R-LIME: LIME法の矩形制約と最適化

R-LIME: Rectangular Constraints and Optimization for Local Interpretable Model-agnostic Explanation Methods

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

深層学習やランダムフォレストに代表される機械学習モデルは，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，モデルが複雑化・ブラックボックス化したために，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，決定根拠の不透明性が実装への大きな障害となっている．そのため，機械学習モデルの事後説明 (*post hoc* explanation) に関する研究が広く行われるようになっている [1, 2, 3, 4]．これは，既に学習された複雑な機械学習モデル f の決定プロセスについての説明 e を，ユーザが理解しやすい形式で提示することを目的とする．

事後説明に関する既存研究は, モデルの構造への依存性によってモデル依存の手法 (model-specific method) とモデル非依存の手法 (model-agnostic method) に分類される．モデル依存の手法は主に深層ニューラルネットワークを対象として研究されており，モデル内部のパラメータを利用してモデルの振る舞いを解釈する．モデルへの直接的なアプローチが可能な一方で，同一の手法を他の構造のモデルに適用することはできない．モデル非依存の手法はモデルの内部構造やパラメータを利用せず，モデルの出力のみを用いる．利用される情報が制約される一方で，モデルの構造に依存しないため汎用性が高い．モデル非依存の手法はさらに，入力空間における局所性によって大域的手法 (global method) と局所的手法 (local method) に分類される．大域的手法は入力空間全体におけるモデルの振る舞いについての説明を出力する．これは任意の入力に妥当する説明を出力することを目指す一方で，モデルが複雑であるほどその振る舞いを大域的に解釈することは困難になる．局所的手法は，特定の入力 x に対するモデル f の出力 $f(x)$ について，その決定根拠を x の近傍で解釈する．大域的手法に比べて単純で精度の高い説明を出力することができる一方で，説明の適用可能な範囲は限定される．

本稿では, モデル非依存かつ局所的な手法に着目する. 例えばLIME (Local Interpretable Model-agnostic Explanations) [1] は，複雑で解釈不可能なモデルの決定境界を，局所的に単純で解釈可能なモデルによって近似する．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

本章では，モデル非依存かつ局所的な事後説明の既存研究のうち，特に提案手法と深く関連する研究を概観する．

2.1 LIME[1]

LIME (Local Interpretable Model-agnostic Explanations)[1] は，学習済のブラックボックス分類器 $f : \mathbb{R}^d \to \{0, 1\}$ を，着目点 $x \in \mathbb{R}^d$ の周辺で，線形モデル $g : \mathbb{R}^d \to \{0, 1\}$ によって局所近似する手法である (図 1(a))．

1. x の周辺で摂動サンプルの集合 $\mathcal{Z}_p$ および疑似ラベルの集合 $f(\mathcal{Z}_p) = \{f(z) \mid z \in \mathcal{Z}_p\}$ を得る．

2. $\mathcal{Z}_p$ および $f(\mathcal{Z}_p)$ を用いて解釈可能なモデル g を学習する．

解釈可能なモデル g は以下の損失関数を最小化することで学習される:

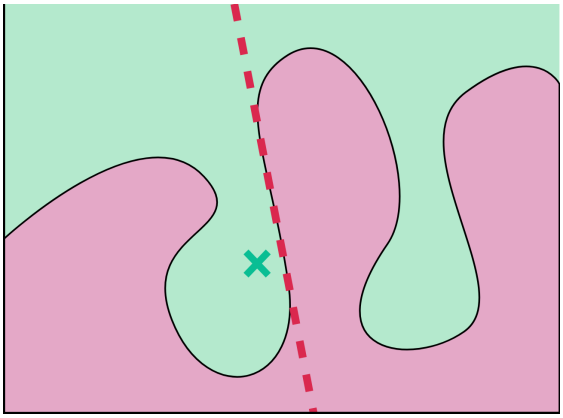
$$\mathcal{L}(f, g, \pi_x) = \sum_{z \in \mathcal{Z}_p} \pi_x(z)(f(z) - g(z))^2. \tag{1}$$

ただし $\pi_x(z)$ は，z が x に近いほど大きくなるように設計された重みづけ関数であり，指数カーネル $\pi_x(z) = \exp\left(-\frac{d(x,z)^2}{\sigma^2}\right)$ が用いられる．

