決定木アルゴリズム: 発展

経済学のための機械学習入門

川田恵介

Table of contents

交差推定

- Cross fitting
- “サンプル分割によるサブサンプルサイズ減少” を緩和
  - そこそこのサンプルサイズ $n \leq 50000$ で通常推奨される (Bischl et al. 2021)
- 格差/因果推論への応用においても重要
  - “すべての” 機械学習 (+ 因果/格差推定) の包括パッケージで実装されている

ポイント

- 誤差項 $u := Y - E_P[Y|X]$ 分布 (“データ固有”) が、推定されたモデルにも、評価用事例にも入り込む
  - 相関が生じ、正しく評価できない
- 誤差項分布が、Training/Validation データで無相関であれば OK
  -「役割の固定」は本質的ではない

交差推定

1. データをいくつか (2,5,10,20 など) に分割
2. 第 1 サブデータ **以外** を用いて予測モデルを試作
3. 第 1 サブデータに予測値を適用
4. 全てのサブデータに 2,3 を繰り返す

交差検証

- Cross validation

5. 交差推定で導出した予測値と実現値について、予測誤差を推定

数値例: 単純平均 VS 決定木 (深さ 2)

```
# A tibble: 6 x 3
```

```
  Group     Y     X
  <dbl> <dbl> <dbl>
1     1     6     3
2     1     7     1
3     2     4     3
4     2     5     2
5     3     4     1
6     3     4     1
```

数値例: **単純平均** VS **決定木** (**深さ** 2)

```
# A tibble: 6 x 5
  Group     Y     X PredMean PredTree
  <dbl> <dbl> <dbl>    <dbl>    <dbl>
1     1     6     3     4.25        4
2     1     7     1     4.25        4
3     2     4     3    NA          NA
4     2     5     2    NA          NA
5     3     4     1    NA          NA
6     3     4     1    NA          NA
```

数値例: **単純平均**

```
# A tibble: 6 x 5
  Group     Y     X PredMean PredTree
  <dbl> <dbl> <dbl>    <dbl>    <dbl>
1     1     6     3     4.25        4
2     1     7     1     4.25        4
3     2     4     3     5.25        6
4     2     5     2     5.25        6
5     3     4     1    NA          NA
6     3     4     1    NA          NA
```

数値例: **単純平均**

```
# A tibble: 6 x 5
  Group     Y     X PredMean PredTree
  <dbl> <dbl> <dbl>    <dbl>    <dbl>
1     1     6     3     4.25        4
```

```
2     1     7     1     4.25        4
3     2     4     3     5.25        6
4     2     5     2     5.25        6
5     3     4     1     5.5         7
6     3     4     1     5.5         7
```

数値例: 単純平均

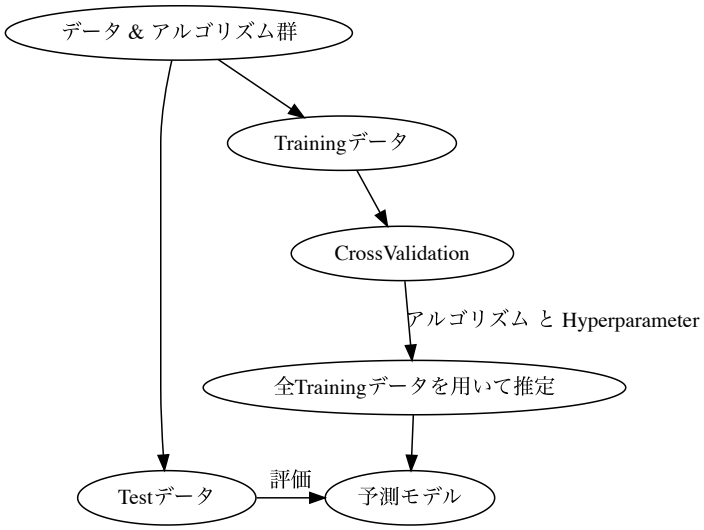
```
# A tibble: 6 x 7
  Group     Y     X PredMean PredTree ErrorMean ErrorTree
  <dbl> <dbl> <dbl>    <dbl>    <dbl>     <dbl>     <dbl>
1     1     6     3     4.25        4    3.06           4
2     1     7     1     4.25        4    7.56           9
3     2     4     3     5.25        6    1.56           4
4     2     5     2     5.25        6    0.0625         1
5     3     4     1     5.5         7    2.25           9
6     3     4     1     5.5         7    2.25           9
```

