# 最適かつ解釈可能な近傍における
# ブラックボックス分類器の線形近似

# Linear Approximation of Black-box Classifier in Optimal and Interpretable Neighborhood

北海道大学 工学部
情報エレクトロニクス学科
情報理工学コース
情報認識学研究室

大原玄嗣

2023年3月

# 目 次

# 図 目 次



# 表 目 次

# 1　はじめに

機械学習モデルは，その精度の飛躍的な向上に伴って，近年では産業の様々な場面において活用されている．しかしその一方で，これらのモデルが複雑化・ブラックボックス化しているために，医療・金融など決定が重大な結果をもたらすような場面においては，その不透明性が実装への大きな障害となっている [1, 2, 3, 4].

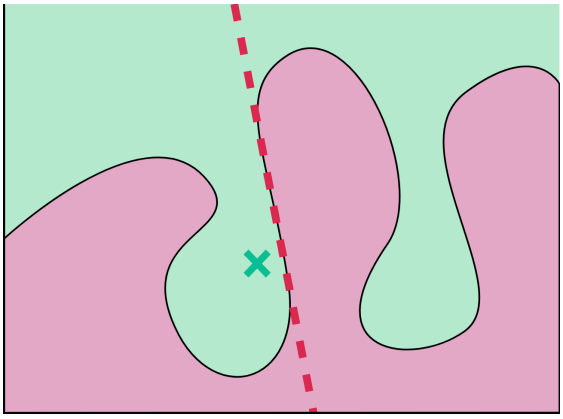
こうした課題を解決するために，機械学習モデルを解釈するための技術に関する研究が広く行われている．その一つの手法として提案された LIME[1] は，複雑で解釈不可能なモデルの決定境界を，単純で解釈可能なモデルによって局所的に近似する．この手法は，ユーザが複雑なモデルの局所的な振る舞いを解釈することを可能にする一方で，近似領域を最適化していないこと [3] や，近似領域がユーザに解釈可能な形式で提示されないこと [2] が指摘されてきた.

本稿では，別の説明手法である Anchor[2] を参考として，上記の LIME の問題点を解決するためのアルゴリズムを提案する．提案手法は，矩形の近似領域を，その内部で学習される近似モデルの精度に関する制約のもとで最大化する．このようにして得られる矩形領域は特徴量に関する述語の連言として表現されるため，ユーザにとって解釈性が高い.

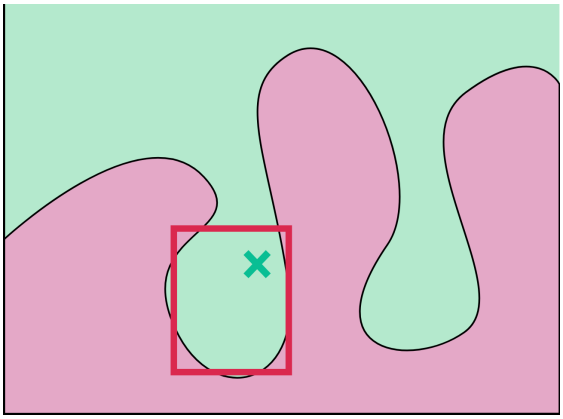
## 2 関連研究

LIME[1] は，既に学習されたブラックボックスモデルを線形モデルや決定木などの解釈可能なモデルによって局所的に近似する手法である．まず，説明対象のインスタンスの周辺で摂動ベクトルを生成し，それらをブラックボックスモデルに入力することで擬似ラベルを得る．その後，得られた局所的なサンプル集合を用いて解釈可能なモデルを学習する (図 1a)．線形 LIME は，ブラックボックスモデルの出力に対する各徴量の寄与度 (学習された線形モデルの重み) を示すことで，モデルの局所的な振舞いについての有用な示唆をユーザに与える．しかしその一方で，摂動ベクトルを生成した領域がユーザに明示されないため，説明を他のどのインスタンスに適用してよいかをユーザが判断することができないことが指摘されている [2]．図 2b は，LSTM ネットワークによる感情予測モデルおよび図 2a の 2 つのインスタンスに対して，LIME が生成した説明の例である．左の説明は “not” の語がモデルの正の予測 (その文章がポジティブな文章であるという予測) に寄与することを示しているが，これは右のインスタンスには当てはまらない．しかしユーザは生成された説明だけを見てもその適用範囲を判断できないため，誤って左の説明を右のインスタンスに適用してしまい，ブラックボックスモデルの振舞いについての誤解を生じてしまう可能性がある [2].

