雑誌会

大原玄嗣

October 1, 2024

情報認識学研究室 M1

0.

発表の概要

- 卒業研究
  R-LIME: LIME 法の矩形制約と最適化

1. 卒業研究「R-LIME: LIME 法の矩形制約と最適化」

研究の背景 解釈可能な機械学習 (Interpretable Machine Learning)

- 複雑な機械学習モデル (Black Box)
  - 深層ニューラルネットワーク
  - アンサンブルモデル

  → 出力の根拠が解釈困難

- 単純な機械学習モデル (White Box)
  - 線形モデル
  - 決定木

  → 出力の根拠が解釈可能

局所近似によって解釈

関連研究

- LIME[1]
- Anchor[2]
- BELLA[3] (Appendix: Page 2, 3)

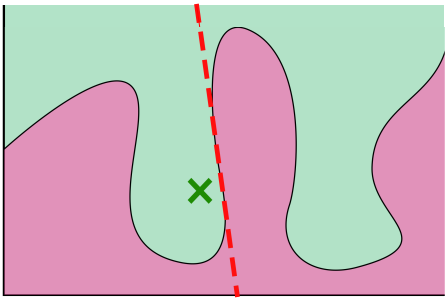
[1] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. **""Why Should I Trust You?": Explaining the Predictions of Any Classifier".** In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, 2016, pp. 1135–1144. ISBN: 978-1-4503-4232-2.

[2] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. **"Anchors: High-Precision Model-Agnostic Explanations".** In: *Proceedings of the AAAI Conference on Artificial Intelligence* 32.1 (Apr. 2018), pp. 1527–1535.

[3] Nedeljko Radulovic, Albert Bifet, and Fabian Suchanek. ***BELLA: Black box model Explanations by Local Linear Approximations.*** 2023. arXiv: `2305.11311 [cs.LG]`.

関連研究 1 — LIME (Local Interpretable Model-agnostic Explanations)[4]

1. 着目点の周辺で摂動サンプルを生成
2. 得られたサンプル集合で
   線形モデルを学習

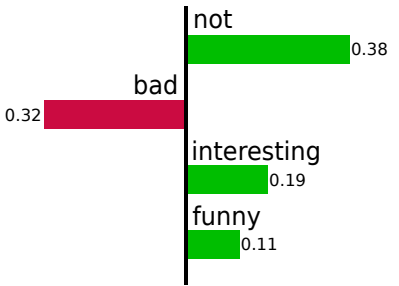


[4] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. **""Why Should I Trust You?": Explaining the Predictions of Any Classifier".** In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* KDD '16. San Francisco, California, USA: Association for Computing Machinery, 2016, pp. 1135–1144. ISBN: 978-1-4503-4232-2.

関連研究 1 — LIME の出力例

This book is not bad.
It is funny and interesting.

着目点の例．感情予測モデルはこの文章を
“Positive” な文章と予測した．

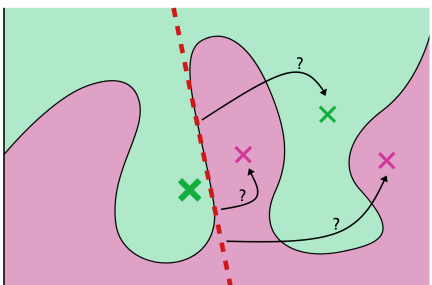


着目点に対する感情予測モデルの予測に関して，LIME が
出力する説明の例．

関連研究 1 — LIME の問題点

説明の適用範囲が示されない

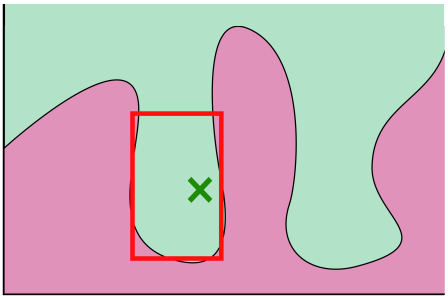
- 説明から得られた知見は
  どの程度一般的なのか？



関連研究 2 — Anchor[5]

- モデルの出力が高い確率で
  一致する矩形領域を探索
- 矩形領域を特徴量に関する
  述語の連言として表現

  ex. Gender = 'Male' AND 20 <= Age < 30



[5] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. **"Anchors: High-Precision Model-Agnostic Explanations".** In: *Proceedings of the AAAI Conference on Artificial Intelligence* 32.1 (Apr. 2018), pp. 1527–1535.