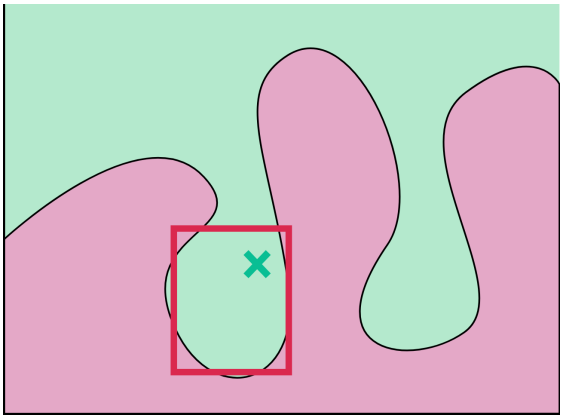
LIME は，出力 $f(x)$ に対する各特徴量の寄与度 (線形近似モデルの重み) を示すことで，モデルの局所的な振舞いについての有用な示唆をユーザに与える．しかし一方で，摂動領域がユーザに明示されないため，説明の有効範囲をユーザが判断することができない [2]．図 2(b) に LSTM ネットワークによる感情予測モデルに対する LIME の説明例を示す．左の説明は “not” の語がモデルの正の予測 (その文章がポジティブな文章であるという予測) に寄与することを示しているが，これは右のインスタンスには当てはまらない．しかしユーザは生成された説明だけを見てもその適用範囲を判断できないため，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]．

2.2 Anchor[2]

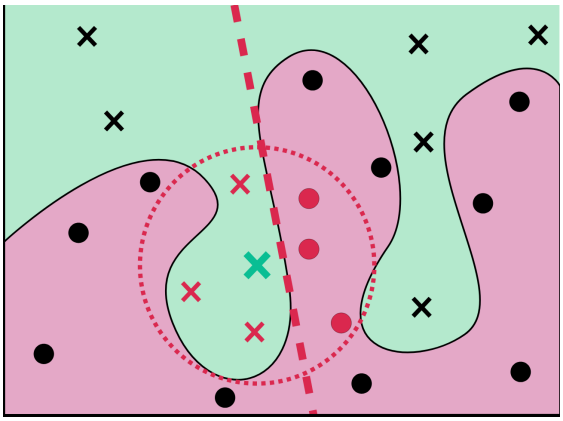
Anchor[2] は，特徴量に関する述語の連言 (ルール) として表現される矩形領域を，その内部でのブラックボックス分類器 f の出力 $f(z)$ が $f(x)$ に一致する確率が閾値を超える限



(a) LIME．着目点の周辺で決定境界を局所線形近似する．



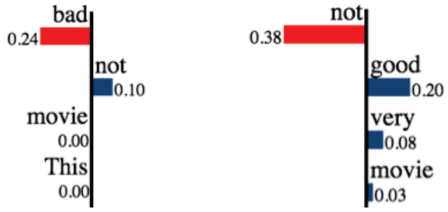
(b) Anchor．着目点を含む矩形領域の被覆度を精度制約の下で最大化する．



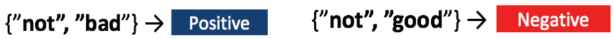
(c) BELLA．データセットの部分集合を探索し，局所線形近似モデルとブラックボックス分類器の類似度を最大化する．

図 1: LIME，Anchor，BELLA の 3 手法の視覚的な比較．

(a) 着目点のインスタンス



(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]

りで最大化する．これによって，出力に大きく関与する重要な特徴量がユーザに提示される (図 1(b))．

離散特徴量のみで構成される m 次元入力空間 $\mathbb{D}^m$ において，学習済みのブラックボックス分類器 $f : \mathbb{D}^m \to \{0, 1\}$，着目点 $x \in \mathbb{D}^m$，入力空間 $\mathbb{D}^m$ 上の分布 $\mathcal{D}$ が与えられているものとする．t 個の述語の連言 $A(z) = a_{i_1}(z) \wedge a_{i_2}(z) \wedge \cdots \wedge a_{i_t}(z)$ を “ルール” と呼ぶ．ただし述語 $a_i(z)$ は，$z_i = x_i$ であるときに真 $(= 1)$ となり，$z_i \neq x_i$ のときに偽 $(= 0)$ となる．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:

