予測問題と予測木

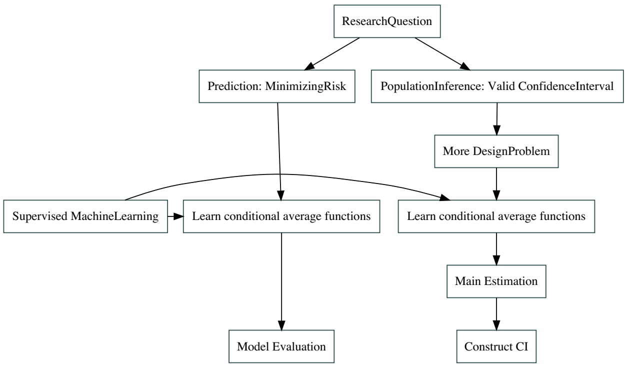
機械学習の経済学への応用

川田恵介

本スライドの内容

- 母平均関数に" 適合する関数" $f(X)$ を推定するアルゴリズムとして、予測木 (回帰木 | 分類木) アルゴリズムの紹介
- Motivation として予測問題の概論を紹介
  - 母集団の推論問題は後日
- 比較対象として、Naive なアルゴリズムも紹介

全体像



予測: 一般問題

- 教師付き学習の予測問題への応用を紹介

典型的問題設定

- データ $\{Y, X = [X_1, .., X_L]\}$ が活用可能
  - ランダムサンプリング元の母集団を想定
  - **同じ**母集団から**新たに**抽出された事例について、Y を予測
- データから Y の予測モデル $f(X)$ を推定 (学習)

例

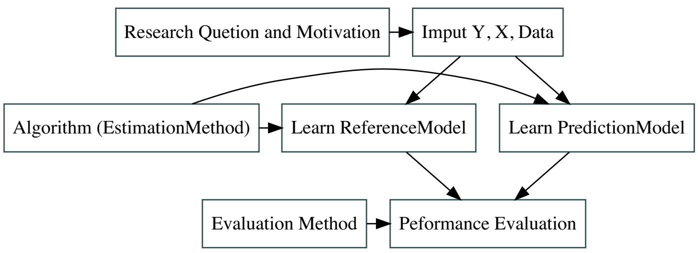
- 需要予測: $X =$ 店舗の属性、気象予測、カレンダー, $Y =$ 販売量
- 皮膚癌: $X =$ 写真、$Y =$ 犬 $|$ 猫

- 滞納予測: $X =$ 個人属性、$Y =$ 返済を滞納するかどうか
- キャッチーな議論: 予測するマシンの世紀

経済学における応用例

- 「新しいアルゴリズムを用いると、予測性能がこのくらい改善する」的な研究は少ない
  - 研究動機を工夫したものが多い
- 1 年後生存の予測 (Einav et al. 2018)
  - 「終末期医療論争」の前提条件は成り立っているのか？
- 経済モデルの評価 (Fudenberg et al. 2022)
  - 「構造モデル」の評価

