more adjustments

Algorithm 1 One-sided Distance-based Likelihood Ratio Test (`CFD LRT`)

1: **Inputs:** point (x, y), recourse output s = $\texttt{GetRecourse}(x, f_{\boldsymbol{\theta}}), \mathcal{D}$; FP-Rate: α, # Shadow Models: N, $\mathcal{T}$ = $\texttt{TrainClassifier}(\cdot)$

2: teststats = []

3: **Compute**: $t_0 = T(s) = c(x, x')$ ← compute CFD on input

Attack 2: CFD LRT

again, similar to preliminaries, but more adjustments

Algorithm 1 One-sided Distance-based Likelihood Ratio Test (`CFD LRT`)

Inputs: point (x, y), recourse output $s = \texttt{GetRecourse}(x, f_{\boldsymbol{\theta}}), \mathcal{D}$; FP-Rate: α, # Shadow Models: N, $\mathcal{T} = \texttt{TrainClassifier}(\cdot)$
teststats = []
Compute: $t_0 = T(s) = c(x, x')$
for $i = 1 : N$ **do**
  Sample $\mathcal{D}_t^{(i)} \sim \mathcal{D}$
  $f_{\boldsymbol{\theta}^{(i)}} = \texttt{TrainClassifier}(\mathcal{D}^{(i)})$
  $s^{(i)} = \texttt{GetRecourse}(x, f_{\boldsymbol{\theta}^{(i)}})$
  teststats $\leftarrow T(s^i) = c(x, x'^{(i)})$
end for

compute CFD on input

train shadow models (only need to do once!) and recourses, collect their CFDs on input

ATTACK 2: CFD LRT

again, similar to preliminaries, but more adjustments

Algorithm 1 One-sided Distance-based Likelihood Ratio Test (`CFD LRT`)

Inputs: point (x, y), recourse output $s = \texttt{GetRecourse}(x, f_{\boldsymbol{\theta}}), \mathcal{D}$; FP-Rate: α, # Shadow Models: N, $\mathcal{T} = \texttt{TrainClassifier}(\cdot)$
teststats = []
Compute: $t_0 = T(s) = c(x, x')$

compute CFD on input

for $i = 1 : N$ **do**
    Sample $\mathcal{D}_t^{(i)} \sim \mathcal{D}$
    $f_{\boldsymbol{\theta}^{(i)}} = \texttt{TrainClassifier}(\mathcal{D}^{(i)})$
    $s^{(i)} = \texttt{GetRecourse}(x, f_{\boldsymbol{\theta}^{(i)}})$
    teststats $\leftarrow T(s^i) = c(x, x'^{(i)})$
end for

train shadow models (only need to do once!) and recourses, collect their CFDs on input

$\hat{\mu}_{\text{MLE}} = \frac{1}{N} \sum_{i=1}^{N} \left(\log c(x, \mathbf{x}'^{(i)}) \right))$
$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{\mu}_{\text{MLE}} - \log \left(c(x, \mathbf{x}'^{(i)}) \right) \right)^2$

estimate params of normal distribution

ATTACK 2: CFD LRT

again, similar to preliminaries, but more adjustments

Algorithm 1 One-sided Distance-based Likelihood Ratio Test (`CFD LRT`)

Inputs: point (x, y), recourse output $s = \texttt{GetRecourse}(x, f_{\boldsymbol{\theta}}), \mathcal{D}$; FP-Rate: α, # Shadow Models: N, $\mathcal{T} = \texttt{TrainClassifier}(\cdot)$
teststats = []
Compute: $t_0 = T(s) = c(x, x')$ — compute CFD on input
for $i = 1 : N$ **do**
    Sample $\mathcal{D}_t^{(i)} \sim \mathcal{D}$
    $f_{\boldsymbol{\theta}^{(i)}} = \texttt{TrainClassifier}(\mathcal{D}^{(i)})$
    $s^{(i)} = \texttt{GetRecourse}(x, f_{\boldsymbol{\theta}^{(i)}})$
    teststats $\leftarrow T(s^i) = c(x, x'^{(i)})$
end for

train shadow models (only need to do once!) and recourses, collect their CFDs on input