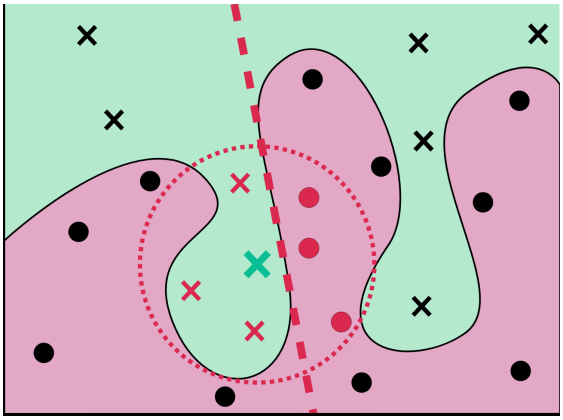
Anchor[2] は，生成された説明の有効範囲がユーザに解釈可能な形式で提示されない，という上記の LIME の欠点を克服するために提案された，LIME とは異なるアプローチによる手法である．特徴量に関する述語の連言 (ルール) として表現される矩形領域を，ブラックボックス分類器の摂動ベクトルに対する出力と説明対象のインスタンスに対する出力が高い確率で一致する限りで最大化することで，出力に大きく関与する重要な特徴量をユーザに提示する (図 1b)．Anchor は，生成された説明の有効範囲をユーザに明確に提示する一方で，説明の含む情報が LIME に比べて少ないため説明の有用性が限定される．図 2c は，LSTM ネットワークによる感情予測モデルおよび図 2a の 2 つのインスタンスに対して，Anchor が生成した説明の例である．左の説明は，“not” と “bad” を文章中に固定して他の単語を別の語に置換しても，ブラックボックス分類器の出力が変化しにくいことを示している．この説明を右のインスタンスに適用できないことは，その 2 つの語を右の文章が含んでいないことからユーザにとっても明確である．しかし説明そのものは，“not” や “bad” の語が出力に与える影響の大きさについて言及していないため，図 2b に示した LIME の説明と比較すると，モデルの振舞いについてユーザが得られる知見はより少なくなる．



(a) LIME のイメージ．説明対象のインスタンスの周辺で，ブラックボックス分類器の決定境界を局所的に線形近似する．(図は線形モデルを用いた例)



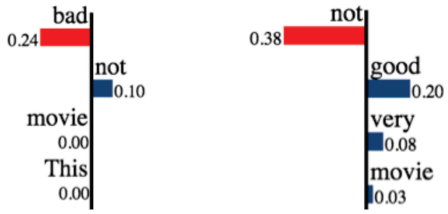
(b) Anchor のイメージ．説明対象のインスタンスを含む矩形領域を，精度 (内部におけるブラックボックス分類器の出力の純粋度) に関する制約を満たす限りで最大化する．



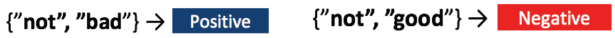
(c) BELLA のイメージ．データセットの部分集合を探索し，その部分集合で学習された線形モデルの出力とブラックボックス分類器の出力の類似度を最大化する．

図 1: LIME，Anchor，BELLA の 3 手法の視覚的な比較．


(a) 説明対象のインスタンス




(b) LIME による説明




(c) Anchor による説明

図 2: LIME と Anchor による出力結果の例 [2]

BELLA[3] は，近似領域を最適化しない，という LIME の欠点を克服するために提案された手法である．ブラックボックスモデルの学習に用いたデータセットの部分集合を線形探索し，近似モデルとブラックボックスモデルの出力の類以度 (Berry-Mielke universal R value [5]) を最大化する (図 1c)．BELLA は，精度の高い近似モデルを学習するために近傍を最適化することで，生成された説明の信頼性を高める．しかしその一方で，ブラックボックスモデルの学習に用いたデータセットに直接アクセスできることを前提としており，病院での診断結果などプライバシー保護の観点からデータセットに直接アクセスできない場合には使用することができない．また，BELLA は近似領域をデータセットの部分集合という形で提示するため，一般のユーザがそれを十分に解釈することができるとはいいがたい．

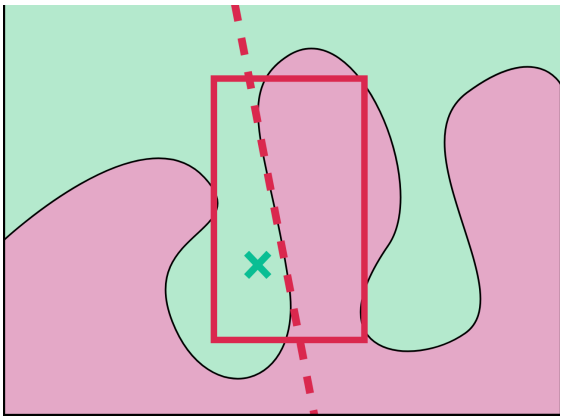


図 3: 提案手法のイメージ．説明対象のインスタンスを含む矩形領域を，内部で学習された近似モデルの精度に関する制約のもとで最大化する．

# 3　提案手法

## 3.1　概要

提案手法は，前章で紹介した LIME およびその改善手法である Anchor や BELLA が抱える欠点を克服することを目標とする．提案手法は，LIME と同様に，所与のブラックボックスモデルを線形モデルによって局所的に近似する．ただし，近似モデルを学習するための摂動ベクトルを生成する領域を矩形領域とし，特徴量に関する述語の連言 (ルール) として表現する．またその領域で生成される説明の信頼性をルールの “精度”，説明の一般性をルールの “被覆度” として定義し，精度に関する制約のもとで被覆度を最大化するようなルールを探索する (図 3).

