

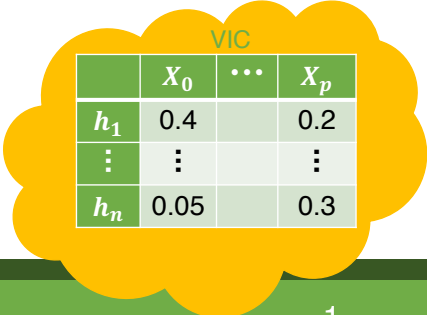
Variable Importance Cloud の 要約方法と決定木に対する実験的評価

○又 康太, 金森 憲太郎, 有村 博紀 (北海道大学)

発表概要

Dong と Rudin らが提案した Variable Importance Cloud
(羅生門集合上の特徴量重要度ベクトルの集合) の
要約方法を提案する

決定木のクラスに対して実データセット上で提案した
要約情報を求め, その有用性について検証を行う



VIC

	X_0	\dots	X_p
h_1	0.4		0.2
\vdots	\vdots		\vdots
h_n	0.05		0.3

研究背景：特徴量重要度と羅生門効果

- **特徴量重要度** (VI : Variable Importance)

機械学習モデルの予測に対して各特徴量がどの程度
寄与したかを評価する指標

- **羅生門効果** [Breiman+, 2001]

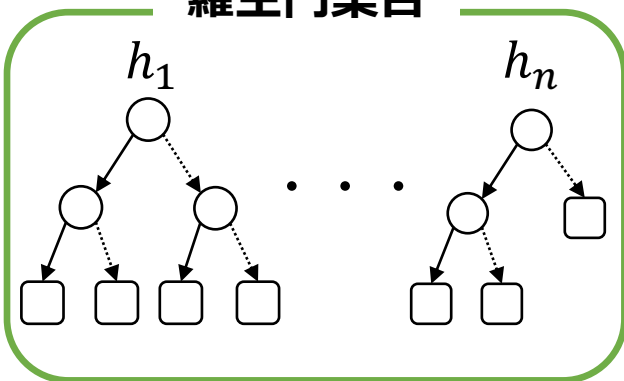
一つの予測タスクに対して, **同等の予測精度を持つ
異なるモデル**が複数存在し得ることが知られている

単一のモデルの VI だけに着目すると,
本来は重要な**知見を見落とし**たり,
不都合な**発見が隠蔽される**リスクが指摘されている

研究内容：Variable Importance Cloud

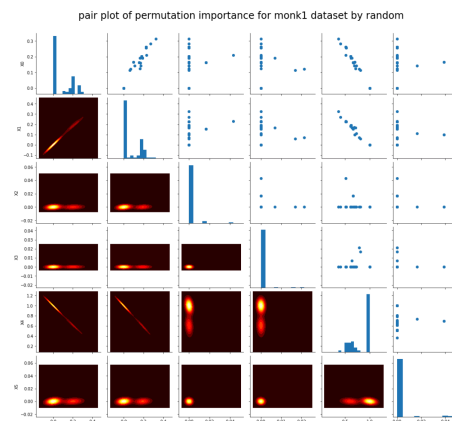
- **羅生門集合**：予測精度が準最適であるモデルの集合
(今回の実験ではこのモデル集合を求める)
- **Variable Importance Cloud (VIC)** [Dong+, 2019]
 - 羅生門集合に含まれるモデルの特微量重要度ベクトルの集合
 - 特微量重要度ベクトル：全特微量の重要度(VI)を並べたベクトル
- **VIC のメリット**：特微量重要度に関して、高精度なモデル全てに共通する性質（分布，平均，分散，最大/最小など）がわかること