$\hat{\mu}_{\text{MLE}} = \frac{1}{N}\sum_{i=1}^{N}\left(\log c(x, \mathbf{x}'^{(i)})\right))$
$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\mu}_{\text{MLE}} - \log\left(c(x, \mathbf{x}'^{(i)})\right)\right)^2$

estimate params of normal distribution

if $t_0 > z_{1-\alpha}$ **then** ▷ $z_{1-\alpha}$ is the $1-\alpha$-quantile of $Z \sim \mathcal{LN}(\hat{\mu}_{\text{MLE}}, \hat{\sigma}^2_{\text{MLE}})$
    Output: G = `NON-MEMBER`
else
    Output: G = `MEMBER`
end if

threshold τ

Is privacy leakage through recourses inevitable?

Is privacy leakage through recourses inevitable?

Privacy community thinks DP in training can bound the success of any adversary.

Bounding success of $\mathscr{A}$ with DP

Theorem 1. *Let $\mathcal{T} : (\mathcal{X} \times \mathcal{Y})^n \to \Theta$ denote the training algorithm, draw $D_t \sim \mathcal{D}^n$ and and $\mathcal{A}$ be an arbitrary adversary that receives $z = (x, y), s \sim \mathcal{R}(f_\theta, x, D_t)$ from the recourse inference game, and produces a guess $G \in \{MEMBER, NON\text{-}MEMBER\}$. Then, if $\mathcal{R}$ is $(\epsilon, 0)$-differentially private, we have for all $\mathcal{A}$:*

$$BA_{\mathcal{A}} \leq \frac{1}{2} + \frac{1 - e^{-\epsilon}}{2}.$$

Implications:

- Using DP in training, we can strongly bound the adversary's **balanced accuracy** success ((TPR + TNR) / 2) — not just excess accuracy broadly
- For a small FPR α, TPR of $\mathscr{A}$ is also close to α

BOUNDING SUCCESS OF $\mathscr{A}$ WITH DP

Theorem 1. *Let $\mathcal{T} : (\mathcal{X} \times \mathcal{Y})^n \to \Theta$ denote the training algorithm, draw $D_t \sim \mathcal{D}^n$ and and $\mathcal{A}$ be an arbitrary adversary that receives $z = (x, y), s \sim \mathcal{R}(f_\theta, x, D_t)$ from the recourse inference game, and produces a guess $G \in \{MEMBER, NON\text{-}MEMBER\}$. Then, if $\mathcal{R}$ is $(\epsilon, 0)$-differentially private, we have for all $\mathcal{A}$:*

$$BA_{\mathcal{A}} \leq \frac{1}{2} + \frac{1 - e^{-\epsilon}}{2}.$$

Implications:

- Using DP in training, we can strongly bound the adversary's **balanced accuracy** success ((TPR + TNR) / 2) – not just excess accuracy broadly
- For a small FPR α, TPR of $\mathscr{A}$ is also close to α

Proof:

- Appendix A; mostly applies definitions and expands integrals

Bounding success of $\mathscr{A}$ with DP

Theorem 1. *Let $\mathcal{T} : (\mathcal{X} \times \mathcal{Y})^n \to \Theta$ denote the training algorithm, draw $D_t \sim \mathcal{D}^n$ and and $\mathcal{A}$ be an arbitrary adversary that receives $z = (x, y), s \sim \mathcal{R}(f_\theta, x, D_t)$ from the recourse inference game, and produces a guess $G \in \{MEMBER, NON\text{-}MEMBER\}$. Then, if $\mathcal{R}$ is $(\epsilon, 0)$-differentially private, we have for all $\mathcal{A}$:*

$$BA_{\mathcal{A}} \leq \frac{1}{2} + \frac{1 - e^{-\epsilon}}{2}.$$

Implications:

- Using DP in training, we can strongly bound the adversary's **balanced accuracy** success ((TPR + TNR) / 2) – not just excess accuracy broadly
- For a small FPR α, TPR of $\mathscr{A}$ is also close to α