Standard Prediction RoadMap



理想的かつ実現不可能な評価

- 論点整理に有益

- 理想的な評価は、既知の損失関数 L についての母平均

$$E[L(Y, f(X_1, .., X_L))]$$

- よく用いられるのは、二乗誤差

$$L = (Y - f(X_1, .., X_L))^2$$

含意

$$E[(Y, f(X))^2] = \underbrace{E[(Y - \mu_Y(X))^2]}_{Irreducibel=\text{個人差}}$$

$$+ \underbrace{E[(\mu_Y(X) - f(X))^2]}_{Reducible}$$

- ただし $\mu_Y(X) = E[Y|X]$
- 最善の予測モデル: $f(X) = \mu_Y(X)$

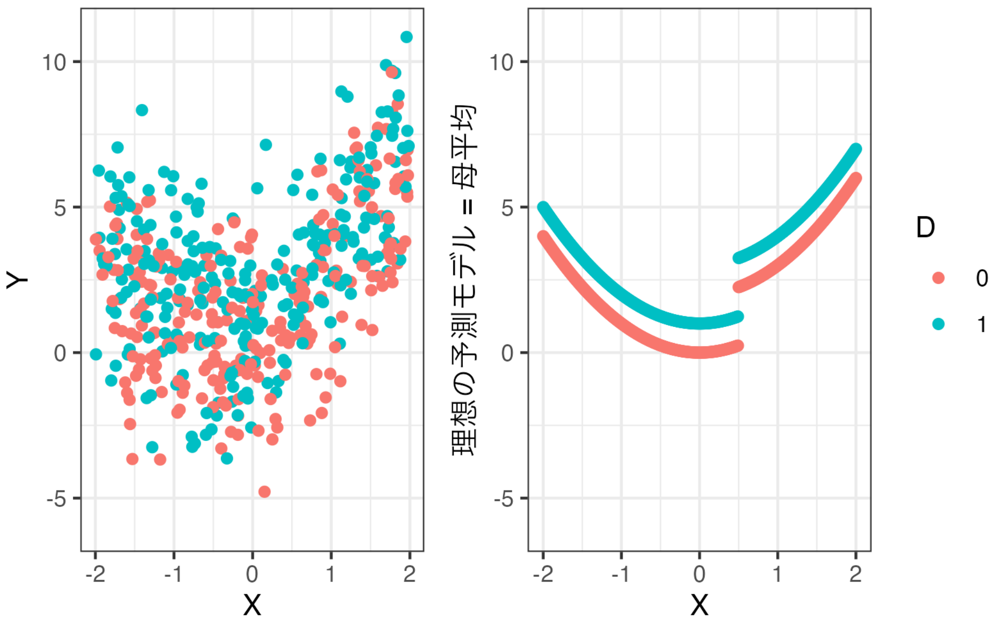
含意

- 母集団上で定義される評価を、データ上でどのように行うか？
  - AIC|BIC などの活用, **サンプル分割**
  - 後日
- 予想誤差 $= Y - f(X) = \underbrace{Y - \mu_Y(X)}_{Irreducible} + \underbrace{\mu_Y(X) - f(X)}_{Reducible}$ をどのように削減するか？
  - Irreducible: 有効な予測変数 X が活用できるデータの探索
  - Reducible: Algorithm の改善