羅生門集合



VIC

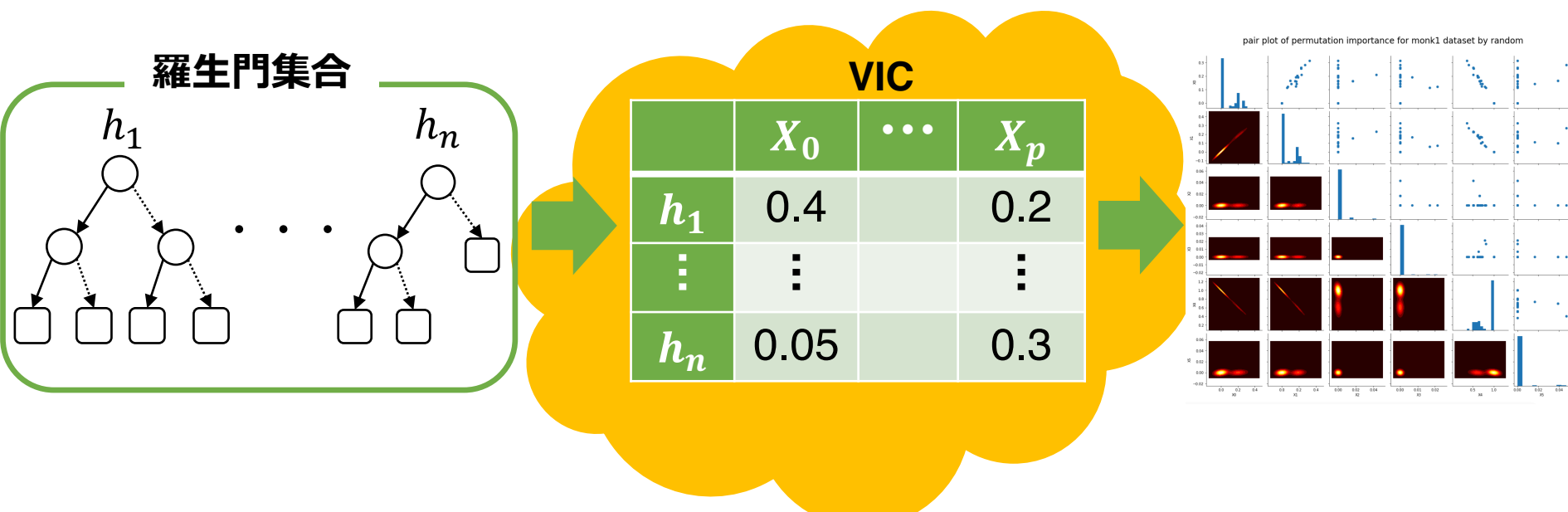
	X_0	\dots	X_p
h_1	0.4		0.2
\vdots	\vdots		\vdots
h_n	0.05		0.3



Dong, J. and Rudin, C. Variable importance clouds: A way to explore variable importance for the set of good models. *arXiv preprint arXiv:1901.03209*, 2019.

今日の発表内容

- 決定木に対する VIC (Variable Importance Cloud) の構築方法を提案
- 実データセット上での実験を行い, VIC の要約方法とその有用性を検証



Dong, J. and Rudin, C. Variable importance clouds: A way to explore variable importance for the set of good models. *arXiv preprint arXiv:1901.03209*, 2019.

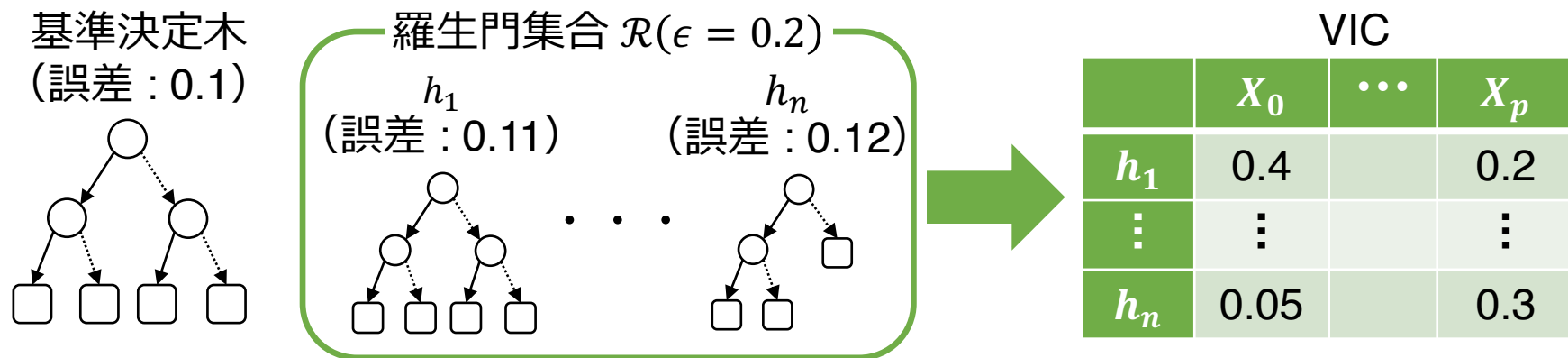
VIC の構築法

既存手法と提案手法の共通手順

1. **基準決定木**として最適決定木を一つ学習する（誤差は L^* ）
2. データセット上で決定木を列挙し**訓練誤差が $(1 + \epsilon)L^*$ 以下の決定木の集合を羅生門集合 \mathcal{R}** とする
3. 羅生門集合 \mathcal{R} 内の各決定木に対して、特徴量重要度を計算する