- 平均二乗誤差 (Mean) 2.79
- 平均二乗誤差 (Tree) 6

トレードオフの緩和

- サンプル分割法では、Training データに多くの事例を割くと、Validation データに割ける事例が減り、評価の精度が下がる (推計誤差の拡大 $\iff$ Validation データへの依存)
- 交差検証では、すべての事例について予測値を計算し、その平均を取るので、評価の精度を確保できる
- 理論的検討: アルゴリズムの相対比較について有効 (Wager 2019)
  - 最終的な予測モデルの性能検証には使えない

予測研究の典型的ワーク



正則化

- Hyperparameters $\simeq$ EmpricialRisk 最小化では決定できないパラメータ
- 決定木については、木の深さ、最小サンプルサイズ、“剪定度合い” などなど

剪定

- 最大分割回数は、自然な Hyper parameter だが、、、
- 浅い木は、将来の重要な分割を見逃してしまう可能性がある
- 剪定: 一旦非常に深い木を推定 (Approximation error を減らす) した後に、単純化 (正則化) を行う
  - 重要ではないサブグループについて、再結合

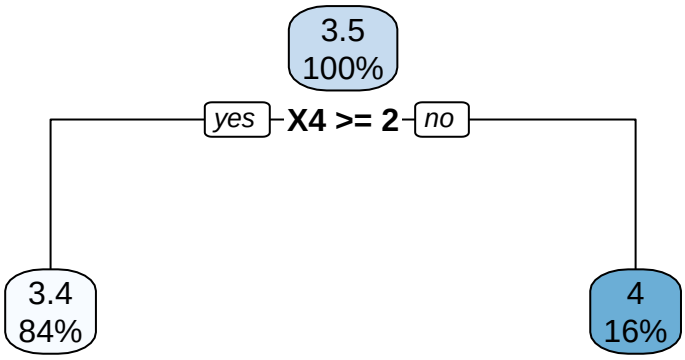
Step 1. 深い木の推定

- 停止条件を緩めると、一般にどこまでもサブサンプル分割が行われる
  - 平均値が異なるサブグループが見つかる限り止まらない

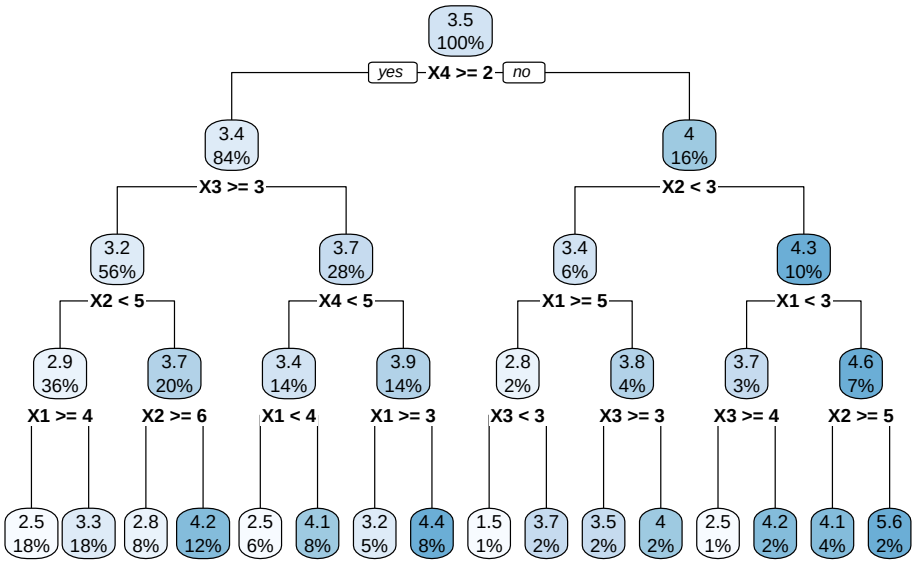
数値例: サイコロゲーム

- ディーラーは、サイコロを 5 つふり、4 つ $(X_1, .., X_4)$ プレイヤーに見せる
  - プレイヤーは残り一つの出目 Y を予測
- サイコロの出目は、uniform 分布 (完全無相関) に決定
  - 理想の予測モデル $g(X_1, .., X_4)$
- “見” を 200 回行いデータ収集

例


3.5
100%
yes
X4 >= 2
no
3.4
84%
4
16%

例


3.5
100%
yes
X4 >= 2
no
3.4
84%
4
16%
X3 >= 3
X2 < 3
3.2
56%
3.7
28%
3.4
6%
4.3
10%
X2 < 5
X4 < 5
X1 >= 5
X1 < 3
2.9
36%
3.7
20%
3.4
14%
3.9
14%
2.8
2%
3.8
4%
3.7
3%
4.6
7%
X1 >= 4
X2 >= 6
X1 < 4
X1 >= 3
X3 < 3
X3 >= 3
X3 >= 4
X2 >= 5
2.5
18%
3.3
18%
2.8
8%
4.2
12%
2.5
6%
4.1
8%
3.2
5%
4.4
8%
1.5
1%
3.7
2%
3.5
2%
4
2%
2.5
1%
4.2
2%
4.1
4%
5.6
2%