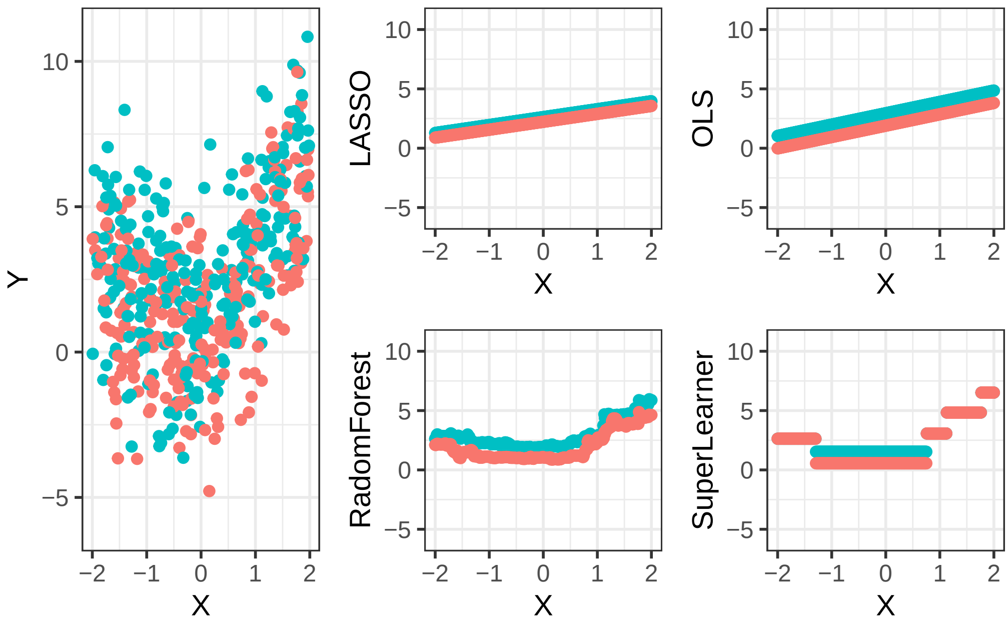
Algorithm

- 推定手順の大枠
- データを予測モデルに変換
- 母平均に近い予測モデルを" 得やすい" アルゴリズム = 優れたアルゴリズム

数値例: 理想のアルゴリズム



数値例: 実際のアルゴリズム



まとめ

- 母平均が最善の予測モデル
- 頑張って母平均を推定する

Naive algorithm

- 単純平均法と丸暗記法

単純平均法

- 全データについての平均値

$$f(x) = \sum_i Y_i / N$$

- X は完全無視だが、大量の事例について平均を取れる

丸暗記法

- 全く同じ X の値を持つ事例についての平均値
  - "最も近い" X の事例について平均値

$$f(x) = \sum_{i|X_i \simeq x} Y_i / N_{X_i \simeq x}$$

- 一般に、少数事例について平均

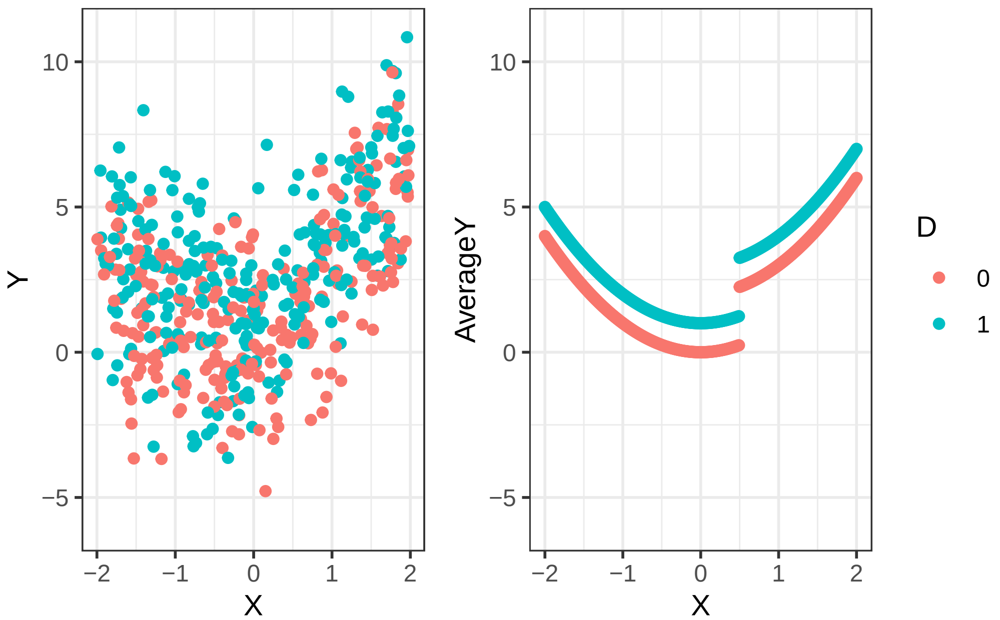
数値例

- $\{D, X\}$ から Y を予測
- データ生成プロセス

$$Y = D + 2 \times I(X >= 0.5) + X^2 + u$$

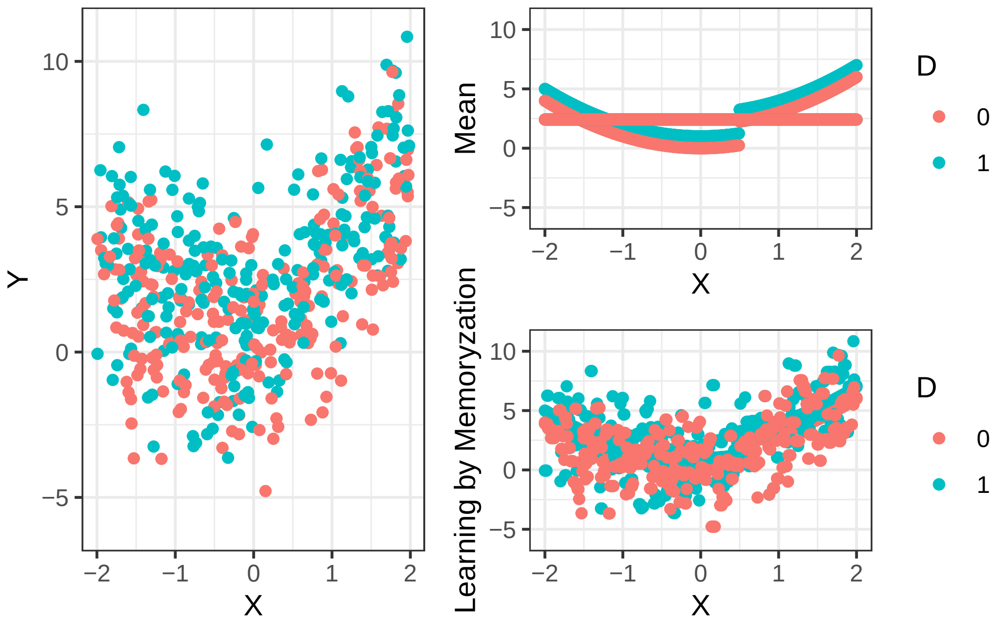
- $\Pr[D = 1] = \Pr[D = 0] = 0.5$, $X \sim U(-2, 2)$, $u \sim N(0, 2)$
- 理想の予測モデル: $f(D, X) = D + 2 \times I(X >= 0.5) + X^2$

数値例



Y
X
AverageY