Proof:

- Appendix A; mostly applies definitions and expands integrals

However:

- DP is not a silver bullet! Training with DP causes significant drop in accuracy

Experimental Evaluation: Setup

Datasets

1. Adult (A)
   - Label: whether income > 50,000
2. Home Equity Line of Credit (H)
   - Label: whether individuals will repay HELOC
3. Diabetes (D)
   - Label: whether patient will be readmitted within next 30 days
4. Synthetic
   - Label: comes from Gaussian samples

Recourse Algorithms

1. SCFE (Wachter et al.)
   - Gradient-based objective
2. Growing Spheres (GS)
   - Random search in the input space
3. CCHVAE
   - Trains a variational autoencoder (VAE)
   - VAE searches in a lower-dimensional latent space

Experimental Evaluation: Procedure

Subsampling

- Subsampling 10,000 data points
- 5,000 points: owner trains private model
- 5,000 points: adversary trains shadow model, for CFD likelihood ratio (LRT) attack

Training

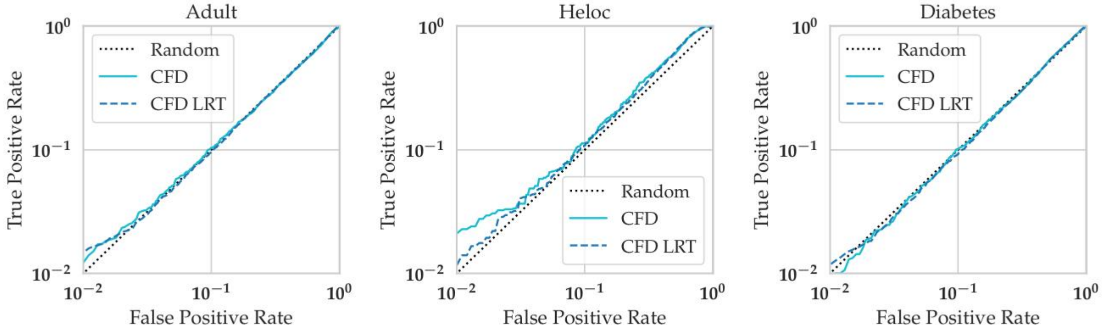
- Fully connected classifier neural network
- 1 hidden layer: 1000 nodes, ReLu activation
- ADAM optimizer (lr=0.0001)
- 250 epochs

Evaluation

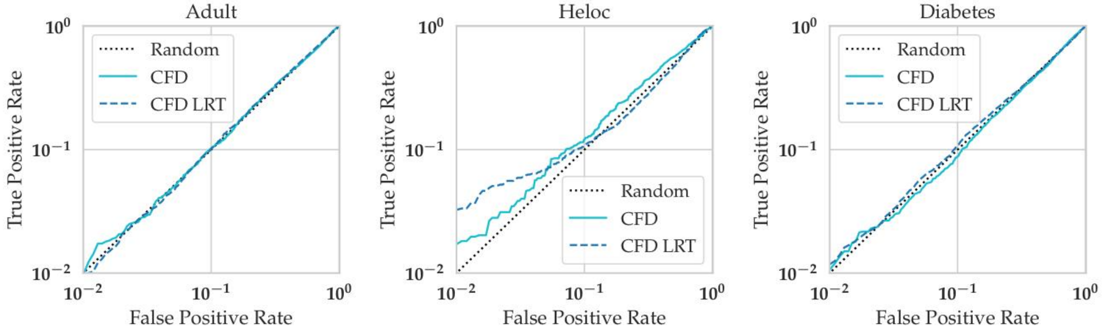
- Balanced accuracy
- Receiver operating characteristic AUC score
- Log-scale ROC curves
- True positive rates (TPRs) at low false positive rates (FPRs)

