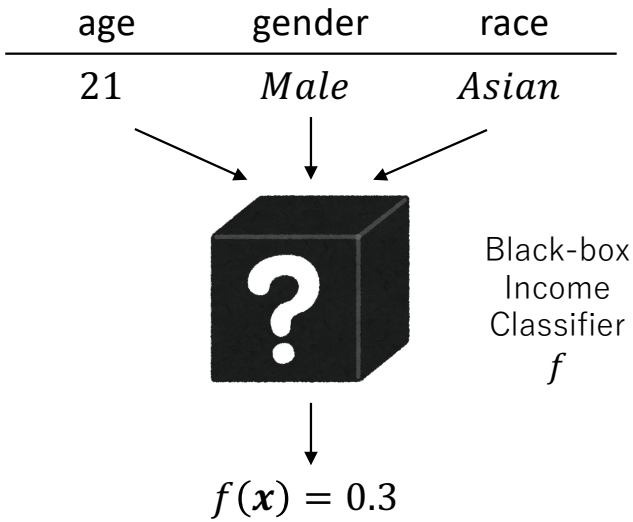
差分プライバシーを保証したモデル説明DPGD-Explainに対するレコード再構築リスクの実験評価

當麻 僚太郎, 菊池 浩明
明治大学

背景

- 機械学習モデルはブラックボックス
  - ニューラルネットワーク
  - ランダムフォレスト
  - SVM

- XAIによって保証したいこと
  - 透明性
  - 公平性
  - モデルのふるまい

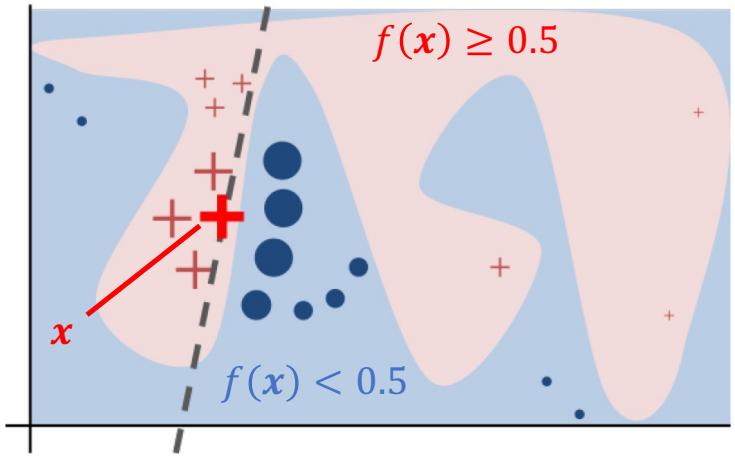


LIME [Ribeiro et al. 2016]