本手法の特長を以下に列記する．

- 摂動ベクトルの生成領域が最適化され，ユーザに解釈可能な形式で提示される．
- ブラックボックス分類器の学習に用いたデータセットに直接アクセスすることなく，データセットの分布のみを利用する．
- 摂動ベクトルの個数を事前に決定することなく，精度の推定値に基づいて動的に決定する．

これによって，プライバシーに関する制約が大きい場面においてもデータに直接アクセスすることなくモデルを解釈することが可能になる．また説明の適用可能な範囲が最適化・提示され，それによってユーザは生成された説明の信頼性や一般性を評価することが可能になる．

以下では，提案手法に応用するアイデアとして，前章で紹介した Anchor についてその問題設定の概要を述べる．

### 3.2 Anchor

Anchor[2] の目標は，ブラックボックスモデルの摂動ベクトルに対する出力と説明対象のインスタンスに対する出力が一致する確率が高い矩形領域のうち，最も適用範囲の広いものを見つけることである．

離散特徴量のみで構成される $m$ 次元入力空間 $X^{(m)}$ において，学習済みのブラックボックス分類器 $f : X^{(m)} \to \{0, 1\}$，説明対象のインスタンス $x \in X^{(m)}$，入力空間 $X^{(m)}$ 上の分布 $\mathcal{D}$ が与えられているものとする．$l$ 個の述語の連言 $A(z) = a_1(z) \wedge a_2(z) \wedge \cdots \wedge a_l(z)$ を “ルール” と呼ぶ．ただし，述語 $a_i$ は，属性 $f_i \in \{0, 1, \ldots, m\}$ と対応するインスタンス $x$ の値 $x_{f_i} \in X^{(m)}_{f_i}$ の組 $\langle f_i, x_{f_i} \rangle$ として表現され，$z_{f_i} = x_{f_i}$ であるときに $a_i(z) = 1$ となる．また，ルール $A$ の精度 $\mathrm{acc}(A)$ および被覆度 $\mathrm{cov}(A)$ を以下のように定義する:

$$\mathrm{acc}(A) = \mathbb{E}_{z \sim \mathcal{D}(z|A)}[\mathbb{1}_{f(z)=f(x)}], \tag{1}$$

$$\mathrm{cov}(A) = \mathbb{E}_{z \sim \mathcal{D}(z)}[A(z)]. \tag{2}$$

ここで $\mathcal{D}(z|A)$ は，ルール $A$ のもとでの条件つき分布である．分布 $\mathcal{D}$ に従って摂動ベクトル $z$ をサンプリングするとき，“精度” は $A$ の示す領域において $z$ と $x$ に対する $f$ の出力が一致する確率を，“被覆度” は $z$ が $A$ に適合する確率を表現している．

Anchor は，精度が所与の閾値 $\tau$ を上回る限りで被覆度を最大化したい．しかし式 (1) を直接に計算することはできないため，有限回のサンプリングによって確率的に評価する必要がある．信頼係数 $1 - \delta$ を導入し，精度に関する制約を以下のように定める:

$$P(\mathrm{acc}(A) \geq \tau) \geq 1 - \delta. \tag{3}$$

以上より，解くべき最適化問題は以下のように記述される:

$$A^* = \underset{A \ s.t. \ P(\mathrm{acc}(A) \geq \tau) \geq 1-\delta, A(x)=1}{\arg\max} \mathrm{cov}(A). \tag{4}$$

この問題を近似的に解くことで，ブラックボックスモデルの出力が高い確率で固定されるような矩形領域のうち，最も一般性の高いものが得られる．

### 3.3 提案手法への応用

Anchor はブラックボックスモデルの出力を固定する矩形領域を探索するが，提案手法は高い精度の近似モデルを学習することのできる矩形領域を探索する．式 (1) による精度の定義を以下のように変更する:

$$\mathrm{acc}(A) = \max_{g \in G} \mathbb{E}_{z \sim \mathcal{D}(z|A)}[\mathbb{1}_{f(z)=g(z)}]. \tag{5}$$

ただし $G$ は可能な近似モデルの全体を表す.

このように変更された精度の定義のもとで式 (4) の最適化問題を解くことで，精度の高い近似モデルを学習することのできるルールのうち，被覆度が最大のものを選択することができる．以下に示す提案手法のアルゴリズムにおいても，その多くは Anchor において用いられるアルゴリズムに準じている．

### 3.4 アルゴリズム

式 (4) のような非凸の最適化問題に対しては貪欲法が用いられることが多いが，貪欲法は局所最適解に収束しやすいため，提案手法では各反復において複数の候補を選択するビームサーチを用いる．擬似コードを Algorithm 1 に示す．所望のルールが得られるまで以下の処理 (1 から 3) を反復している．

0. 候補ルールの集合 $\mathcal{A}_0$ を空集合 $\emptyset$ で初期化する
1. $\mathcal{A}_{i-1}$ をもとに新しい候補ルールの集合 $\bar{\mathcal{A}}_i$ を生成する (GenerateCands)
2. $\bar{\mathcal{A}}_i$ のうち精度が最大の $B$ 個を $\mathcal{A}_i$ として選択する (B-BestCands)
3. $\mathcal{A}_i$ のうち精度の制約を満たすルールを探索し，存在すればそのうち被覆度が最大の $A^*$ を選択して反復を終了する (LargestCand)

#### 3.4.1 候補ルールの生成

直前の反復において得られた $B$ 個の候補ルールを受け取り，それらに新たな述語を 1 つ付け加えることで，新たな候補ルールを生成する．擬似コードを Algorithm 2 に示す．ただし，$T(x)$ は $x$ のもつ属性と値の組 (述語) の集合 $T(x) = \{\langle i, x_i \rangle \mid i = 1, 2, \ldots, m\}$ を表し，$T(x) \setminus A$ は $T(x)$ のうちルール $A$ に含まれない述語の連言を表す.