備考

- 決定木の学習にはTDIDT学習器（CART等）を用いる
- 手法の違いは羅生門集合の構築方法（決定木の列挙方法）である



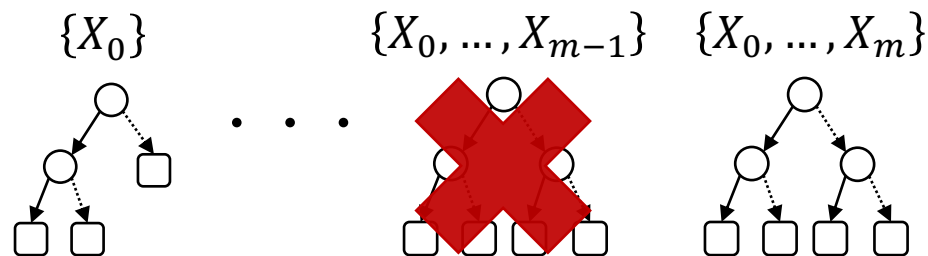
VIC の構築法：既存手法

[Dong+, 2019] は、羅生門集合の構築法として以下の二つの方法を提案した ($m = 4$)

既存手法 1：特徴量集合列挙法

1. m 個の特徴量の組合せを列挙し、それぞれで決定木を学習する
2. 重複した決定木を削除する

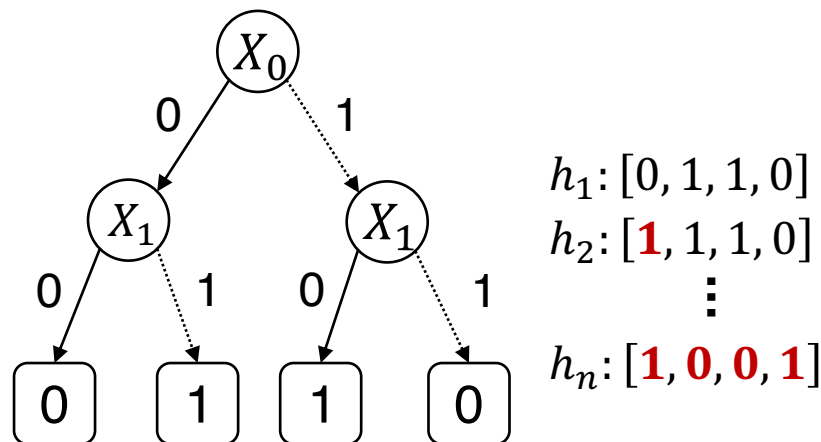
短所： m が大きいと時間がかかる



既存手法2：葉ラベル列挙法

1. m 個の特徴量を固定し、最適な決定表 (decision fern) を構築する
2. 葉のラベルをフリップし、訓練誤差の昇順になるように葉ラベルの組合せを列挙する