ATTACK EFFICIENCY

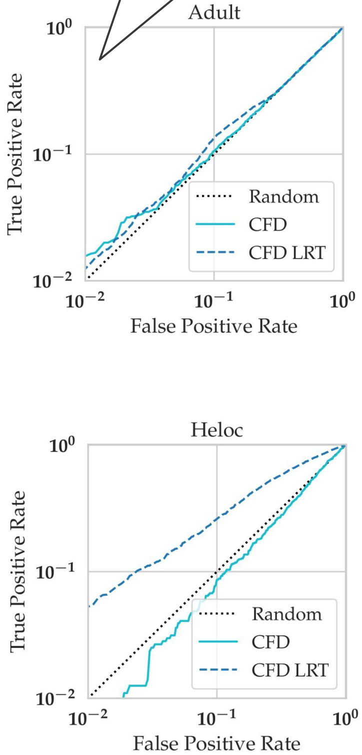
log-log transformation



(a) SCFE



(b) GS



(c) CCHVAE

ATTACK EFFICIENCY: TAKEAWAYS

Small upwards deviations from the diagonal at low FPRs (i.e. 0.01)

↓

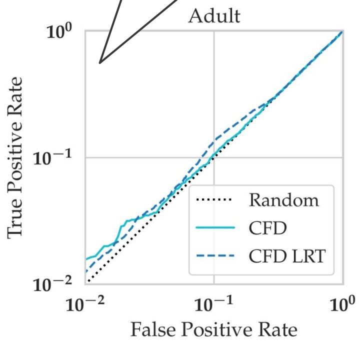
A small fraction of points are accurately identified as members

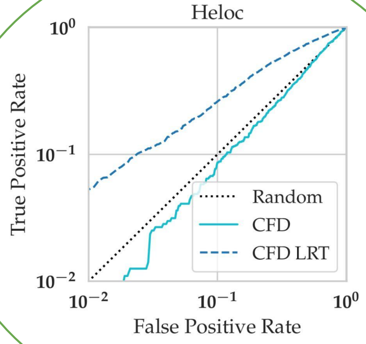
↓

Successful MI attack

- Both methods (CFD, CFD LRT) often outperform the random baseline across all metrics
- CFD LRT generally outperforms CFD