$$\mathrm{acc}(A) = \mathbb{E}_{z \sim \mathcal{D}(z|A)}[\mathbb{1}_{f(z)=f(x)}], \tag{2}$$

$$\mathrm{cov}(A) = \mathbb{E}_{z \sim \mathcal{D}(z)}[A(z)]. \tag{3}$$

ここで $\mathcal{D}(z|A)$ は，ルール A が真となる領域における条件つき分布である．$\mathrm{acc}(A)$ は領域 A において摂動 z と着目点 x に対する f の出力が一致する確率を，$\mathrm{cov}(A)$ は摂動 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 τ を上回る限りで被覆度を最大化する．しかし式 (2) を直接に計算することはできないため，サンプリングによって PAC (Probably Approximately Correct) 評価する．信頼係数 $1 - \delta$ $(0 \leq \delta \leq 1)$ を導入し，精度制約を以下のように緩和する:

$$P(\mathrm{acc}(A) \geq \tau) \geq 1 - \delta. \tag{4}$$

以上より，次の最適化問題を解く:

$$A^* = \underset{A \ s.t. \ P(\mathrm{acc}(A)\geq\tau)\geq 1-\delta \wedge A(x)=1}{\arg\max} \mathrm{cov}(A). \tag{5}$$

Anchor は説明の有効範囲を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有用性が限定される．図 2(c) に LSTM ネットワークによる感情予測モデルに対する Anchor の説明例を示す．左の説明は，“not” と “bad” を固定して他の単語を置換しても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，LIME の説明 (図 2(b)) 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．

2.3 BELLA[3]

BELLA (Black box model Explanations by Local Linear Approximations) [3] は，データセット Z を用いて学習されたブラックボックス分類器 $f : \mathbb{R}^d \to \{0, 1\}$ に対して，摂動サンプルではなく部分データセット $Z' \subset Z$ を用いて局所線形近似モデル $g : \mathbb{R}^d \to \{0, 1\}$ を学習する (図 1(c)).

1. Z に含まれる全てのインスタンス z について，着目点 x との距離 $d_x(z)$ を計算する．
2. $d_x(z)$ が最小の k 個のインスタンスを選択し，部分データセット $Z_s(k) \subseteq Z$ を得る．
3. $Z_s(k)$ を用いて局所線形近似モデル g を学習し，f と g の類似度 $\mathcal{R}(k)$ を計算する．
4. $\mathcal{R}(k)$ を最大化する k^* を得る．

ただし，類似度 $\mathcal{R}(k)$ として Berry-Mielke universal R value [5] を用いる:

$$\mathcal{R}(k) = 1 - \frac{\frac{1}{n}\sum_{z\in Z_s(k)}(f(z) - g(z))^2}{\frac{1}{n^2}\sum_{z\in Z_s(k)}\sum_{z'\in Z_s(k)}(f(z) - g(z'))^2}. \tag{6}$$

BELLA は，高精度の線形近似のために近似領域を最適化することで説明の信頼性を高める．一方で，分類器の学習に用いたデータセット Z に直接アクセスできることを前提としているため，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適用できない．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，近似領域の解釈性は Anchor に比して低い．

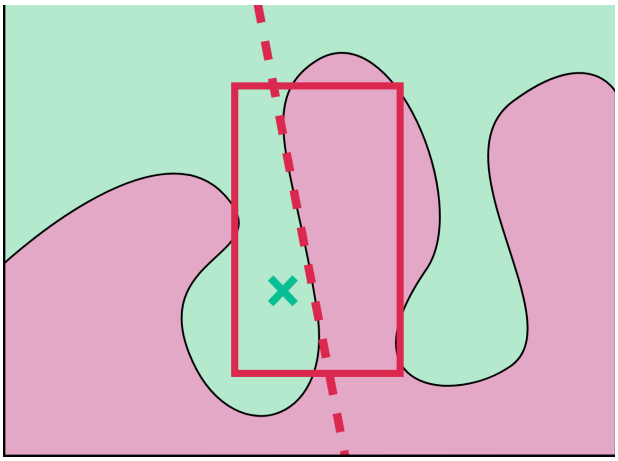


図 3: 提案手法 (R-LIME) のイメージ．着目点を含む矩形領域を，内部で学習された近似モデルの精度に関する制約のもとで最大化する．

3　提案手法

3.1　概要

提案手法は，前章で紹介した LIME[1] およびその改善手法である Anchor[2] や BELLA[3] 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分類器 f を注目点 x の周辺で線形モデル g によって局所近似する．ただし g を学習するための摂動領域を Anchor と同様に矩形とすることで，特徴量に関する述語の連言 (ルール) として表現する．説明の信頼性をルールの “精度” として，説明の一般性をルールの “被覆度” として定義し，精度に関する下限制約の下で被覆度を最大化する (図 3)．