短所：特徴量と木の形が固定される



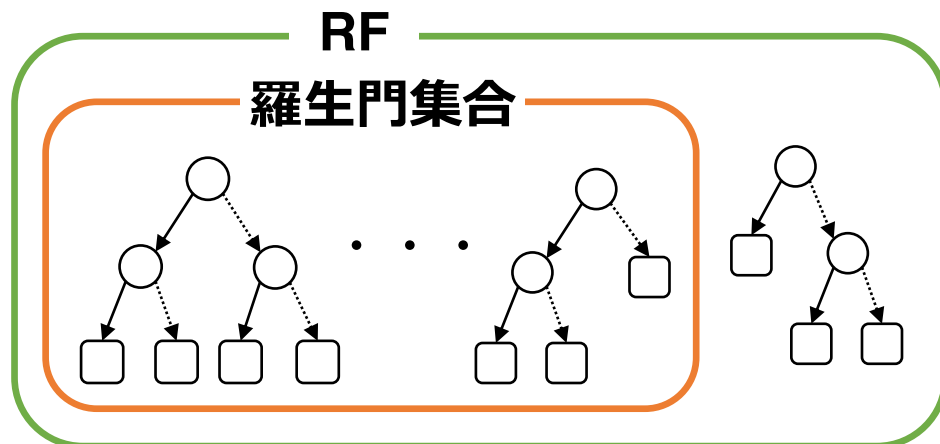
VIC の構築法：提案手法

提案手法 1：ランダム木生成法

1. ランダムフォレストや ExtraTrees で森を作成
2. 森から使用した決定木を取り出す

長所：既存手法をそのまま使える

短所：多くの繰り返しが必要

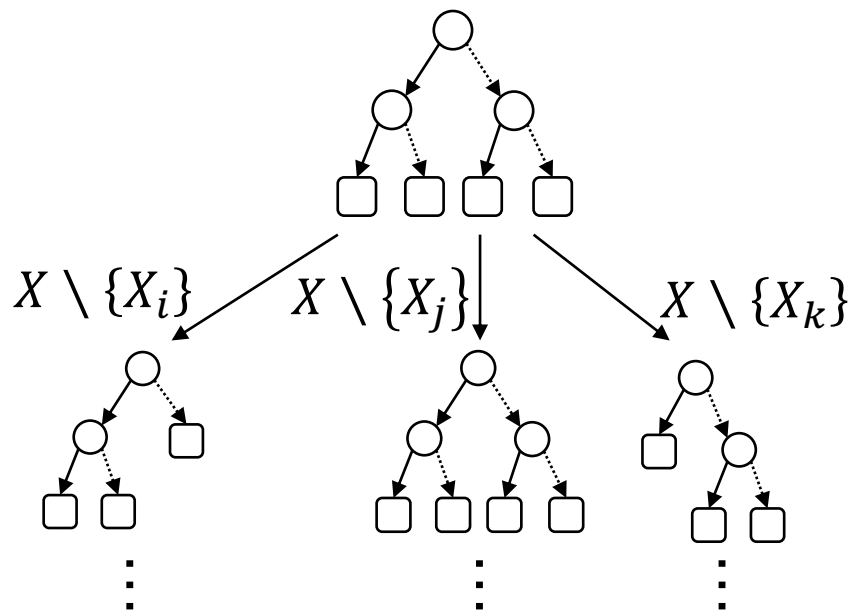


提案手法 2：Lawler 法 [Ruggieri+, ICML2017] による決定木列挙

1. 全特徴量を用いて基準決定木を学習
2. 以下を再帰的に行う：決定木で使った特徴量を一つ削除したデータセットを作成し、決定木を学習

長所：異なる特徴量集合を重複なく列挙可能

短所：特徴量数が少ないと多くの木を作れない



実験設定

使用したデータセット (VIC構築法)

- monk1 (ランダム木生成法)
- Adult (Lawler 法)

実験手順

1. 決定木列挙法を用いて, 誤差が小さい**トップ k ($k = 100$) の決定木から羅生門集合 \mathcal{R} を求める** (準最適解の代用)
2. \mathcal{R} に対応する VIC 上で二つの特徴量の重要度に対して, **相関係数と, 散布図, ヒートマップを計算する**

特徴量重要度は次を使用した

- Information gain に基づく特徴量重要度 (VI)
- Permutation Importance (PI)

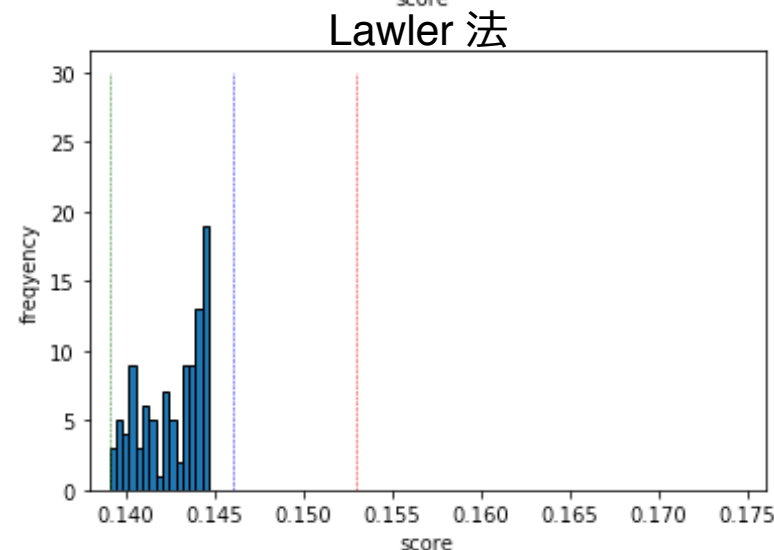
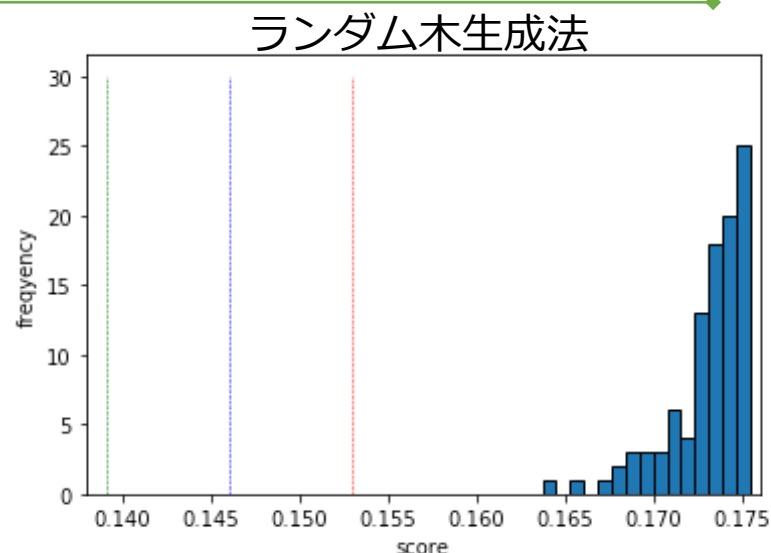
実験 1：提案アルゴリズムの比較