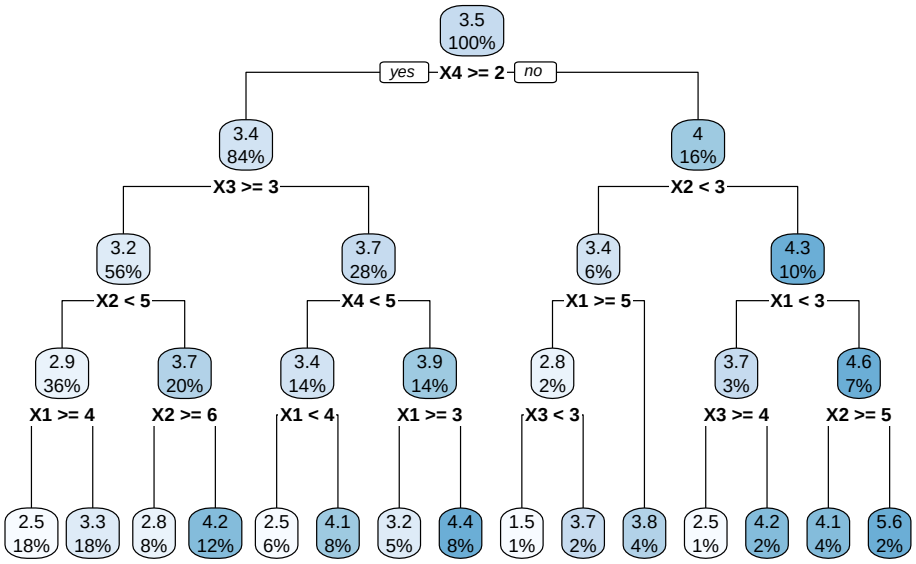
Setp 2. **剪定**

- 以下を最小化するようにサブグループを再結合

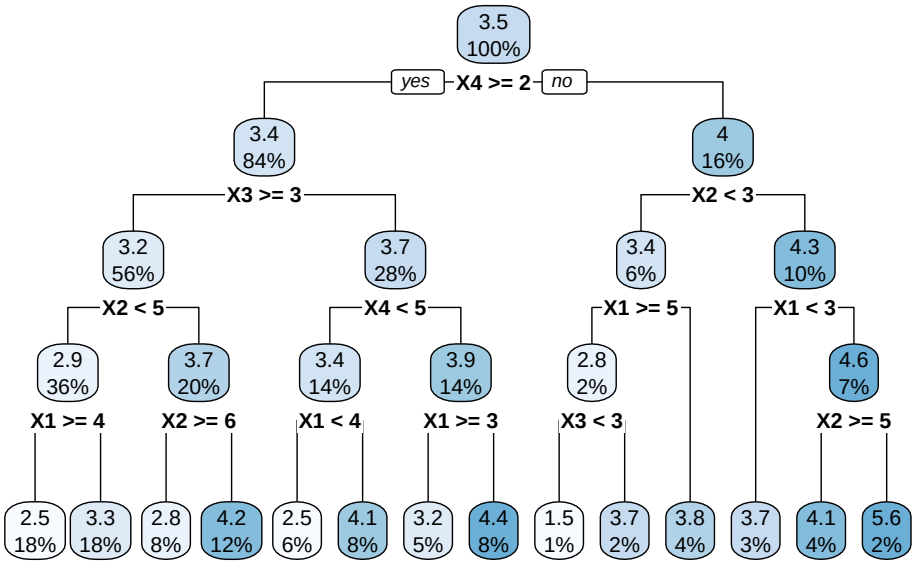
$$EmpiricalRisk(MeanSquiaredError) + \lambda \times |T|$$

- λ : Hyper Parameter (rpart 関数では cp)
  - 交差推定で選択
- 分割しても平均二乗誤差があまり減らないサブグループから結合していく