- モデル f のふるまいを特定の入力 $\boldsymbol{x}$ の周りで近似する
- 各特徴量が出力に与えた影響を評価

Approximation

$$g(\boldsymbol{x}) = w_1 x_1 + w_2 x_2 + w_3 x_3 + w_0$$



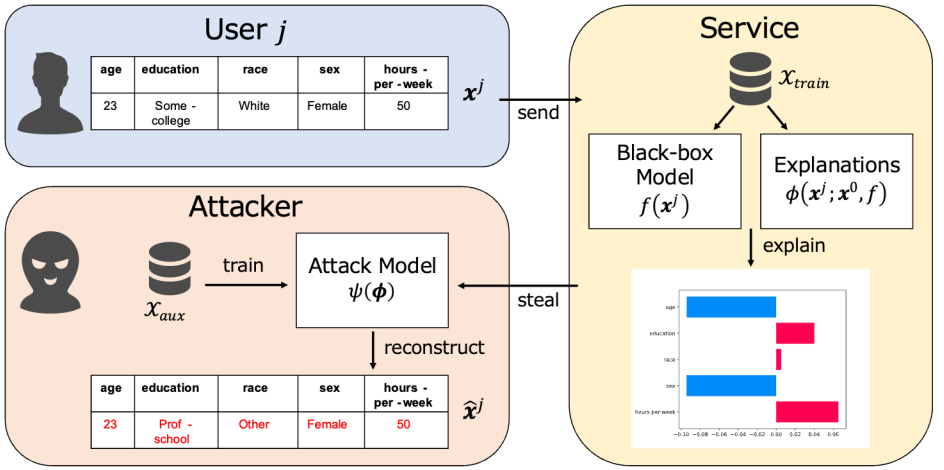
	x_1	x_2	x_3	$f(\boldsymbol{x})$
$\boldsymbol{x}$	1.5	True	A	0.8
$\boldsymbol{x}^1$	-0.4	False	B	0.6
$\boldsymbol{x}^2$	0.1	False	A	0.3
$\boldsymbol{x}^3$	0.8	True	C	0.9
$\boldsymbol{x}^4$	-1.1	True	A	0.2

⇩

	w_1	w_2	w_3
$\boldsymbol{w}$	0.23	0.13	-0.30

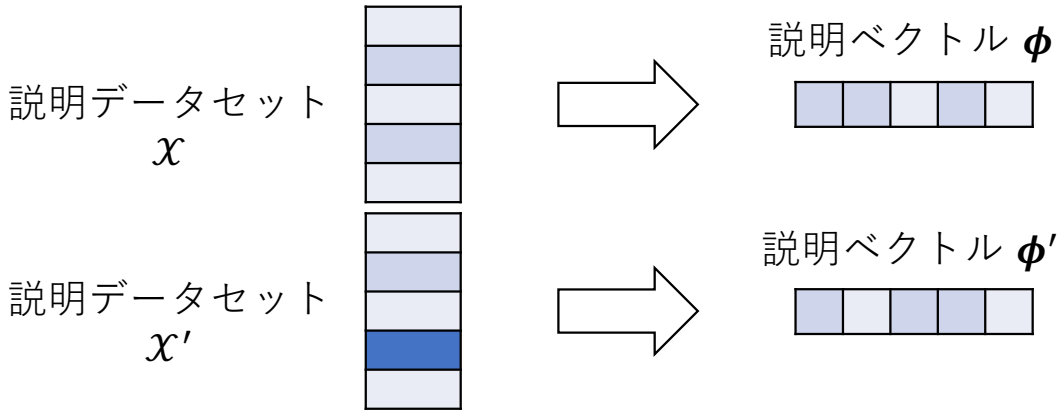
LIMEの問題点： モデル説明からのプライバシーリスク

- Shapley値による説明から元の入力ベクトルを再構築する攻撃
  - Feature Inference Attack on Shapley Values [Luo et al. 2022]
- LIMEも同様の攻撃に対して脆弱
  - Combination of AI Models and XAI Metrics Vulnerable to Record Reconstruction Risk [Toma et al. 2024]



DPGD-Explain [Patel et al. 2022]

- $(\epsilon, \gamma\delta)$-差分プライバシーを保証する逐次学習アルゴリズム
- プライバシー費用を最小化する動的計画法アルゴリズム
- 収束に必要なイテレーション数を求める最適化
- 事後解析（post-hoc）なモデル説明



$$\Pr[\boldsymbol{\phi} \in S] \leq e^{\epsilon} \Pr[\boldsymbol{\phi}' \in S] + \delta$$

ただし，$S \subseteq \mathbb{R}^n$ は説明ベクトル集合

Patel, N., Shokri, R., Zick, Y.: Model Explanations with Differential Privacy, ACM FAccT'22.