Adult データ上で、ランダム木生成法と Lawler法を比較した

- 求めた羅生門集合から、訓練誤差の分布、最小、最大、計算時間を調べた
- 誤差分布と最小最大**からは、ランダム木生成法より、**Lawler法の方がより小さい誤差の仮説**を見つけていた
- 計算時間**は、ランダム木生成法より、**Lawler法の方が x50 (Adult データの場合) 程度低速だった**

表1: 提案アルゴリズムの最小、最大誤差

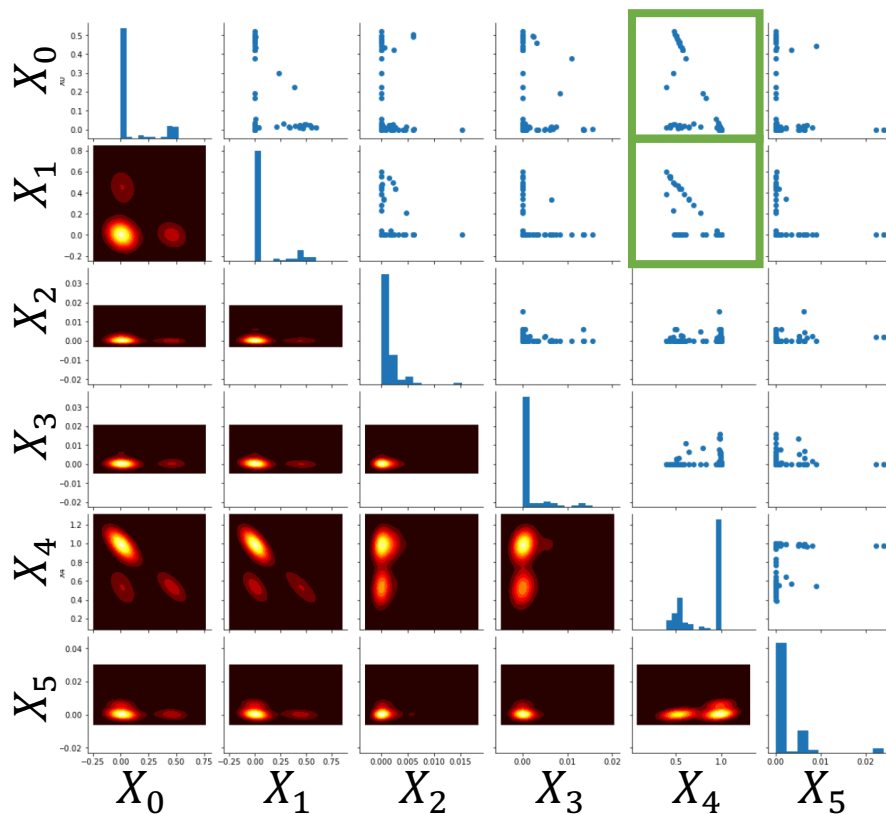
	ランダム木生成法	Lawler 法
基準決定木	0.139	0.139
最小誤差	0.164	<u>0.139</u>
最大誤差	0.175	<u>0.145</u>