Anchor は，矩形領域 A の内部でブラックボックス分類器 f の出力が $f(x)$ に一致する確率に関する下限制約の下で A の被覆度を最大化する (第 2.2 節)．提案手法は矩形領域 A の内部で近似モデル g を学習し，g の精度に関する下限制約の下で A の被覆度を最大化する．式 (2) による精度の定義を以下のように変更する:

$$\mathrm{acc}(A) = \max_{g \in G} \mathbb{E}_{z \sim \mathcal{D}(z|A)}[\mathbb{1}_{f(z)=g(z)}]. \tag{7}$$

ただし G は可能な線形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 (5) 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．第 3.2 節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 で用いられるアルゴリズムに準じている．

Algorithm 1 R-LIME

Input: Black-box model f, Target instance x, Distribution $\mathcal{D}$,
Threshold τ, Beam width B, Tolerance ϵ, Confidence level $1-\delta$

Output: Rule A^* satisfying Eq. (5)

$A^* \leftarrow$ **null**, $\mathcal{A}_0 \leftarrow \emptyset$, $t \leftarrow 0$ ▷ 候補ルールの集合 $\mathcal{A}_0$ を空集合で初期化

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

  $t \leftarrow t+1$

  $\bar{\mathcal{A}}_t \leftarrow \textsc{GenerateCands}(\mathcal{A}_{t-1})$ ▷ $\mathcal{A}_{t-1}$ を用いて新たな候補ルールを生成

  $\mathcal{A}_t \leftarrow \text{B-}\textsc{BestCands}(\bar{\mathcal{A}}_t, \mathcal{D}, B, \epsilon, \delta)$ ▷ $\bar{\mathcal{A}}_t$ から B 個の最良の候補ルールを選択

  $A^* \leftarrow \textsc{LargestCand}(\mathcal{A}_t, \tau, \delta)$ ▷ 式 (5) を満たす被覆度最大のルールを選択

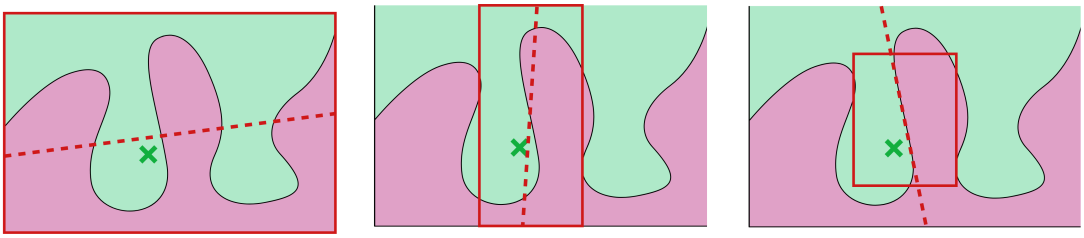


図 4: R-LIME のアルゴリズムの概要．左から右へアルゴリズムが進行する様子を示している．赤の実線は矩形領域 A を表し，赤の破線は A の内部で学習された線形近似モデル g を表す．A の初期値は空のルール (入力空間全体) であり，A に述語を加えていくことで被覆度が小さくなっていく．$\mathrm{acc}(A) \geq \tau$ となった時点で被覆度が最大のルールが出力される．

本稿では，提案手法を “R-LIME” (Ruled LIME) と呼称する．R-LIME 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．
- データセットに直接アクセスすることなく，分布のみを利用する．
- サンプリング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．

3.2 アルゴリズム

式 (5) 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．

Algorithm 2 候補ルールの生成

function GENERATECANDS($\mathcal{A}, x$)
  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 $\bar{\mathcal{A}}$