DPGD-Explainの近似モデル

- 説明対象のブラックボックスモデルを f，入力ベクトルを $\boldsymbol{z} = (z_1, \ldots, z_n)$，$\boldsymbol{z}$ に対するモデル説明を $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)$ とする

- 説明モデル $g(\boldsymbol{x}, \boldsymbol{z}) = \boldsymbol{\phi}^T(\boldsymbol{x} - \boldsymbol{z})$ の損失は重み付きの最小二乗法

$$\mathcal{L}(g, f, \boldsymbol{z}, \mathcal{X}) = \sum_{\boldsymbol{x} \in \mathcal{X}} \alpha(\|\boldsymbol{x} - \boldsymbol{z}\|)\left(\boldsymbol{\phi}^T(\boldsymbol{x} - \boldsymbol{z}) - f(\boldsymbol{x})\right)^2$$

  - $\mathcal{X}$ は m 行 n 列の説明データセット
  - 重み関数 $\alpha: \mathbb{R}_+ \to \mathbb{R}_+$ は距離 $\|\boldsymbol{x} - \boldsymbol{z}\|$ が大きくなるほど小さい値を取る
    - 例） $\alpha(\|\boldsymbol{x} - \boldsymbol{z}\|) = \exp(-\|\boldsymbol{x} - \boldsymbol{z}\|^2)$

差分プライバシーを見たす説明生成

- 勾配降下法でモデル説明 $\boldsymbol{\phi}$ を学習

- LIMEとの主な違い
  - DPの保証
  - イテレーティブな説明生成

- DP-SGDとの主な違い
  - Sensitivityの理論的保証
    →勾配のクリッピングなし

Algorithm 1: Interactive DP Model Explanation

Input: POI $\vec{z} \in \mathbb{R}^n$, explanation dataset $\mathcal{X}$, DP parameters $\epsilon > 0, \delta > 0$, weight function $\alpha(\|\vec{x} - \vec{z}\|) \in \mathcal{F}(1, \vec{z})$, learning rate function $\eta(t) \in \mathbb{R}_+$, and the number of GD iterations T.

Output: ϕ^{Priv}

Set the parameter for the Gaussian mechanism $\sigma \leftarrow \frac{1}{m\epsilon}\sqrt{16T \log\left(\mathrm{e} + \frac{\sqrt{T}\epsilon}{\delta}\right) \log\frac{T}{\delta}}$;
Initialize ϕ with an arbitrary vector in $\mathcal{C}_{2,1}$;
return `DPGD-Explain`(ϕ, σ, T);
Procedure `DPGD-Explain`(ϕ, σ, T)
    $\phi^{\{1\}} \leftarrow \phi$
    for $t = 1, \ldots, T-1$ **do**
        $\xi_t \leftarrow \left(\phi^{\{t\}} - \eta(t)\left[\nabla\mathcal{L}(\phi, \mathcal{X}) + \mathcal{N}(0, \sigma^2\mathrm{I})\right]\right)$
        $\phi^{\{t+1\}} \leftarrow \arg\min_{\phi \in \mathcal{C}_{2,1}} \|\phi - \xi_t\|$
    end
    return $\phi^{\{T\}}$