例: **剪定**



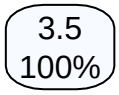
例: 剪定



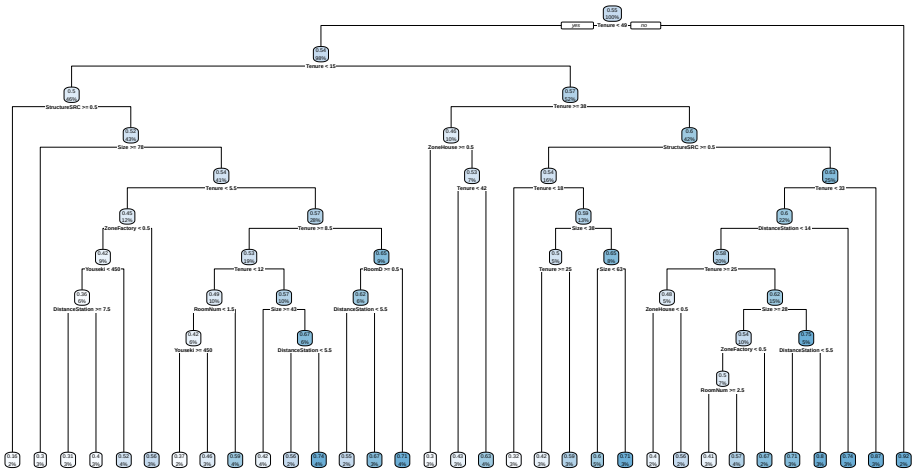
Tuning space

- 多くのアルゴリズムは、複数の Hyper paramter を持つ
  - 有界の範囲から探す必要がある
  - どの範囲で探すか?
- mlr3tuningspaces
  - λ (cp) , 最小サンプルサイズ (minsplit), 分割を試みる最小サンプルサイズ (minbucket) を交差推定で最適化