**Algorithm 1** ビームサーチによる探索

**Input:** Black-box model $f$, Target instance $x$, Distribution $\mathcal{D}$, Threshold $\tau$

**Output:** Rule $A^*$ satisfying Eq. (4)

**hyperparameters** $B, \epsilon, \delta$

$A^* \leftarrow$ **null**, $\mathcal{A}_0 \leftarrow \{\emptyset\}$

**while** $A^* =$ **null do**

$\bar{\mathcal{A}}_t \leftarrow \textsc{GenerateCands}(\mathcal{A}_{t-1})$

$\mathcal{A}_t \leftarrow \text{B-}\textsc{BestCands}(\bar{\mathcal{A}}_t, \mathcal{D}, B, \epsilon, \delta)$

$A^* \leftarrow \textsc{LargestCand}(\mathcal{A}_t, \tau, \delta)$

**Algorithm 2** 候補ルールの生成

**function** $\textsc{GenerateCands}(\mathcal{A})$

**if** $\mathcal{A} = \emptyset$ **then return** $\{\emptyset\}$

$\bar{\mathcal{A}} \leftarrow \emptyset$

**for all** $A \in \mathcal{A}$ **do**

**for all** $a_i \in (T(x) \setminus A)$ **do**

$\bar{\mathcal{A}} \leftarrow \bar{\mathcal{A}} \cup (A \wedge a_i)$

**return** $\bar{\mathcal{A}}$

#### 3.4.2 精度が最大の候補ルールの選択

生成された候補ルールの集合 $\bar{\mathcal{A}}$ を受け取り，その中から精度が最大の $B$ 個を選択したい．これを多腕バンディット問題における最適腕識別として解く．すなわち，各候補ルール $A_i \in \bar{\mathcal{A}}$ をアーム，精度 $\mathrm{acc}(A_i)$ をそれらの報酬の分布とみなし，$z \sim \mathcal{D}(\cdot|A_i)$ をサンプリングして報酬 $\mathbb{1}_{f(z)=g_i(z)}$ を得ることを，1 回の試行とみなす．ただし $g_i$ は候補ルール $A_i$ において学習された近似モデルであり，各試行の直後にはサンプリングされた摂動ベクトル $z$ と擬似ラベル $f(z)$ を用いて $g_i$ を更新する．精度が最大のルール (アーム) を効率的に選択するために，最適腕識別の手法の一つである KL-LUCB アルゴリズム [6] を用いる．擬似コードを Algorithm 3 に示す．詳細は省略するが，許容誤差 $\epsilon \in [0, 1]$ のもとで得られた解 $\mathcal{A}$ について，以下が成り立つことが保証される [6]:

$$P(\min_{A \in \bar{\mathcal{A}}} \mathrm{acc}(A) \geq \min_{A' \in \mathcal{A}} \mathrm{acc}(A') - \epsilon) \geq 1 - \delta. \tag{6}$$

しかし，KL-LUCB アルゴリズムは各アームの報酬の分布が不変であることを前提としている一方で，提案手法はこの前提を満たさない．この問題については 5.2 節で考察する．

**Algorithm 3** 精度が最大の候補ルールの選択

**function** B-BestCands($\bar{\mathcal{A}}, \mathcal{D}, B, \epsilon, \delta$)

**initialize** $\mathrm{acc}, \mathrm{acc}_{ub}, \mathrm{acc}_{lb}$ for $\forall A \in \bar{\mathcal{A}}$

$\mathcal{A} \leftarrow$ B-ProvisionallyBestCands($\bar{\mathcal{A}}$)

$A \leftarrow \arg\min_{A \in \mathcal{A}} \mathrm{acc}_{lb}(A, \delta)$

$A' \leftarrow \arg\max_{A' \notin (\bar{\mathcal{A}} \setminus \mathcal{A})} \mathrm{acc}_{ub}(A', \delta)$

**while** $\mathrm{acc}_{ub}(A', \delta) - \mathrm{acc}_{lb}(A, \delta) > \epsilon$ **do**

**sample** $z \sim \mathcal{D}(z|A), z' \sim \mathcal{D}(z'|A')$

**update** $\mathrm{acc}, \mathrm{acc}_{ub}, \mathrm{acc}_{lb}$ for $A$ and $A'$

$\mathcal{A} \leftarrow$ B-ProvisionallyBestCands($\bar{\mathcal{A}}$)

$A \leftarrow \arg\min_{A \in \mathcal{A}} \mathrm{acc}_{lb}(A, \delta)$

$A' \leftarrow \arg\max_{A' \notin (\bar{\mathcal{A}} \setminus \mathcal{A})} \mathrm{acc}_{ub}(A', \delta)$

**return** $\mathcal{A}$

**Algorithm 4** 制約を満たす被覆度最大のルールの選択

**function** LargestCand($\mathcal{A}, \tau, \delta$)

$A^* \leftarrow$ **null**

**for all** $A \in \mathcal{A}$ s.t. $\mathrm{acc}_{lb}(A, \delta) > \tau$ **do**

**if** $\mathrm{cov}(A) > \mathrm{cov}(A^*)$ **then** $A^* \leftarrow A$

**return** $A^*$

### 3.4.3 制約を満たす被覆度最大のルールの選択

ルール $A$ が式 (3) による制約を満たすためには,

$$\mathrm{acc}_{lb}(A, \delta) > \tau \tag{7}$$

が成り立てばよい．ただし $\mathrm{acc}_{lb}(A, \delta)$ は $\mathrm{acc}(A)$ に関する $100(1-\delta)$%信頼区間の下限である．受け取った候補ルールの集合 $\mathcal{A}$ が式 (7) を満たすルールを含む場合は，そのなかで被覆度が最大のものを選択し，反復を終了する．$\mathcal{A}$ に式 (7) を満たすルールが含まれない場合は **null** を返し，次の反復に進む．擬似コードを Algorithm 4 に示す.