Algorithm 3 精度が最大の候補ルールの選択 (KL-LUCB アルゴリズム [6])

function B-BESTCANDS($\bar{\mathcal{A}}, \mathcal{D}, B, \epsilon, \delta$)
  initialize $\mathrm{acc}, \mathrm{acc}_u, \mathrm{acc}_l$ for $\forall A \in \bar{\mathcal{A}}$
  $\mathcal{A} \leftarrow$ B-PROVISIONALLYBESTCANDS($\bar{\mathcal{A}}$)
  $A \leftarrow \arg\min_{A \in \mathcal{A}} \mathrm{acc}_l(A, \delta)$
  $A' \leftarrow \arg\max_{A' \notin (\bar{\mathcal{A}} \setminus \mathcal{A})} \mathrm{acc}_u(A', \delta)$
  while $\mathrm{acc}_u(A', \delta) - \mathrm{acc}_l(A, \delta) > \epsilon$ **do**
    sample $z \sim \mathcal{D}(z|A), z' \sim \mathcal{D}(z'|A')$
    update $\mathrm{acc}, \mathrm{acc}_u, \mathrm{acc}_l$ for A and A'
    $\mathcal{A} \leftarrow$ B-PROVISIONALLYBESTCANDS($\bar{\mathcal{A}}$)
    $A \leftarrow \arg\min_{A \in \mathcal{A}} \mathrm{acc}_l(A, \delta)$
    $A' \leftarrow \arg\max_{A' \notin (\bar{\mathcal{A}} \setminus \mathcal{A})} \mathrm{acc}_u(A', \delta)$
  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. $\mathrm{acc}_l(A, \delta) > \tau$ **do**
    if $\mathrm{cov}(A) > \mathrm{cov}(A^*)$ **then** $A^* \leftarrow A$
  return A^*

3.2.1 候補ルールの生成

直前の反復において得られた B 個の候補ルールそれぞれに新たな述語を 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 $T(x) = \{a_1, a_2, \ldots, a_m\}$ 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．

3.2.2 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 $f(z)$ の組を用いて 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．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]:

$$P(\min_{A \in \bar{\mathcal{A}}} \mathrm{acc}(A) \geq \min_{A' \in \mathcal{A}} \mathrm{acc}(A') - \epsilon) \geq 1 - \delta. \tag{8}$$

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第 5.2 節で考察する．

3.2.3 制約を満たす被覆度最大のルールの選択

ルール A が式 (4) による制約を満たすためには，下限制約

$$\mathrm{acc}_l(A, \delta) > \tau \tag{9}$$

が成り立てばよい．ただし $\mathrm{acc}_l(A, \delta)$ 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 (9) 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 (9) 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 (7) による精度の定義のもとで式 (5) 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．連続特徴量は離散化され，二値および順序特徴量のみになっている．

属性	概要	可能な値の個数
Race	人種 (黒人または白人)	2
Alcohol	アルコールに関する深刻な問題の有無	2
Junky	薬物使用の有無	2
Supervised Release	保護観察の有無	2
Married	結婚の有無	2
Felony	重罪か否か	2
WorkRelease	仮釈放プログラムへの参加の有無	2
Crime against Property	財産に対する罪か否か	2
Crime against Person	人間に対する罪か否か	2
Gender	性別 (女性または男性)	2
Priors	前科の数	4
YearsSchool	正式な学校教育を修了した年数	4
PrisonViolations	刑務所の規則への違反回数	3
Age	年齢	4
MonthsServed	収監期間 (月)	4
Recidivism	再収監の有無	2

4 実験

提案手法の有用性を確かめるために，一つの実データセットにおいて LIME および R-LIME を比較した．

4.1 定性的評価

4.1.1 実験設定

実験には recidivism データセット [7] を用いた．これは 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．

欠損値の除去を行った 8594 個のデータを訓練データ (7639) とテストデータ (955) に分割し，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 5) に対して LIME および R-LIME による説明を生成した．ただし，R-LIME では線形近似モデルとしてロジスティック回帰を使用し，分布 $\mathcal{D}$ は訓練データから推定した多変量正規分布を用いた．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 τ は $\tau = 0.70, 0.80, 0.90$ の 3 つの値を用いた．

4.1.2 実験結果

図 6, 7 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) であり，絶対値の和が 1 になるように正規化した．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 6(a), 7(a)) は，前科が多いこと (Prior) や財産に関する罪を犯したこと (Crime against Property)，刑務所の規則違反が多いこと (Prison Violations) が主に正の予測 (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)，教育を受けた年数が長いこと (YearsSchool) が主に負の予測 (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 は説明の適用範囲を提示しておらず，ユーザは説明がどの受刑者に適用可能であるのかを判断することができない．