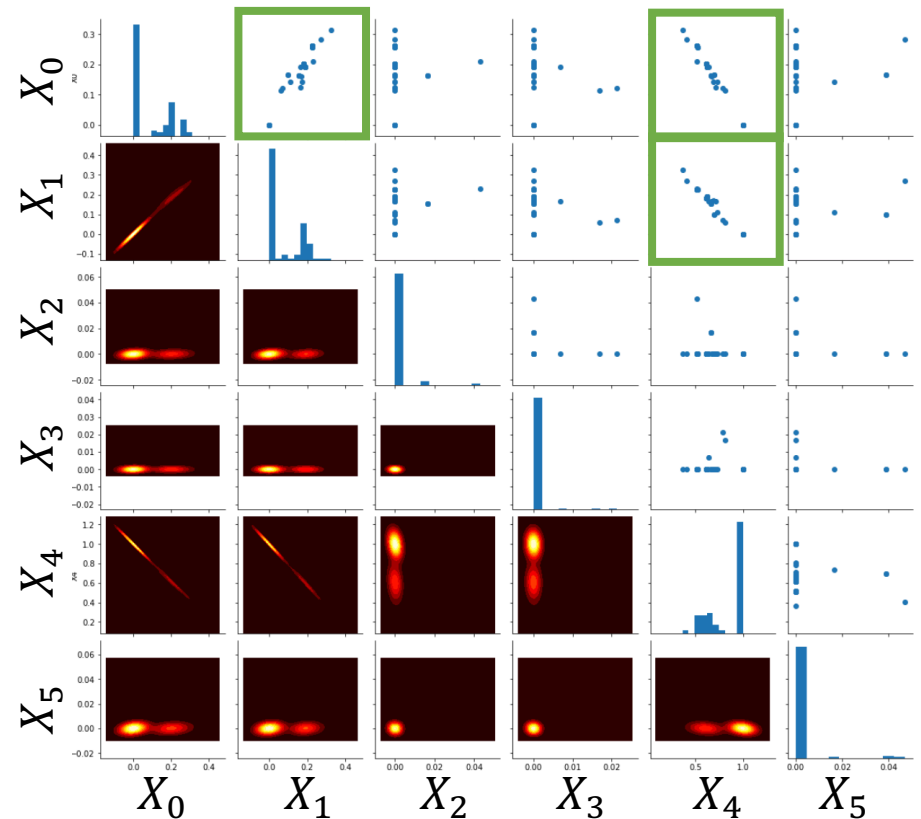
実験 2：実験結果（monk1 データ）

- 二つの特徴量について特徴量重要度の散布図とヒートマップを示す
- X_0, X_1 間の相関係数の値は **0.99** であった
- X_0, X_4 と X_1, X_4 間の相関係数の値は **-0.6** 以上であった

pair plot of gain importance for monk1 dataset by random



pair plot of permutation importance for monk1 dataset by random



実験 2：結果の考察（monk1 データ）

monk1 は $X_0 = X_1 \vee X_4 = 1$ のとき出力が 1 となり、
その他の場合、出力が 0 となるデータセットである

考察

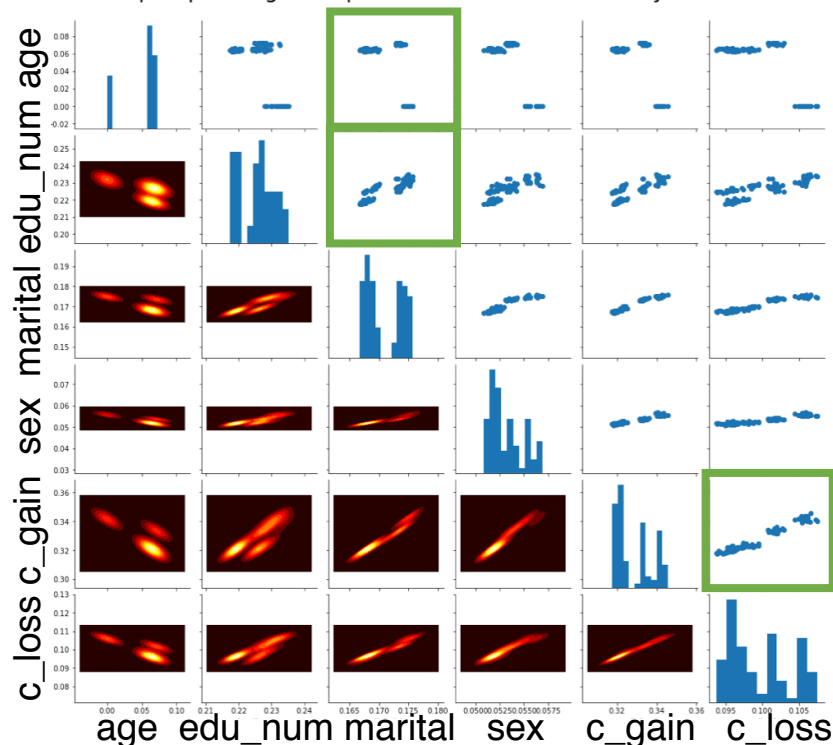
- X_0, X_1 間には**正の相関**があった
→ **同時に使うこと**が予測性能の向上に寄与することを示唆していると考えられる
- X_0, X_4 と X_1, X_4 間には**負の相関**があった
→ **いずれか片方の特徴量さえあれば予測には十分**であることを示唆していると考えられる