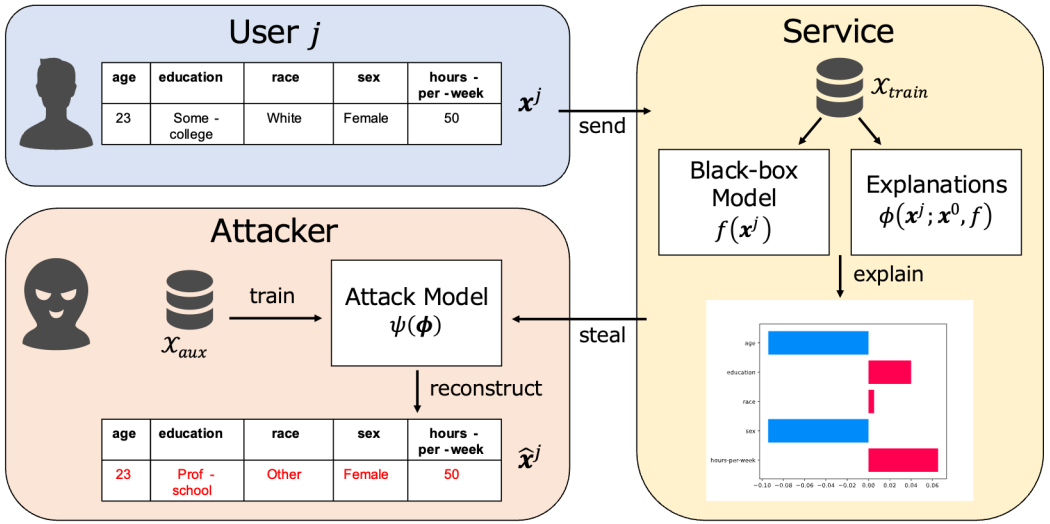
本研究の新規性

- LIMEとDPGD-Explainを独自実装し，Luoらのレコード再構築攻撃を適用してPatelらの主張を確かめる

- Research Question
  - DPGD-Explainは本当にXAIを安全にするか？

Threat Model

- あるユーザのプライベートな入力ベクトル $\boldsymbol{z} = (1.5, \text{True}, A)$ に対して，モデル説明 $\boldsymbol{\phi} = (0.23, 0.13, -0.3)$ が得られたとする
- 攻撃者は補助データセット $\mathcal{X}_{aux}$ を用いて訓練した攻撃モデル ψ を用いて，元の入力ベクトルの予測 $\hat{\boldsymbol{x}} = \psi(\boldsymbol{\phi})$ を得る



実験設定

- 2種類のオープンデータ
  - AdultデータセットはOne-Hot Encodingして14次元→119次元
  - IMDBデータセットは出現頻度上位500語がそれぞれ存在するかどうかのバイナリ（500次元）
    - 例） amazing, bad, characterに対して”There is amazing character”は $(1, 0, 1)$ となる

Dataset	Data	Records	Classes
UCI Adult	Tabular	48,842	$\leq 50\text{K}/< 50\text{K}$
IMDB Movie Reviews	Text	50,000	pos/neg

評価指標

- 安全性評価
  - m 行 n 列の入力データセット X と再構築したデータセット $\hat{X}$
  - レコード再構築攻撃の攻撃成功率

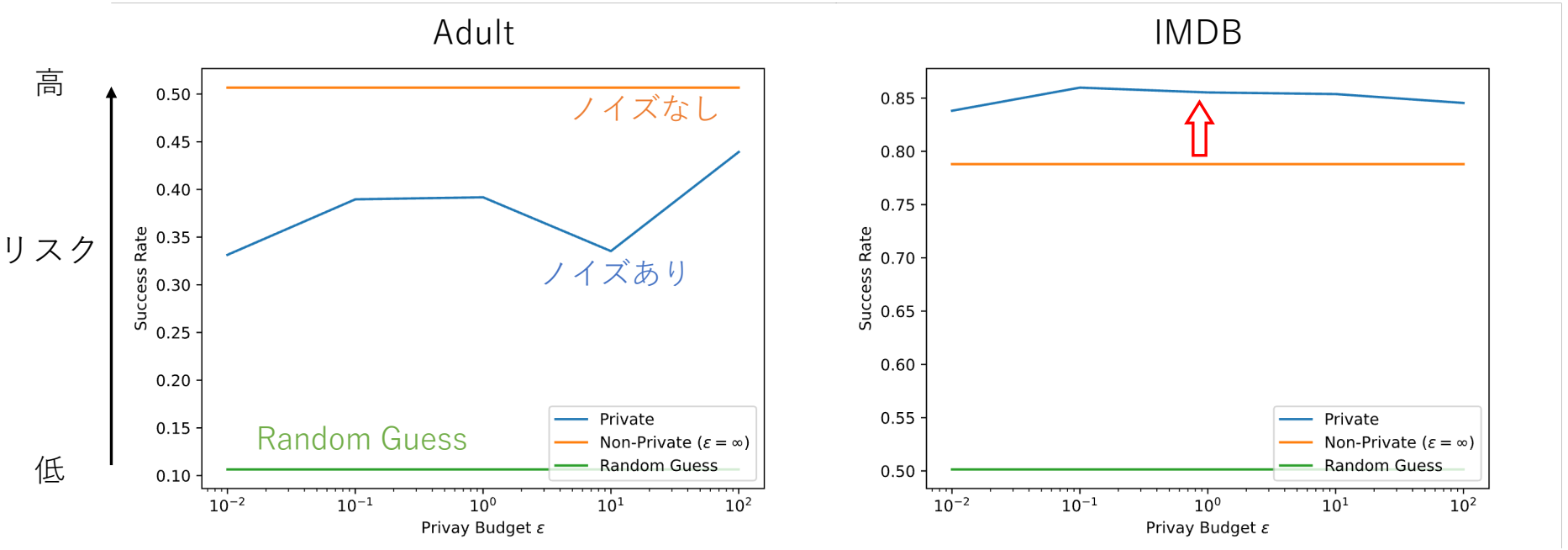
$$SR(X, \hat{X}) = \frac{success(X, \hat{X})}{mn}$$