以上のアルゴリズムによって，式 (5) による精度の定義のもとで式 (4) の最適化問題を近似的に解くことで，学習された近似モデルの精度が所与の閾値を上回るような矩形領域 (ルール) のうち，被覆度が最大のものを近似的に選択することができる.

表 1: 実験で使用した recidivism データセットの属性とその概要．連続特徴量は全て離散化され，二値および順序特徴量のみによって構成される．

| 属性 | 概要 |
|---|---|
| Race | 人種 (黒人または白人) |
| Alcohol | アルコールに関する深刻な問題の有無 |
| Junky | 薬物使用の有無 |
| Supervised Release | 保護観察の有無 |
| Married | 結婚の有無 |
| Felony | 重罪か否か |
| WorkRelease | 仮釈放プログラムへの参加の有無 |
| Crime against Property | 財産に対する罪か否か |
| Crime against Person | 人間に対する罪か否か |
| Gender | 性別 (女性または男性) |
| Priors | 前科の数 |
| YearsSchool | 正式な学校教育を修了した年数 |
| PrisonViolations | 刑務所の規則への違反回数 |
| Age | 年齢 |
| MonthsServed | 収監期間 (月) |
| Recidivism | 再収監の有無 |

# 4 実験

提案手法の有用性を確かめるために，表形式データセットに対して LIME および提案手法による説明を生成し，両者を比較した．

## 4.1 実験設定

本実験では recidivism データセット [7] を用いた．recidivism データセットは，1979 年 7 月 1 日から 1980 年 6 月 30 日までの 1 年間に North Carolina 刑務所から釈放された 9549 名の受刑者に関するデータである．人種，性別，アルコール依存症の有無，前科の数，再収監の有無など 19 の項目についての値が記録されている．本実験では，再収監の有無 (Recidivism) を予測ラベル，その他の 18 項目に対して離散化を含む前処理を施した 15 項目を特徴量とする二値分類問題を対象とした (表 1)．この問題設定は，受刑者の保釈を決定する場面において機械学習モデルを導入したケースとして考えることができる．このような決定は受刑者の人生に非常に大きな影響を与えうるため，ユーザがブラックボックスモデルの出力を適切に解釈することは必要不可欠である．

| | |
|---|---|
| Race | White (1) |
| Alcohol | Yes (1) |
| Junky | Yes (1) |
| Supervised Release | Yes (1) |
| Married | No (0) |
| Felony | No (0) |
| WorkRelease | No (0) |
| Crime against Property | No (0) |
| Crime against Person | No (0) |
| Gender | Male (1) |
| Priors | 1 |
| YearsSchool | 8.00 < YearsSchool <= 10.00 (1) |
| PrisonViolations | 1 |
| Age | Age <= 21.00 (0) |
| MonthsServed | 4.00 < MonthsServed <= 9.00 (1) |
| Recidivism | Re-arrested (1) |

(a) インスタンス A

| | |
|---|---|
| Race | Black (0) |
| Alcohol | Yes (1) |
| Junky | No (0) |
| Supervised Release | Yes (1) |
| Married | No (0) |
| Felony | Yes (1) |
| WorkRelease | No (0) |
| Crime against Property | Yes (1) |
| Crime against Person | No (0) |
| Gender | Male (1) |
| Priors | 0 |
| YearsSchool | YearsSchool > 11.00 (3) |
| PrisonViolations | 2 |
| Age | Age > 33.00 (3) |
| MonthsServed | MonthsServed > 24.00 (3) |
| Recidivism | Re-arrested (1) |

(b) インスタンス B

図 4: recidivism データセットから抽出された 2 つのインスタンス (括弧内の数字は特徴量として付与された整数値を表す．)

まず，上記の前処理と欠損値の除去を行った 8594 のデータを訓練データ (7639) とテストデータ (955) に分割した．次に，訓練データを用いてランダムフォレスト (木の個数は 50) を学習した．その後，テストデータから抽出した 2 つのインスタンス (図 4) に対して LIME および提案手法による説明を生成した．ただし，提案手法では線形近似モデルとしてロジスティック回帰を用いた．また，ビーム幅を $B = 10$，信頼係数を $1 - \delta = 0.95$，KL-LUCB アルゴリズムの許容誤差を $\epsilon = 0.05$ とし，精度の閾値 $\tau$ は $\tau = 0.60, 0.70, 0.80$ の 3 つの値を用いた．

### 4.2 実験結果

図 5–8 に実験の結果を示す．各属性名に付与された値は，ブラックボックス分類器の出力に対する寄与度 (学習された線形近似モデルの重み) である．また，図には寄与度の絶対値が大きい 5 つの属性を示している．