実験結果は、特徴量エンジニアリングや特徴量選択のための足掛かりとしてVICが有用であることを示している

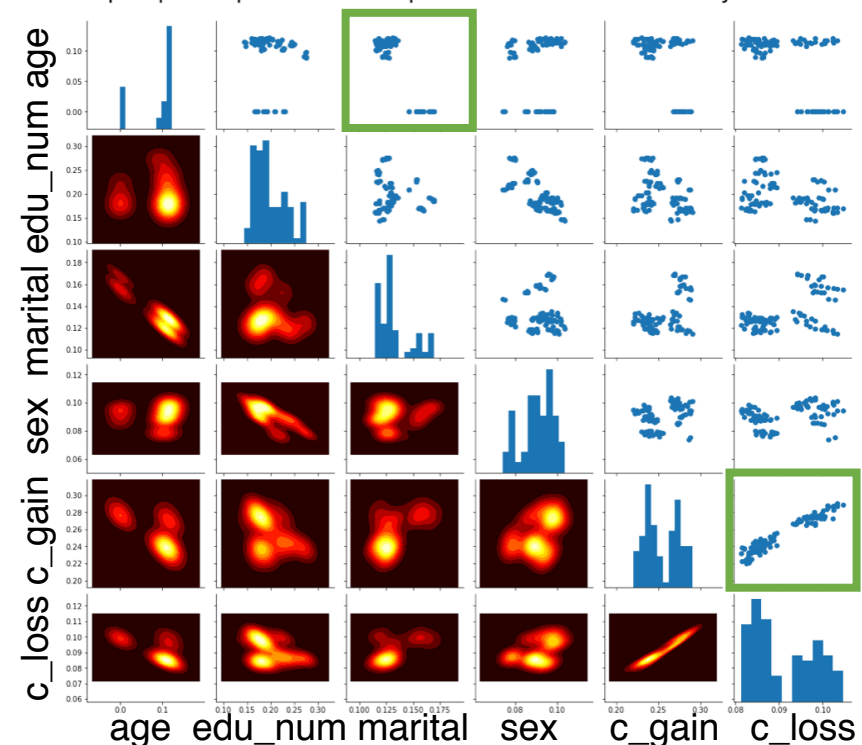
実験 3 : 実験結果 (Adult データ)

- 二つの特徴量について特徴量重要度の散布図とヒートマップを示す
- capital_gain, capital_loss 間の相関係数の値は **0.97** であった
- age, marital_status 間の相関係数は **-0.6** 以上であった
- education_num, marital_status 間の相関係数は **0.8** であった