関連研究 2 — Anchor の出力例

This book is not bad.
It is funny and interesting.

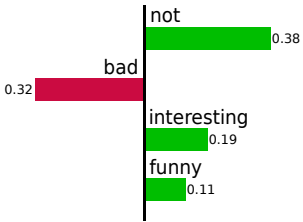
着目点の例．感情予測モデルはこの文章を“Positive” な文章と予測した．

着目点に対する感情予測モデルの予測に関して，Anchor が出力する説明の例．

関連研究 2 — Anchor の問題点

説明から得られる知見が少ない

- 各特徴量の影響の大きさが示されない



感情予測モデルに対する LIME と Anchor の出力の比較

研究の目的

	LIME	Anchor	提案手法
特徴量重要度の提示	✓	×	✓
近似領域の最適化	×	✓	✓
近似領域の解釈性	×	✓	✓

説明の解釈性と説明の適用範囲の解釈性を両立

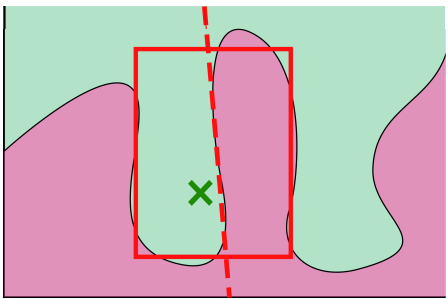
→ ユーザは説明から得られた知見を正当な範囲で活用できる

提案手法: R-LIME (Ruled LIME)

R-LIME = LIME + Anchor

- 近似領域を矩形に
- 近似領域を特徴量に関する
  述語の連言として表現

  ex. Gender = ‘Male’ AND 20 <= Age < 30



提案手法: 問題設定

m 次元入力空間（離散化済み）	$\mathbb{D}^m$
ブラックボックス分類器	$f : \mathbb{D}^m \to \{0, 1\}$
着目点	$x \in \mathbb{D}^m$
入力空間上の分布	$\mathcal{D}$
可能な線形近似モデルの全体	G

ルール … 特徴量に関する述語の連言

$$A(z) = a_{i_1}(z) \wedge a_{i_2}(z) \wedge \cdots \wedge a_{i_k}(z), \quad a_i(z) = \mathbb{1}_{z_i = x_i}$$

提案手法: 問題設定

矩形領域で学習された
近似モデルの精度の期待値

ルール A の精度　$\mathrm{acc}(A) = \max_{g \in G} \mathbb{E}_{z \sim \mathcal{D}(z|A)}[\mathbb{1}_{f(z)=g(z)}]$

ルール A の被覆度　$\mathrm{cov}(A) = \mathbb{E}_{z \sim \mathcal{D}(z)}[A(z)]$

摂動サンプルが矩形領域に
含まれる確率の期待値

最適化問題　$\tilde{A} = \underset{A \ s.t. \ P(\mathrm{acc}(A) \geq \tau) \geq 1-\delta, A(x)=1}{\arg\max} \mathrm{cov}(A)$