LIME が生成した説明 (図 5, 7) は，前科が多いこと (Prior) や刑務所の規則違反が多いこと (Prison Violations)，財産に関する罪を犯したこと (Crime against Property) などがモデルの正の予測 (その受刑者が再収監されるという予測) に貢献しており，また年齢が高いこと (Age) や白人であること (Race)，結婚していること (Married) などがモデルの負の予測 (その受刑者が再収監されないという予測) に貢献していることを示している．このような LIME の説明はブラックボックスモデルの振舞いについての重要な示唆をユーザに与える．例えば，本実験で用いたモデルは recidivism データセットのいずれのインスタンスに対する予測においても人種 (Race) を重視しており，白人の受刑者に有利な予測 (再収監されないという予測) を下していることから，このモデルを保釈の決定の場面に導入することは公平でないと判断できる．しかし LIME はこのような有用性を具える一方で，説明が他のどのようなインスタンスに適用可能であるのかについては提示しない．

他方，提案手法 (図 6, 8) は説明の適用範囲を述語の連言として表現する．例えば図 6 の上段 ($\tau = 0.60$) の説明は，アルコール依存症の受刑者 (Alcohol=Yes) のみに適用可能であることを明確に示している．また，生成された説明の精度と近似領域の被覆度が示されるため，説明がどの程度信頼に足るものであるのかをユーザが評価することができる．例えば図 8 の下段 ($\tau = 0.80$) の被覆度は 0.002 となっている．これはインスタンス B の周辺の決定境界が複雑であり，高い精度で線形近似することが困難であることを示している．これによってユーザは，この説明の適用範囲がごく小さいものであり有用性が限られることを判断することができる．