- 有用性評価
  - ノイズなし（$\epsilon = \infty$）のDPGD-Explainによる説明ベクトルを $\boldsymbol{\phi}^*$ として有用性損失をMSEで評価

$$MSE(\boldsymbol{\phi}, \boldsymbol{\phi}^*) = \frac{1}{n}\left((\phi_1 - \phi_1^*)^2 + \cdots + (\phi_n - \phi_n^*)^2\right)$$

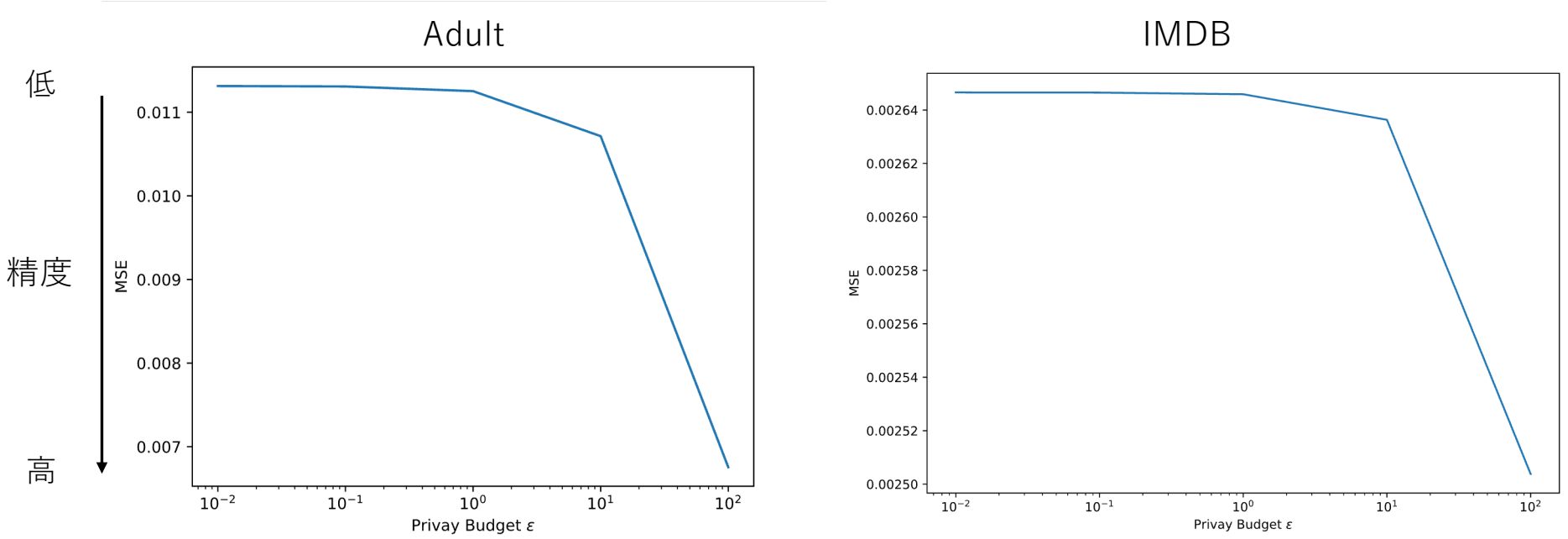