他方，R-LIME (図 6(b), 7(b)) 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(b) の上段 ($\tau = 0.70$) の説明は，軽度な犯罪を犯した女性 (Felony=No, Gender=Female) のみに適用可能であることを示している．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 7(b) の下段 ($\tau = 0.90$) の被覆度は 0.000 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

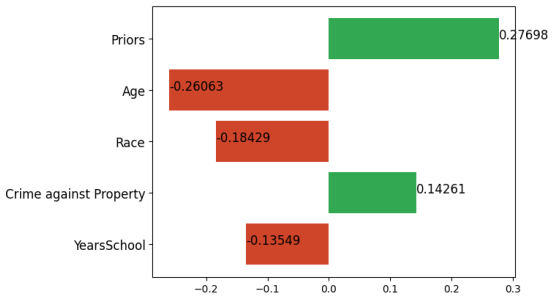
Race	White (1)
Alcohol	No (0)
Junky	No (0)
Supervised Release	Yes (1)
Married	No (0)
Felony	No (0)
WorkRelease	No (0)
Crime against Property	No (0)
Crime against Person	No (0)
Gender	Female (0)
Priors	1
YearsSchool	YearsSchool > 11.00 (3)
PrisonViolations	0
Age	21.00 < Age <= 26.00 (1)
MonthsServed	4.00 < MonthsServed <= 9.00 (1)
Recidivism	No more crimes (0)

(a) インスタンス A

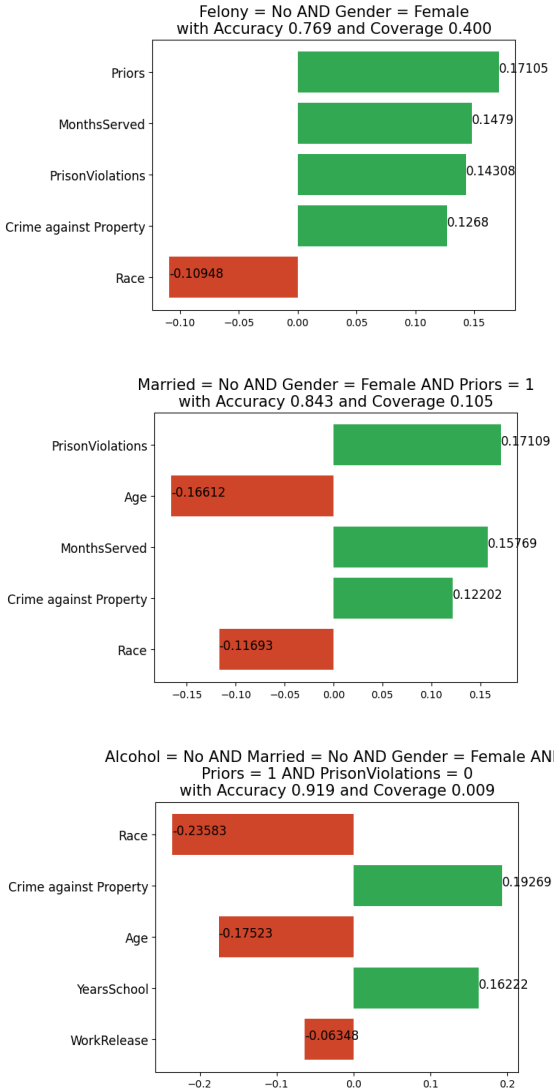
Race	White (1)
Alcohol	No (0)
Junky	No (0)
Supervised Release	Yes (1)
Married	No (0)
Felony	Yes (1)
WorkRelease	Yes (1)
Crime against Property	Yes (1)
Crime against Person	No (0)
Gender	Male (1)
Priors	1
YearsSchool	8.00 < YearsSchool <= 10.00 (1)
PrisonViolations	0
Age	Age <= 21.00 (0)
MonthsServed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

(b) インスタンス B

図 5: 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.)