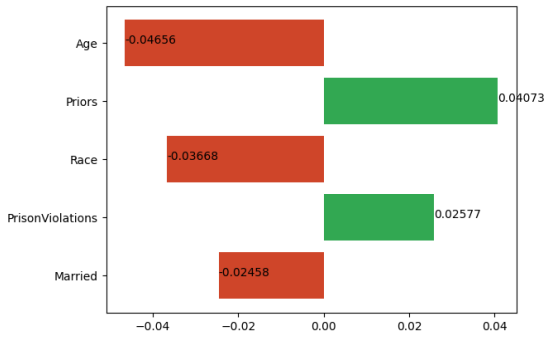



図 5: インスタンス A に対して LIME によって生成された説明

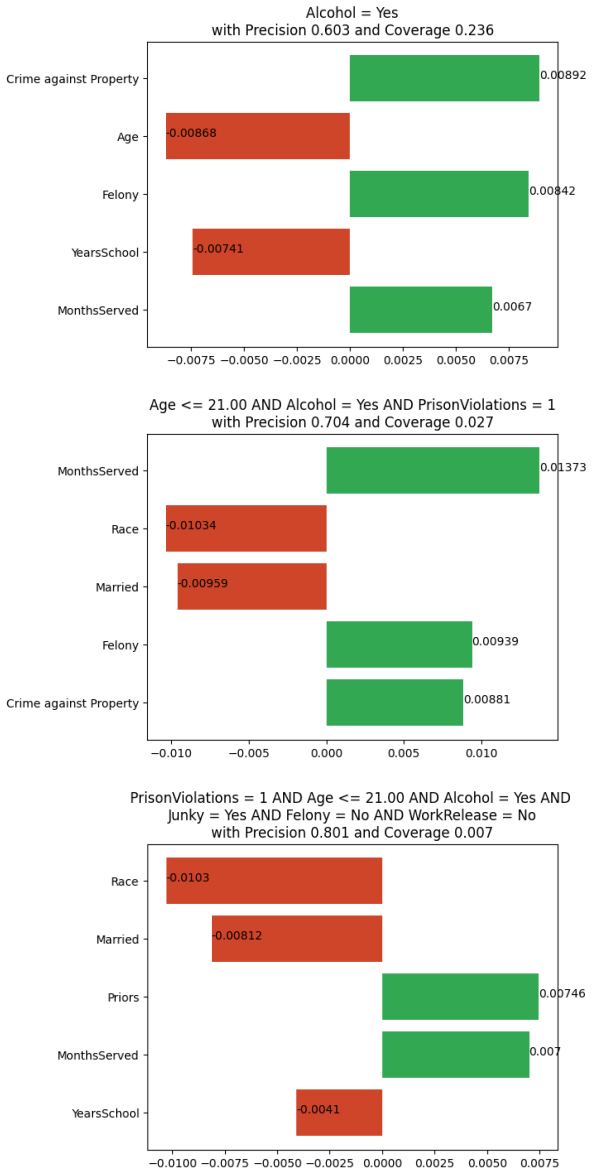



図 6: インスタンス A に対して提案手法によって生成された説明．上から $\tau = 0.60, 0.70, 0.80$

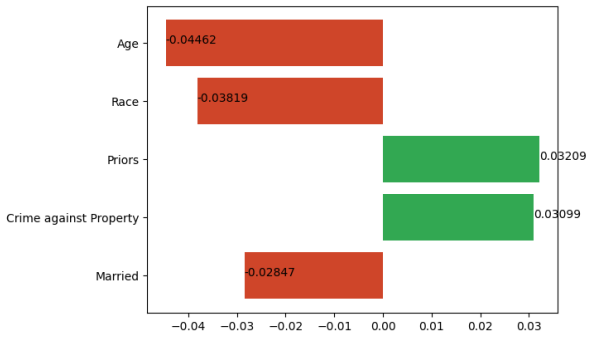



図 7: インスタンス B に対して LIME によって生成された説明

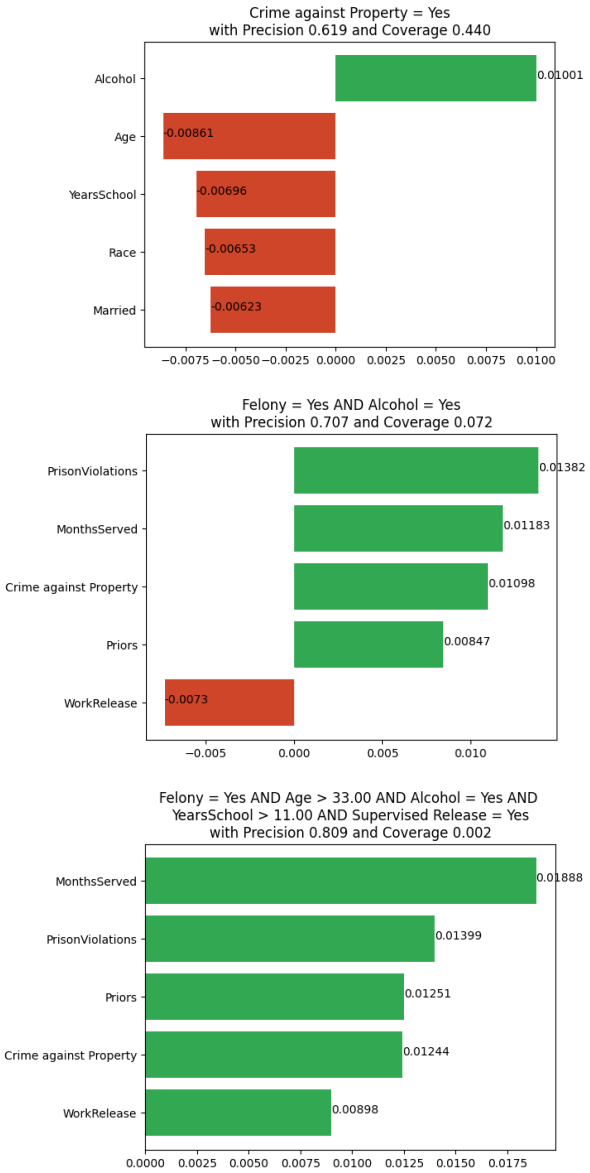
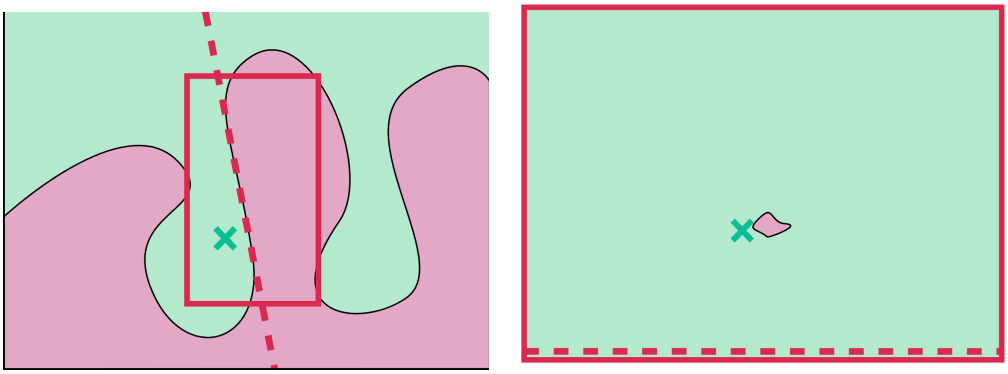



図 8: インスタンス B に対して提案手法によって生成された説明．上から $\tau = 0.60, 0.70, 0.80$



(a) 均衡なラベル分布に対する提案手法の挙動のイメージ (図 3 を再掲).

(b) 不均衡なラベル分布に対する提案手法の挙動のイメージ．近似領域は入力空間全体となり，近似モデルはつねに多数派ラベルを出力する．

図 9: 均衡・不均衡なラベル分布に対する提案手法の挙動のイメージ．

表 2: 均衡・不均衡なラベル分布で生成された説明における，近似モデルの再現率・精度・被覆度の比較．不均衡なラベル分布に対しては，再現率はつねに 0 となった．

| | 均衡データ | | 不均衡データ | |
|---|---|---|---|---|
| | 正例 | 負例 | 正例 | 負例 |
| 真のラベル | 3522 | 3522 | 391 | 3522 |
| ブラックボックスモデルの予測 | 3463 | 3581 | 276 | 3637 |
| 再現率 (Recall) | .727 | | .000 | |
| 精度 (Accuracy) | .842 | | .971 | |
| 被覆度 (Coverage) | .008 | | .869 | |

# 5　課題と今後の展望

## 5.1　不均衡なラベル分布に対する挙動

提案手法は，ブラックボックスモデルの出力の分布に偏りがある場合に，有用性の低い説明を生成する可能性がある．精度の閾値 $\tau$ に対して，ブラックボックスモデルの出力の分布が $\tau : 1 - \tau$ より大きく偏っている (少数派ラベルの割合が $\tau$ 未満である) 場合に，提案手法によって生成される説明の近似領域は入力空間全体となり，学習された線形近似モデルは任意のインスタンスに対して多数派ラベルを出力する (図 9b).