実験結果１： モデル説明の安全性

- Adultでは ϵ が上がるほどSRも上がった
  - ただし単調増加ではなかった
- IMDBではノイズありの方が攻撃成功率が大きい



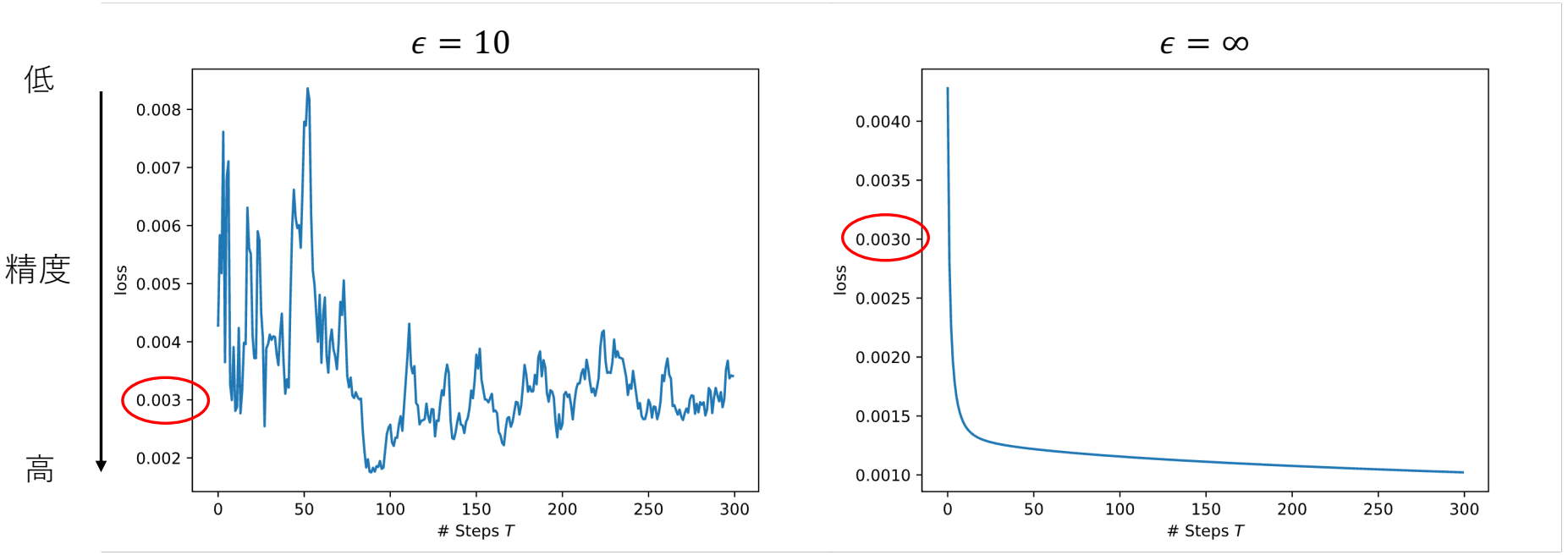
実験結果２： 説明の有用性

- $\epsilon = 0.01, 0.1, 1$ の場合はあまり差がない



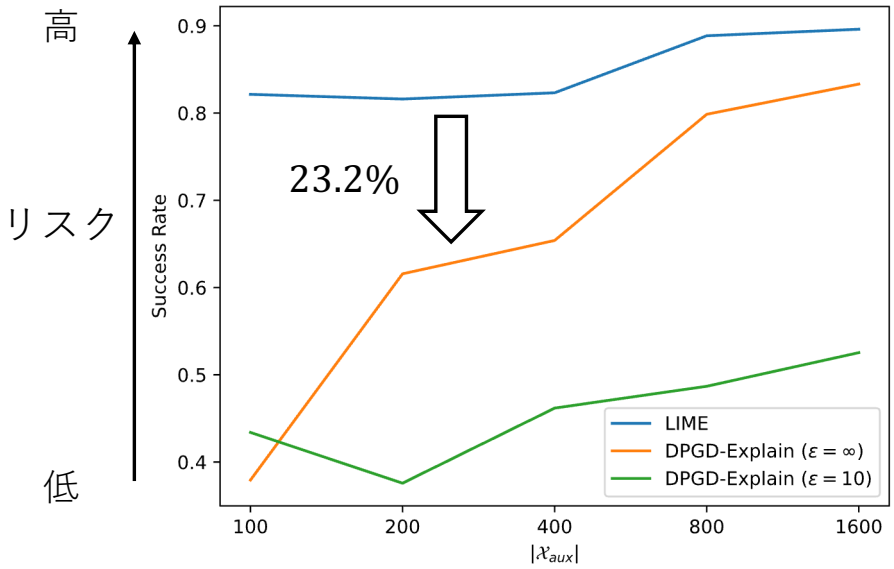