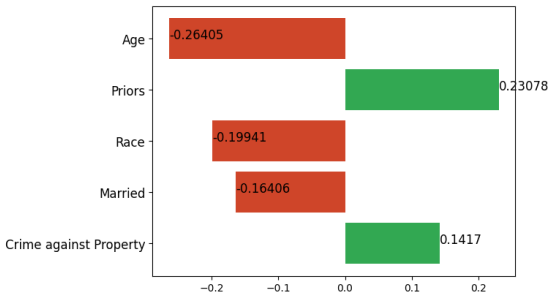


(a) インスタンス A に対して LIME によって生成された説明

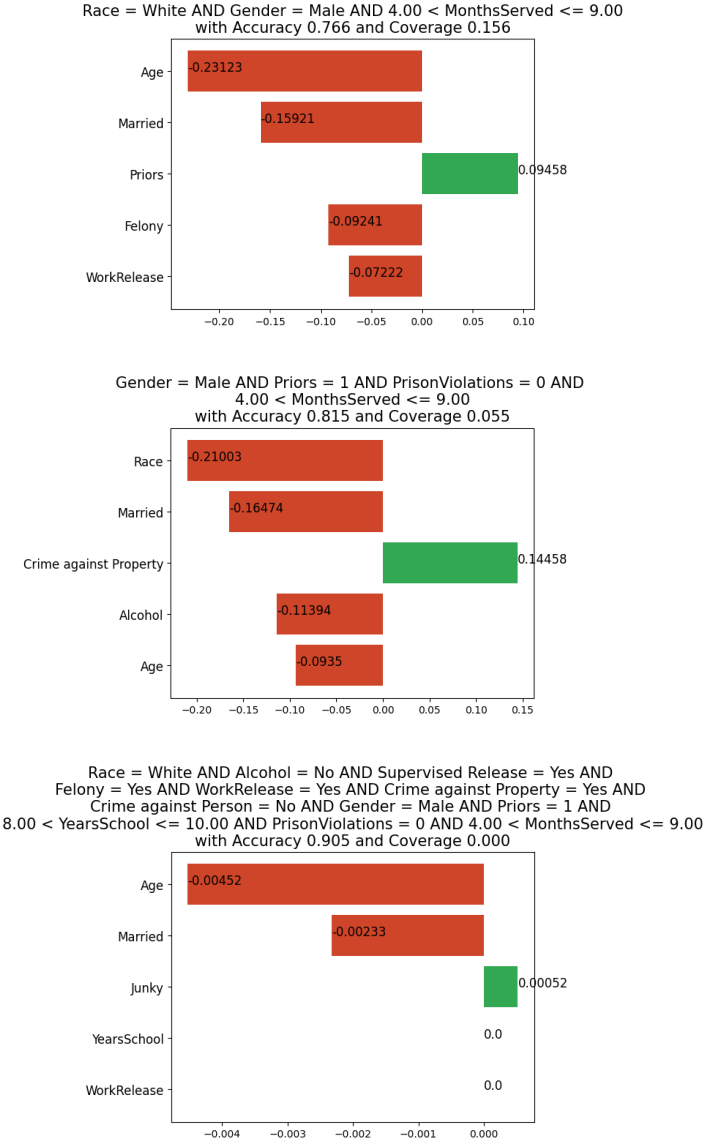


(b) インスタンス A に対して R-LIME によって生成された説明．上から $\tau = 0.70, 0.80, 0.90$

図 6: インスタンス A に対して LIME および R-LIME によって生成された説明.



(a) インスタンス B に対して LIME によって生成された説明



(b) インスタンス B に対して R-LIME によって生成された説明．上から $\tau = 0.70, 0.80, 0.90$

図 7: インスタンス B に対して LIME および R-LIME によって生成された説明.

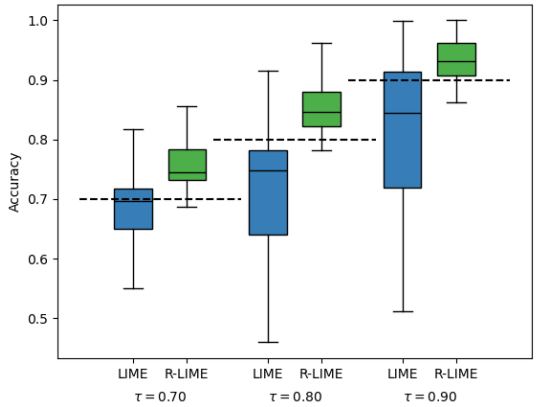


図 8: LIME と R-LIME の局所的な精度の比較．100 個のインスタンスに対して各手法による説明を生成し，R-LIME による矩形領域の内部の 10000 個のインスタンスに対する各手法の精度を計算した．