精度の下限制約の下で被覆度を最大化

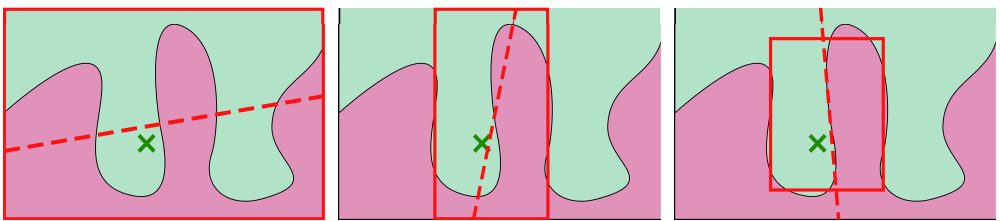
提案手法: アルゴリズム (ビームサーチ)

以下の処理を反復

1. $\mathcal{A}_i \leftarrow$ 候補ルールの集合
2. $\mathcal{A}_i \leftarrow$ 精度が最大の B 個のルール
3. 精度制約を満たす被覆度最大のルールを選択 → 見つかれば終了

$\mathcal{A}_{i-1}$ の中の各ルールに
まだ含まれていない述語を 1 個加える

多腕バンディット問題の
最適腕識別として解く



実験 1 - 設定実データセットについて，LIME と R-LIME を視覚的に比較

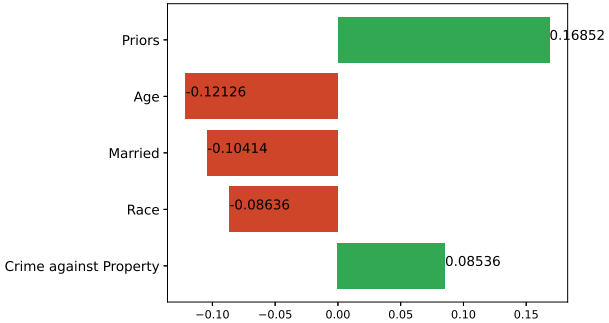
- recidivism データセット[6]を使用
- ブラックボックス分類器 (ランダムフォレスト) を学習
- データセットから抽出したインスタンスに対する LIME と R-LIME の説明を比較

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

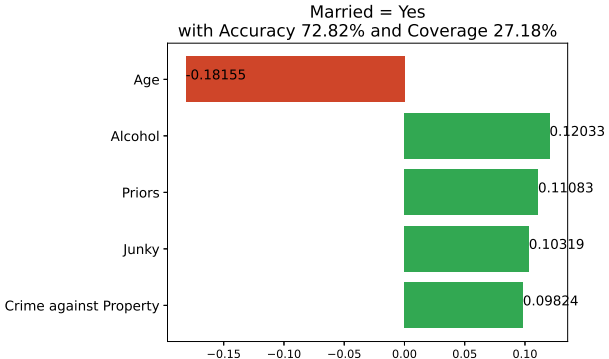
実験 1 - 設定: 着目点となる入力

Race	Black (0)
Alcohol	No (0)
Junky	No (0)
Supervised Release	Yes (1)
Married	Yes (1)
Felony	No (0)
WorkRelease	Yes (1)
Crime against Property	No (0)
Crime against Person	No (0)
Gender	Male (1)
Priors	1
YearsSchool	8.00 < YearsSchool <= 10.00 (1)
PrisonViolations	0
Age	Age > 33.00 (3)
MonthsServed	4.00 < MonthsServed <= 9.00 (1)
Recidivism	No more crimes (0)

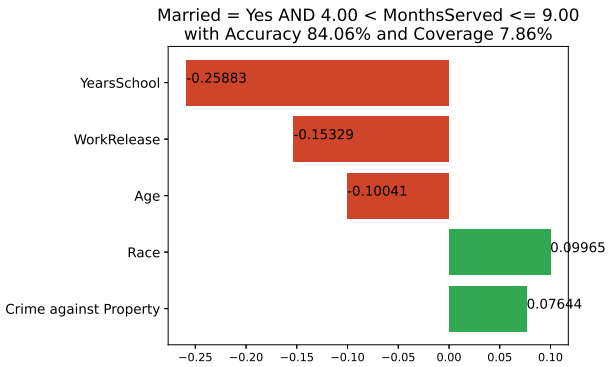
実験 1 - 結果: LIME による説明ランダムフォレストの出力に対する各特徴量の寄与度を提示