不均衡なラベル分布に対する提案手法の挙動を検証するための実験を行った．均衡・不均衡なデータセット (recidivisim データセットをもとに作成) を用いて，1000 のインスタ

表 3: 提案手法による精度の推定値と真値の比較.

| | 推定値 | 真値 | 推定値と真値の差 |
|---|---|---|---|
| 平均 | .811 | .829 | .012 |
| 標準偏差 | .018 | .023 | .017 |

ンスに対して提案手法によって説明を生成し，近似領域内での近似モデルの精度・再現率・被覆度の平均を比較した．ただし不均衡データのラベル比は 1 : 9 とし，また $\tau = 0.80$ とした．表 2 に結果を示す．不均衡なラベル分布で得られる説明は精度と被覆度が高くなった一方で，任意のインスタンスを負例として予測するため，再現率はつねに 0 となった．

この問題に対する第 1 の解決策としては，損失関数を変更することが考えられる．損失関数として重みづけ対数損失や Focal Loss[8] を用いることで，不均衡なラベル分布に対しても有用な説明が生成される可能性がある．また第 2 の解決策としては，近似領域内のラベル分布の偏りを制約することが考えられる．式 (3) に加えて，例えば

$$\left(\mathbb{E}_{z\sim\mathcal{D}(z|A)}[\mathbb{1}_{f(z)=1}] - \frac{1}{2}\right)^2 < \mu \tag{8}$$

という制約を追加することで，近似領域が過度に大きくなることを抑制することができると予想される．

### 5.2 最適腕識別における報酬の分布の変化

提案手法では，精度が最大の候補ルールを選択する問題を多腕バンディット問題における最適腕識別として定式化し，KL-LUCB アルゴリズム [6] を用いて解いた．しかしこのアルゴリズムが報酬の分布が不変であることを前提としている一方で，提案手法においてはサンプリングの直後に近似モデルが更新されるため，報酬の分布 (近似モデルの精度) はサンプリングの度に変化する．そのため，まだ近似モデルの精度が低い段階で得られた報酬が推定値に影響し，精度の真値と乖離する (真値より低くなる) 可能性がある．

精度の推定値と真値の乖離を評価するための実験を行った．データセットからサンプリングした 3200 のデータに対して説明を生成し，精度の推定値と真値を比較した．ただし $\tau = 0.80$ とし，その他の設定は 4.1 節と同様とした．精度の真値としては，近似領域からサンプリングした 1000 のデータに対して，ブラックボックスモデルと近似モデルの出力が一致した割合を用いた．結果を表 3 に示す．精度の推定値と真値の差は平均 0.012，標準偏差 0.017 であった．信頼係数 $1 - \delta = 0.95$ を考慮すると乖離の程度は小さいといえるものの，精度の変化を考慮するために選択アルゴリズムを改善する必要がある．

# 6　おわりに

ブラックボックスモデルの局所的な説明の既存手法に関する課題の解決策として，ブラックボックスモデルの決定境界を線形近似するための領域を解釈可能な形式で定義し，最適化する手法を提案した．近似領域を特徴量に関する述語の連言として表現し，近似精度に関する制約のもとで被覆度を最大化するアルゴリズムを提案した．表形式データセットに対する LIME と提案手法の出力を比較し，説明の適用範囲が明確に示される点や，説明の信頼性・一般性をユーザが評価することが可能な点において，提案手法による説明がより解釈性が高いことを示した．一方で，不均衡なラベル分布に対する挙動が不安定であることや，KL-LUCB アルゴリズムを用いることの理論的な妥当性に疑問が残ることについて，検証・議論した．

## 謝辞

本研究を行うにあたり，北海道大学大学院情報科学研究科情報理工学専攻数理科学講座情報認識学研究室の木村圭吾助教には，研究テーマの設定から方針，内容について貴重な教示を賜りましたことをお礼申し上げます．また，同研究室の工藤峰一教授にも様々な貴重なご意見を頂きましたことをお礼申し上げます．最後に，同研究室の皆様には研究や発表についてご指導・ご協力をいただきましたことに深謝いたします．

# 文献


[1] M. T. Ribeiro, S. Singh and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier.", *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, 1135–1144.

[2] M. T. Ribeiro, S. Singh and C. Guestrin, "Anchors: High-Precision Model-Agnostic Explanations.", *Proceedings of the AAAI Conference on Artificial Intelligence*, **32**-1(2018), 1527–1535.

[3] N. Radulovic, A. Bifet and F. Suchanek. BELLA: Black box model Explanations by Local Linear Approximations, 2023.

[4] R. Guidotti *et al.* Local Rule-Based Explanations of Black Box Decision Systems, 2018.

[5] K. J. Berry and P. W. Mielke, "A Generalization of Cohen's Kappa Agreement Measure to Interval Measurement and Multiple Raters.", *Educational and Psychological Measurement*, **48**(1988), 921 – 933.

[6] E. Kaufmann and S. Kalyanakrishnan, "Information Complexity in Bandit Subset Selection.", *Proceedings of the 26th Annual Conference on Learning Theory*, **30**(2013), 228–251.

[7] P. Schmidt and A. D. Witte, *Predicting Recidivism in North Carolina, 1978 and 1980.* Inter-university Consortium for Political and Social Research, 1988.

[8] T. Y. Lin *et al.*, "Focal Loss for Dense Object Detection.", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **42**-2(2020), 318–327.