例: 交差推定で生成される決定木



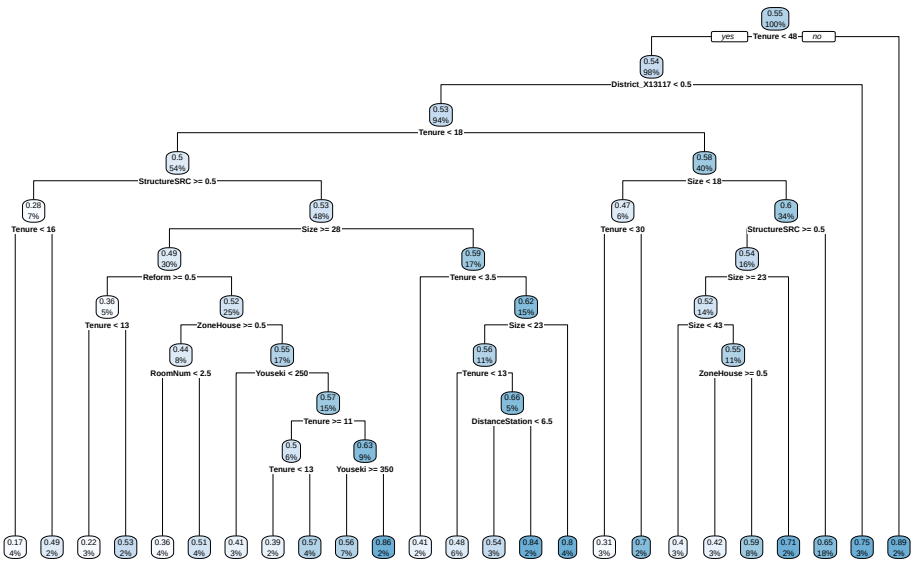
実例: 2000 事例で取引年予測 (シード値 1)



まとめ

- Approximation error の削減は、現代的な PC + アルゴリズムであれば容易
  - 複雑にすればいいだけ!!!
- モデルを適切に単純化 (HyperParameter を適切に選択) することで、Estimation error を削減する (正則化) に工夫が必要
- 正則化を行ったとしても、一般に決定木の EstimationError は大きい
  - 対策: モデル集計 (RandomForest)

実例: 2000 事例で取引年予測 (シード値 2)



補論: 最適化

- 本講義では、Random Search を使用
  - 複雑なシステムについての最適化は、長年の研究課題
  - mlr3tuning (mlr3verse に同梱) では、Grid Search や Iterated Racing なども実装
  - より発展的なアルゴリズムも mlr3mbo (baysian optimization) や mlr3hyperband (hyperband) で実装

  - Hyper parameter のスペースの具体例は、mlr3tuningspace (mlr3verse に同梱) で提案

- サーベイ: Bischl et al. (2021) (mlr3verse の author も含む)

余談: 良性の過剰適合

- 剪定などによる推定パラメタの削減は、教師付き学習の伝統的戦略
  - 伝統的な実証研究でも、研究者が頑張ってやっていた
- パラメタを大幅に増やす (サンプルサイズを超える) と、過剰適合が" 減り!!¡‘、予測性能が改善する場合がある　(Bartlett et al. 2020; Hastie et al. 2022)
  - Benign overfitting

Reference

Bartlett, Peter L, Philip M Long, Gábor Lugosi, and Alexander Tsigler. 2020. “Benign Overfitting in Linear Regression.” *Proceedings of the National Academy of Sciences* 117 (48): 30063–70.

Bischl, B., Martin Binder, Michel Lang, Tobias Pielok, Jakob Richter, Stefan Coors, Janek Thomas, et al. 2021. “Hyperparameter Optimization: Foundations, Algorithms, Best Practices, and Open Challenges.” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 13.

Hastie, Trevor, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. 2022. “Surprises in High-Dimensional Ridgeless Least Squares Interpolation.” *The Annals of Statistics* 50 (2): 949–86.

Wager, Stefan. 2019. “Cross-Validation, Risk Estimation, and Model Selection.” *arXiv: Methodology*.