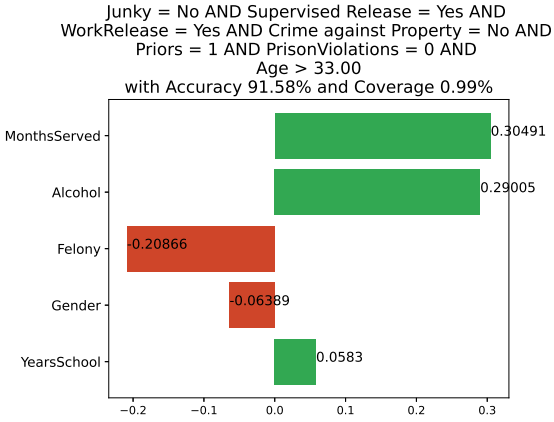
実験 1 - 結果: 提案手法による説明 (閾値 $\tau = 0.70$) 既婚の受刑者のみに適用可能



実験 1 - 結果: 提案手法による説明 (閾値 $\tau = 0.80$) 収監期間による制約が追加され，被覆度 (Coverage) は $\tau = 0.70$ のときに比べて減少



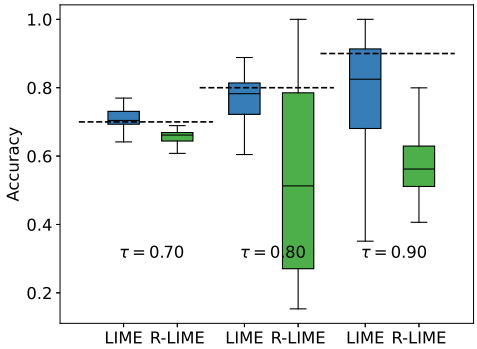
実験 1 - 結果: 提案手法による説明 (閾値 $\tau = 0.90$) 被覆度が小さい → 有用性は限定的



実験 2 - 設定 LIME と R-LIME の局所的な近似精度を比較

- recidivism データセットでランダムフォレストを学習
- データセットから無作為に抽出された 100 個のインスタンスに対して以下の処理を繰り返す
  - LIME と R-LIME の説明を生成
  - R-LIME によって得られた矩形領域から 10000 個サンプリング
  - 両者の近似精度を計算
- 得られた近似精度の分布を比較

実験 2 - 結果



R-LIME は近似領域に適応した高精度の線形近似モデルを学習

領域の取り方によっては，LIME の説明は近似モデルとしてほとんど機能しない

結論

	LIME	Anchor	R-LIME
特徴量重要度の提示	✓	×	✓
近似領域の最適化	×	✓	✓
近似領域の解釈性	×	✓	✓

提案手法によって，ユーザは説明の適用範囲や
信頼性・一般性を評価し正しく活用することができる

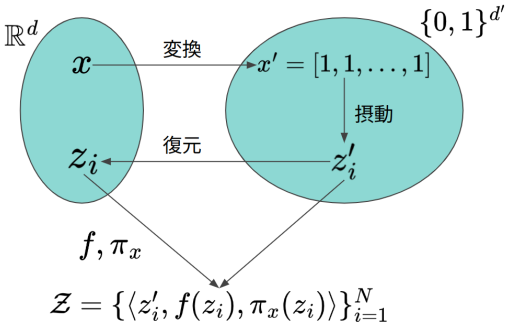
2. Appendix

2. Appendix

関連研究 1 — LIME のサンプリング

1. 着目点 $x \in \mathbb{R}^d$ を解釈可能な表現 $x' \in \{0,1\}^{d'}$ に変換
2. x' の周辺で摂動 $z'_i \in \{0,1\}^{d'}$ を生成
3. z'_i から $z_i \in \mathbb{R}^d$ を復元

$$\pi_x(z_i) = \exp\left(\frac{-d(x, z_i)^2}{\delta^2}\right) : z_i \text{ の重み}$$



2. Appendix

関連研究 3 — BELLA (Black box model Explanations by Local Linear Approximations)[7]

1. データセットに含まれる全てのデータと対象のインスタンスの距離を計算
2. データセットの「最適な部分集合」を探索
3. 得られた集合で，線形モデルを学習

[7] Nedeljko Radulovic, Albert Bifet, and Fabian Suchanek. ***BELLA: Black box model Explanations by Local Linear Approximations.*** 2023. arXiv: 2305.11311 [cs.LG].