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．

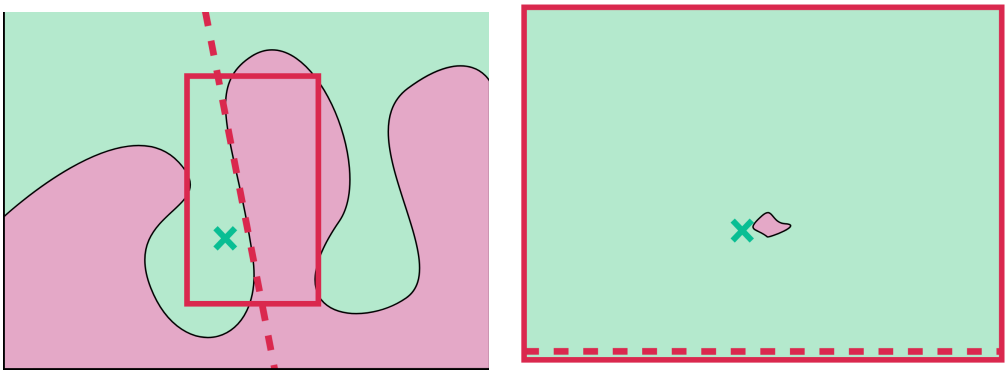
4.2 定量的評価

4.2.1 実験設定

第 4.1.1 節と同様の設定で，recidivism データセットのテストデータから 100 個のインスタンスをランダムに抽出し，それらに対して LIME および R-LIME による説明 ($\tau = 0.70, 0.80, 0.90$) を生成した．R-LIME によって得られた矩形領域の内部で 10000 個のインスタンスをサンプリングし，LIME および R-LIME による線形近似モデルの精度を計算した．

4.2.2 実験結果

結果を図 8 に示す．LIME および R-LIME による線形近似モデルの精度の分布を箱ひげ図で表している．τ のいずれの値に対しても R-LIME の精度は LIME に比べて高くなった．これは LIME と R-LIME によって学習される線形近似モデルが大きく異なっており，R-LIME は矩形領域に適応した高精度の線形近似モデルを学習していることを示している．また，τ の値が大きいほど LIME の精度のばらつきが大きくなっているが，これは LIME で学習される線形モデルは，領域の取り方によっては近似モデルとしてほとんど機能しない場合があることを示している．



(a) 均衡なラベル分布に対する R-LIME の挙動 (図 3 を再掲).

(b) 不均衡なラベル分布に対する R-LIME の挙動．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 R-LIME の挙動の比較.

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

R-LIME 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 τ 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 τ 未満である) 場合に，R-LIME 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 9(b)).

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 (4) に加えて，例えば

$$\left(\mathbb{E}_{z\sim\mathcal{D}(z|A)}[\mathbb{1}_{f(z)=1}] - \frac{1}{2}\right)^2 < \mu \tag{10}$$

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

R-LIME 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

表 2: R-LIME 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

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，R-LIME 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，その他の設定は第 4.1.1 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 2 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．

5.3 ユーザ実験による評価

機械学習モデルを “解釈する” という研究分野の性質上，各手法を定量的に比較することが難しい．本稿では第 4.2 節で定量的な実験を実施したが，それらの結果はあくまで提案手法が LIME に比して局所的に高精度の説明を生成することを示したものであり，提案手法の解釈性の高さを直接に示すものではない．

“解釈性” を定量的に評価する方法の一つとして，ユーザ実験が考えられる．例えば [2] では，説明を提示されたユーザが高い精度でブラックボックスモデルの出力を予測することができれば，その説明はモデルの振舞いについての有用な情報を含んでいる，という前提のもとで，ユーザ実験を実施している．本研究においても，提案手法の解釈性を評価するためにはユーザ実験を実施することが望ましい．

6　おわりに

ブラックボックス分類器のモデル非依存の局所的な事後説明の既存手法に関する課題を提示し，その解決策として，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，最適化する手法である R-LIME を提案した．近似のための矩形領域を特徴量に関する述語の連言として表現し，近似精度の下限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 R-LIME 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，R-LIME 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.