実験結果3： モデル説明値の収束性

- $\epsilon = 10, \infty$ における学習時の損失変化の例
- 学習が不安定になり損失自体も高い



実験結果4： LIMEとDPGD-Explainの比較

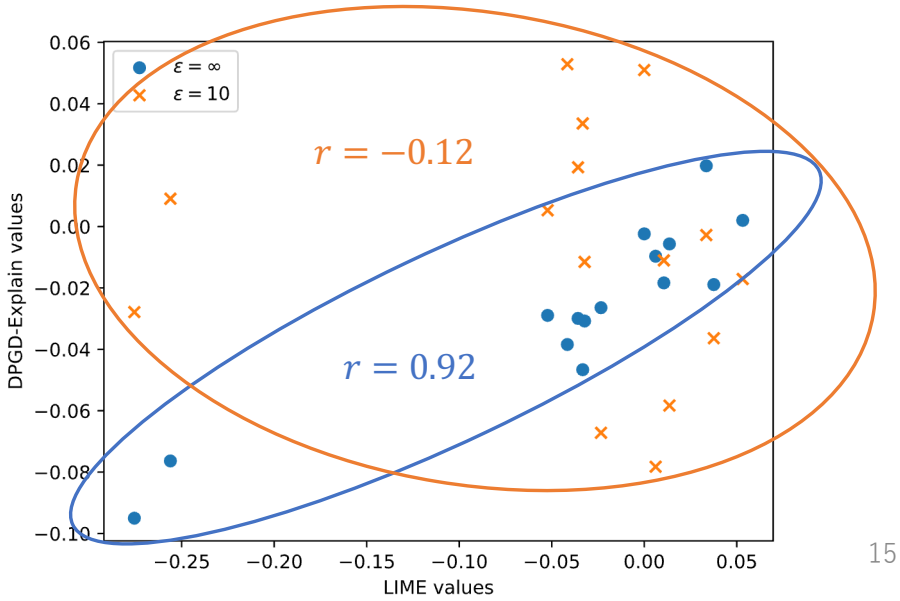
- ノイズなしでもDPGD-ExplainよりLIMEの方が脆弱
  - 平均相対誤差 23.2%
- LIMEとノイズなし（$\epsilon = \infty$）のDPGD-Explainは強い正の相関