2. Appendix

関連研究 3 — BELLA の問題点 ブラックボックスモデルの学習に用いたデータセットを使う

- プライバシーの制約に弱い
- 「どう学習したか」ではなく「どう振る舞うか」が知りたい

近似領域が解釈困難

- データセットの部分集合からは，説明の適用範囲を判断できない

2. Appendix

関連研究と提案手法の比較

	LIME	Anchor	BELLA	**R-LIME**
近似領域の最適化	×	✓	✓	✓
近似領域の解釈性	×	✓	×	✓
特徴量重要度の提示	✓	×	✓	✓
摂動サンプルの生成	✓	✓	×	✓

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. (??)

$A^* \leftarrow$ null, $\mathcal{A}_0 \leftarrow \emptyset$, $t \leftarrow 0$ ▷ *Initialize the set of candidate rules $\mathcal{A}_0$ to $\emptyset$*

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

  $t \leftarrow t+1$

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

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

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

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

Algorithm 2 Generating new candidate rules

1: **function** GENERATECANDS($\mathcal{A}, x$)
2: **if** $\mathcal{A} = \emptyset$ **then return** $\{true\}$ ▷ *An initial empty rule always returns* $true$
3: $\bar{\mathcal{A}} \leftarrow \emptyset$
4: **for all** $A \in \mathcal{A}$ **do**
5: **for all** $a \in (T(x) \setminus A)$ **do**
6: $\bar{\mathcal{A}} \leftarrow \bar{\mathcal{A}} \cup (A \wedge a)$ ▷ *Get a new rule by adding a new predicate* a *to* A
7: **return** $\bar{\mathcal{A}}$

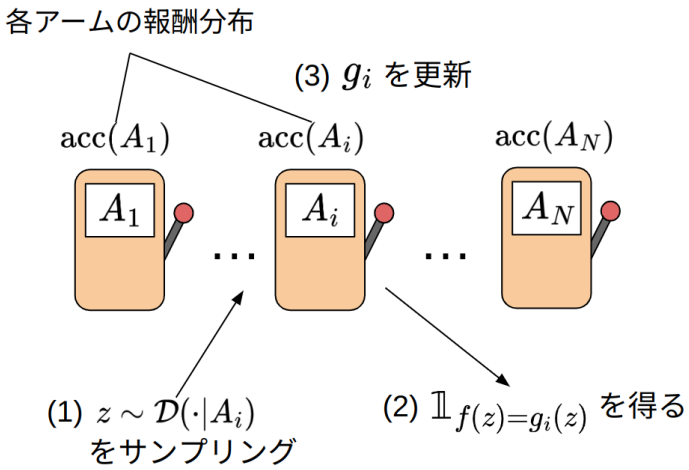
Algorithm 3 Searching rules with highest accuracy (KL-LUCB [5])

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($\bar{\mathcal{A}}$) ▷ B rules with highest accuracy
  $A \leftarrow \arg\min_{A \in \mathcal{A}} \mathrm{acc}_l(A, \delta)$ ▷ The rule with the smallest lower bound
  $A' \leftarrow \arg\max_{A' \notin (\bar{\mathcal{A}} \setminus \mathcal{A})} \mathrm{acc}_u(A', \delta)$ ▷ The rule with the largest upper bound
  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 Searching a rule with highest coverage under constraint