log-log transformation

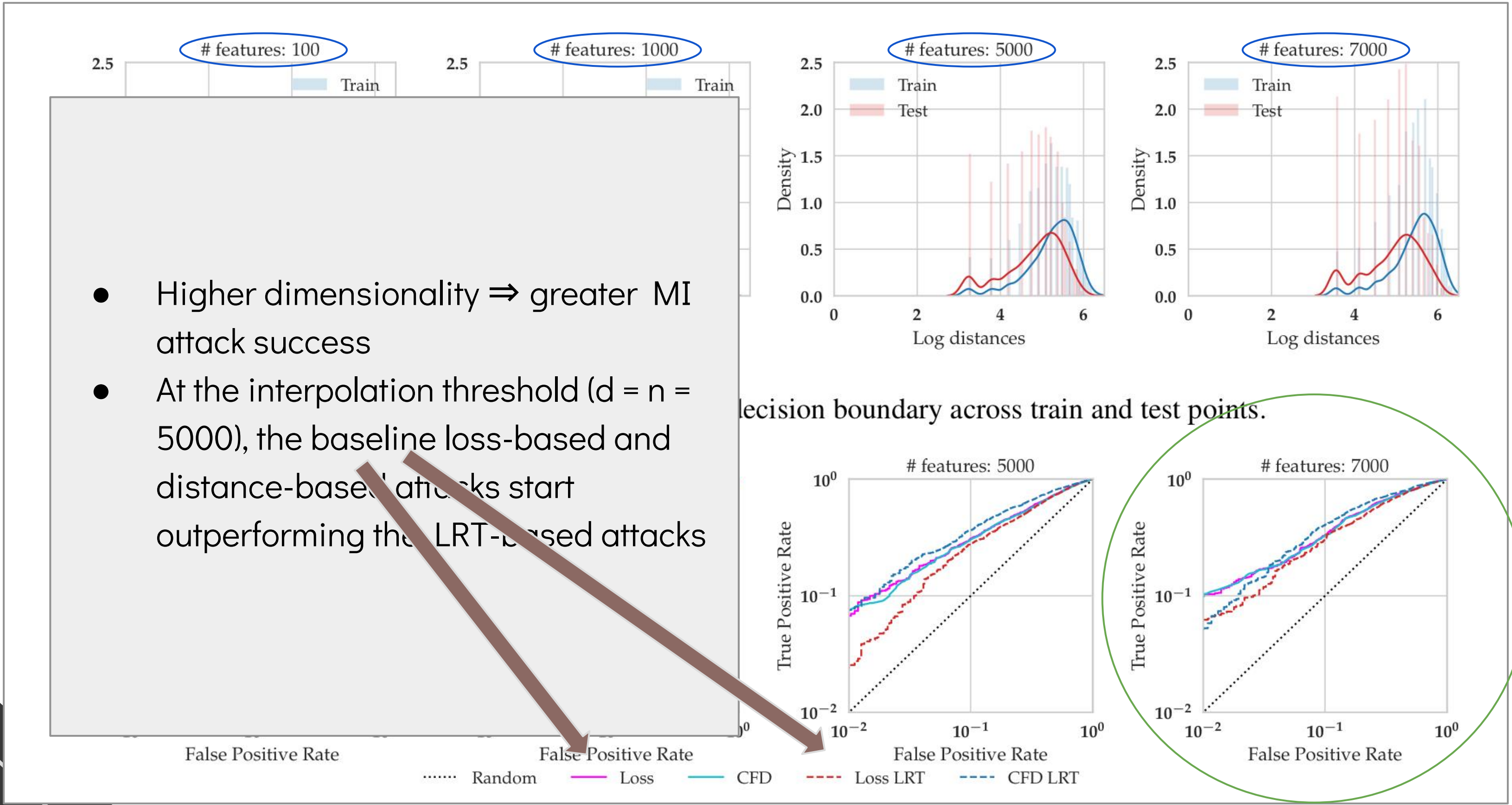




(c) CCHVAE

EFFECTS OF # FEATURES

- Higher dimensionality ⇒ greater MI attack success
- At the interpolation threshold (d = n = 5000), the baseline loss-based and distance-based attacks start outperforming the LRT-based attacks