LIME

DPGD-Explain ($\epsilon = \infty$)

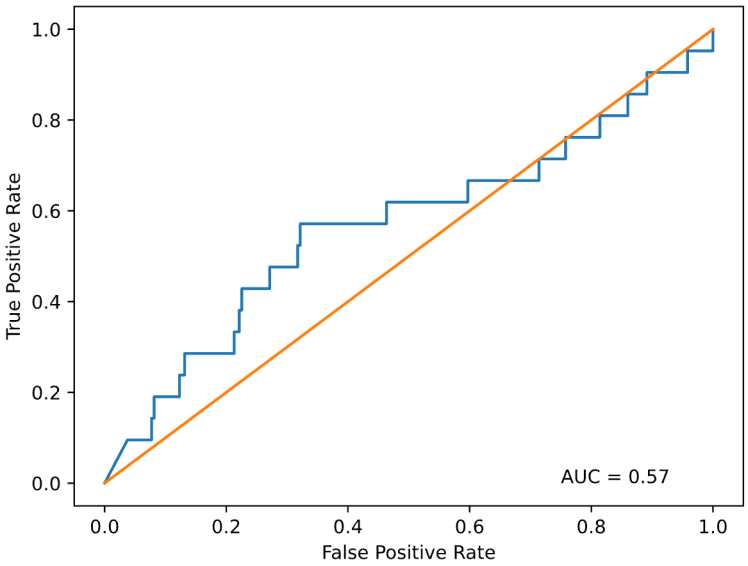
DPGD-Explain ($\epsilon = 10$)



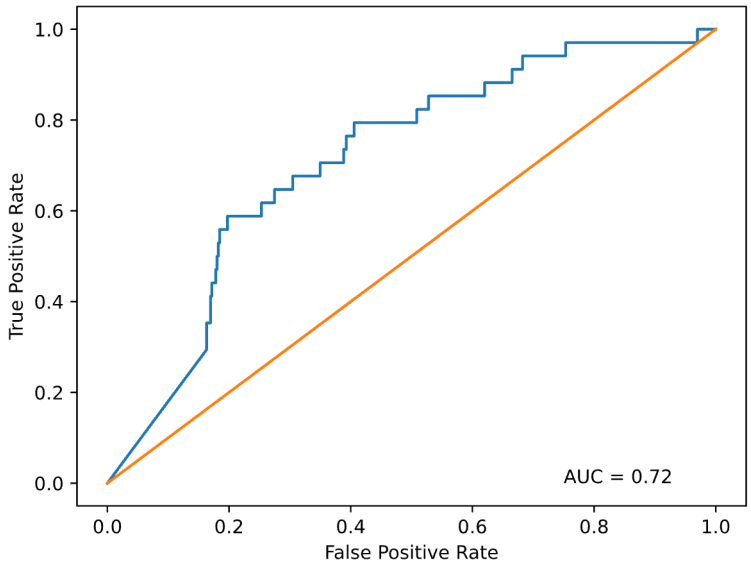
考察： IMDBにおけるプライバシーリスク

- IMDBデータセットは500次元の要素のほとんどが 0
  - 攻撃モデルの学習に失敗して 0 を多く出力しすぎている？
- ROC曲線を描くとノイズなし（$\epsilon = \infty$）の方が脆弱

$\epsilon = 10$



$\epsilon = \infty$



結論

- DPGD-Explainは安全か？
  - (ϵ, δ)-DPでモデル説明を保護しても完全には攻撃を防げない
  - 一方で，プライバシーリスクを減少させることは可能

- DPGD-Explainを他のXAI手法の代わりとして使えるか？
  - ノイズを付加したモデル説明とノイズなしのモデル説明の相関は低い
  - モデル説明の有用性を保ちながら安全性を高める必要