1: **function** LARGESTCAND($\mathcal{A}, \tau, \delta$)
2: $A^* \leftarrow$ **null** ▷ *If no rule satisfies the constraint, return* ***null***
3: **for all** $A \in \mathcal{A}$ s.t. $\mathrm{acc}_l(A, \delta) > \tau$ **do**
4: **if** $\mathrm{cov}(A) > \mathrm{cov}(A^*)$ **then** $A^* \leftarrow A$
5: **return** A^*

最良の候補ルールの選択



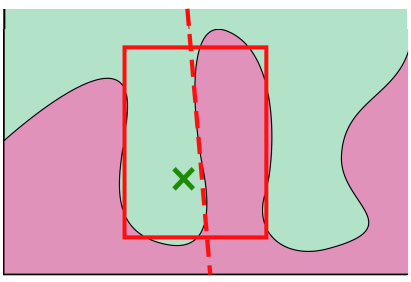
2. Appendix

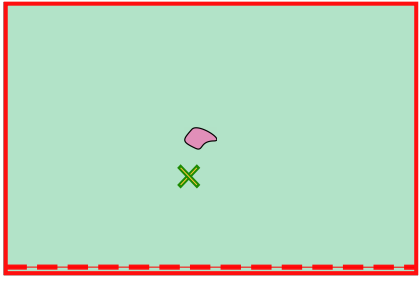
今後の展望

- アルゴリズムの理論的な課題の解決
  - 不均衡なラベル分布に対する挙動
  - 最適腕識別における報酬の分布の変化
- 手法の定量的な比較方法の検討
  - 説明の「解釈性」をどのように評価するか？
  - ユーザ実験を実施

2. Appendix

課題 1: 不均衡なラベル分布に対する挙動ブラックボックス分類器の出力の分布が $\tau : 1 - \tau$ 以上に偏っている場合...





近似領域は入力空間全体になり，近似モデルは任意の入力を多数派クラスに分類

2. Appendix

課題 1 の解決策

- 損失関数を変更
  - クラス比による重みづけ
  - Focal Loss[8]
- ラベル分布の偏りを制約
  - 次の制約を追加

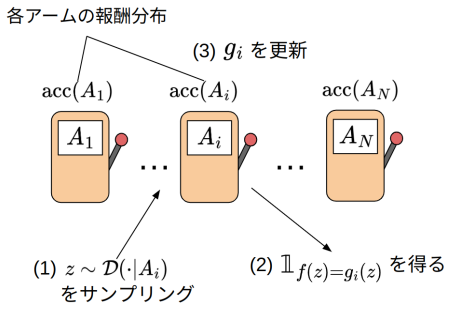
$$\left(\mathbb{E}_{z\sim\mathcal{D}(z|A)}[\mathbb{1}_{f(z)=1}] - \frac{1}{2}\right)^2 < \mu$$

[8] Tsung Yi Lin et al. **"Focal Loss for Dense Object Detection".** In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.2 (2020), pp. 318–327.

2. Appendix

課題 2: 最適腕識別における報酬の分布の変化 KL-LUCB アルゴリズム[9]による最適腕識別

- 報酬の分布が不変であるという前提
- 近似モデルの更新によって，試行の度に報酬の分布が変化



[9] Emilie Kaufmann and Shivaram Kalyanakrishnan. **"Information Complexity in Bandit Subset Selection"**. In: *Proceedings of the 26th Annual Conference on Learning Theory*. Ed. by Shai Shalev-Shwartz and Ingo Steinwart. Vol. 30. Proceedings of Machine Learning Research. Princeton, NJ, USA: PMLR, Dec. 2013, pp. 228–251.

2. Appendix

課題 2 の検証

	Estimated acc.	True acc.	Deviation
Average	.811	.829	.012
Standard Deviation	.018	.023	.017

R-LIME による精度の推定値と真値の比較．

信頼係数 $1 - \delta = 0.95$ を考慮すると，乖離の度合いは許容範囲内

→ 実用上は問題になりにくいが，選択アルゴリズムの改善が必要