Effects of Model Architecture

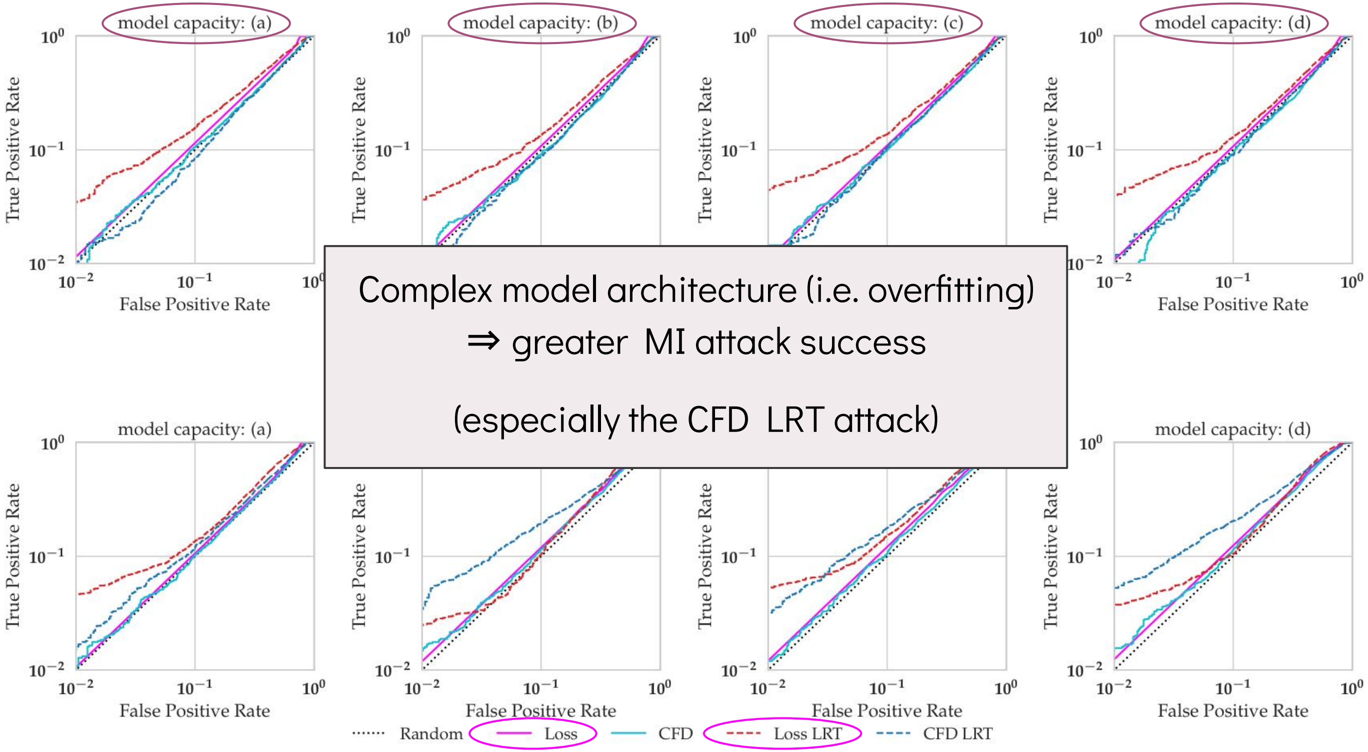
Models

(a) 2 layers, 1000 hidden notes

(b) 3 layers, 100 hidden nodes

(c) 3 layers, 333 hidden nodes

(d) 3 layers, 1000 hidden nodes



(b) # features = 150

Complex model architecture (i.e. overfitting) ⇒ greater MI attack success

(especially the CFD LRT attack)

When are CFD-Based attacks most successful?

- When data dimensionality is high (# of features)
- When underlying model overfits to training data (model architecture)
- Combination of the above ⇒ increased vulnerability of recourses to CFD-based MI attacks

Conclusion

Novel Attacks

- **Idea:** we can leverage recourses to infer *private* training data membership information

- **Contribution:** MI attacks that leverage *counterfactual distances* output by recourse methods

Evidence of privacy Leakage

- **Implications:** privacy leakage is a risk of recourse algorithms; explainability-privacy tradeoff

- **Relevance:** proposed MI attacks are effective in diverse domains (lending, healthcare, law)

Limitations

- CFD is only a heuristic—an approximation of the distance of data point x to the model boundary
  - Recall: CFD = distance from x to its recourse
- Attacks operate under assumption that adversary can only query recourse algorithm once
- Must assume adversary knows optimal threshold that maximizes TPR given a fixed FPR
- This paper highlights a problem (privacy leakage), but not yet a solution
- Assessed on binary classification tasks only
  - Generalizability of results (broadly)

Future Work

- **Generalization:** whether recourse exposes us to other forms of privacy leakage
    - Can algorithmic recourse lead to successful reconstruction attacks? How about attacks on sensitive summary statistics of the training data (or anything else about the data distribution)?
- **Generalization:** which other XAI mechanisms involve privacy violations?
- **Solutions to protect privacy:** whether we can train models that provide recourse while mitigating privacy risks
    - How do we construct faithful model explanations that also do not leak too much information about the underlying training data? What is the privacy-utility trade-off of such models?

Discussion Questions

1. Given the explainability-privacy tradeoff highlighted in this paper, what is the role of each of the following in determining explainability and privacy benchmarks when training a model?
    a. ML practitioners (model builders and model breakers)
    b. End users (model consumers)
2. Both privacy and explainability can cultivate user *trust* in an ML model, and the lack thereof of any of these can break this trust. Both pillars are crucial but cannot fully coexist (this is the crux of our paper)—in what situations would you care about one pillar over the other?
3. Besides recourse, what other XAI mechanisms do you think might lead to privacy violations?
4. Is it even possible to have private explanations? Is this even worth going for?