pair plot of gain importance for adult dataset by lawler



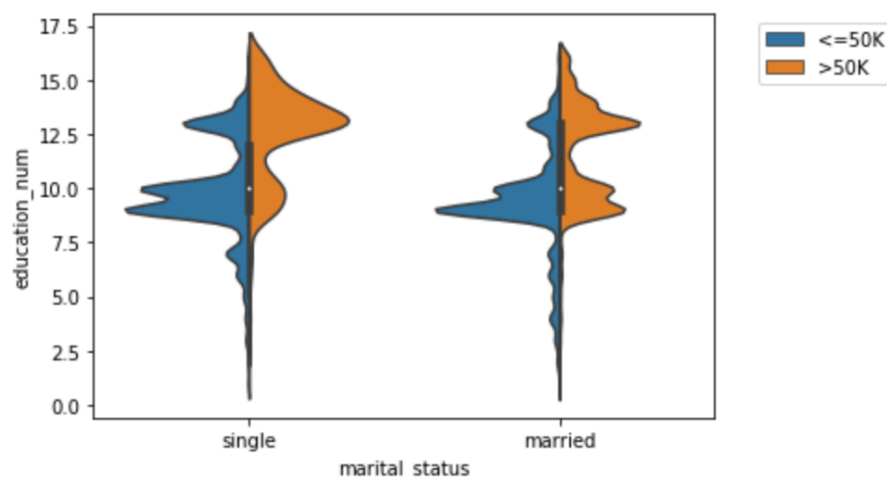
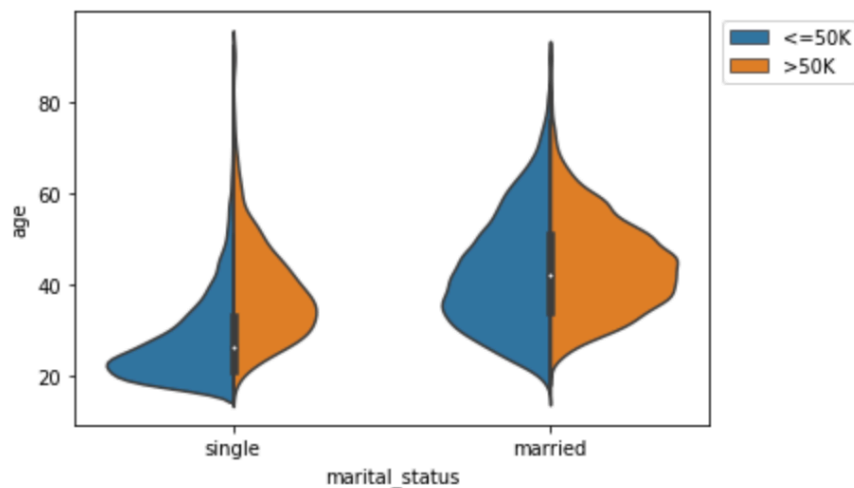
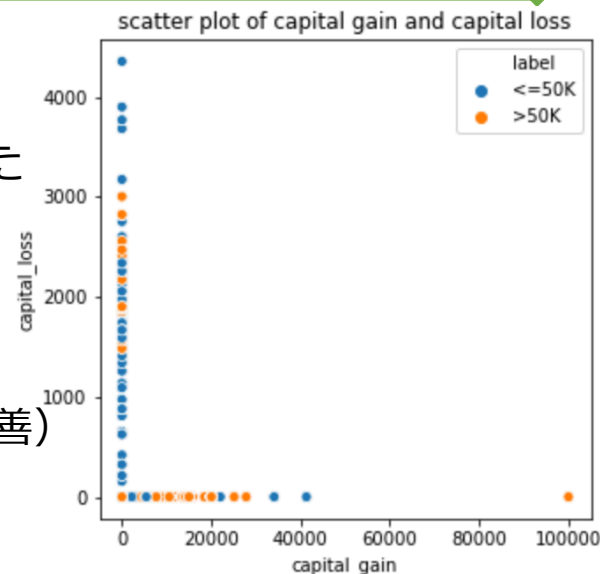
pair plot of permutation importance for adult dataset by lawler



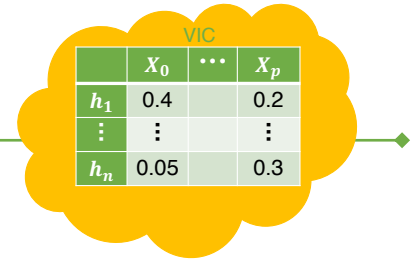
実験 3：結果の考察（Adult データ）

考察

- capital_gain, capital_loss 間に**正の相関 (0.99)**があった
→ 二つの特徴量を用いて、組合せ特徴量を作ることによって**単一で分割できる可能性がある**
- 実験2の仮説に合わない特徴量の組合せがあった
 - age, marital_status 間の**負の相関 (-0.6)**（組合せると改善）
 - education_num, marital_status 間の**正の相関 (0.8)**（組合せても非改善）。



まとめ



A yellow cloud-shaped graphic containing a table labeled 'VIC'. The table has columns X_0 , \dots , and X_p , and rows h_1 , \vdots , and h_n . The values in the table are 0.4, 0.2, 0.05, and 0.3.

	X_0	\dots	X_p
h_1	0.4		0.2
\vdots	\vdots		\vdots
h_n	0.05		0.3

- 決定木に対する VIC の構築方法を提案
- 実データセット上での VIC を構築する実験を行い、VIC の要約方法とその有用性を検証
 - 提案の二つの構築法の間では**計算時間と精度にトレードオフ**を観察
 - 重要度の相関に基づく特徴量の利用法に関して考察
- 今後の課題
 - **羅生門集合を求める厳密アルゴリズムの開発**
(最適決定木の学習アルゴリズム [OSDT, Hu+, NeurIPS 2019] の拡張)
 - **三個以上の特徴量の組に拡張**した要約・分析方法の開発
(多重検定補正が必要になる[LAMP, Terada+ PNAS 2013])