X
D
0
1

数値例



Y
X
Mean
X
D
0
1
Learning by Memoryzation
X
D
0
1

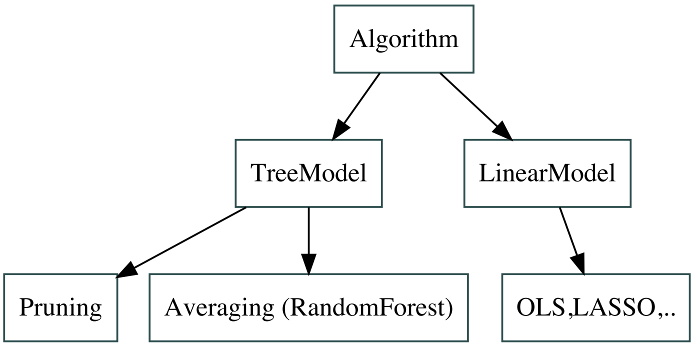
まとめ

- 単純平均法の問題点: “一定” の予測値を決めうつ、荒い近似
  - $E[Y|X]$ と X との関係性を完全無視
- 丸暗記法の問題点: 平均値の推定に、“個人差” が強く反映
  - $X = \{1994年7月4or5日生まれ、男性、岩手県出身\}$ の予測年収は?
  - データにおける最も近い事例が、大谷翔平だと？？？
- 予想: 中間的 Algorithm が良さそう

予測木アルゴリズム

- 非常に” 透明性が高く” 教育的なアルゴリズム
  - コンセプトが明快、モデルが可視化できる場合も
  - 重要な論点を抑えられる

全体像



予測木アルゴリズム

- サブグループの" 平均値" を予測値とする
  - 伝統的方法: 人間がサブグループを決定
  - 本講義: データがサブグループを決定
- トリビア: $Y=$ 連続 であれば回帰木、$Y=$ 離散 であれば分類木|決定木 と呼ばれる

伝統的方法

- データを見る前に推定する (有限個のパラメータからなる) 予測 (母平均) モデルを設定
  - パラメータのみをデータによって決める
- 例:

$$f(D,X)=\beta_1\times I(D=1,X\le 0)+\beta_2\times I(D=1,X>0)$$

$$+\beta_3 \times I(D=0, X \leq 0) + \beta_4 \times I(D=0, X > 0)$$

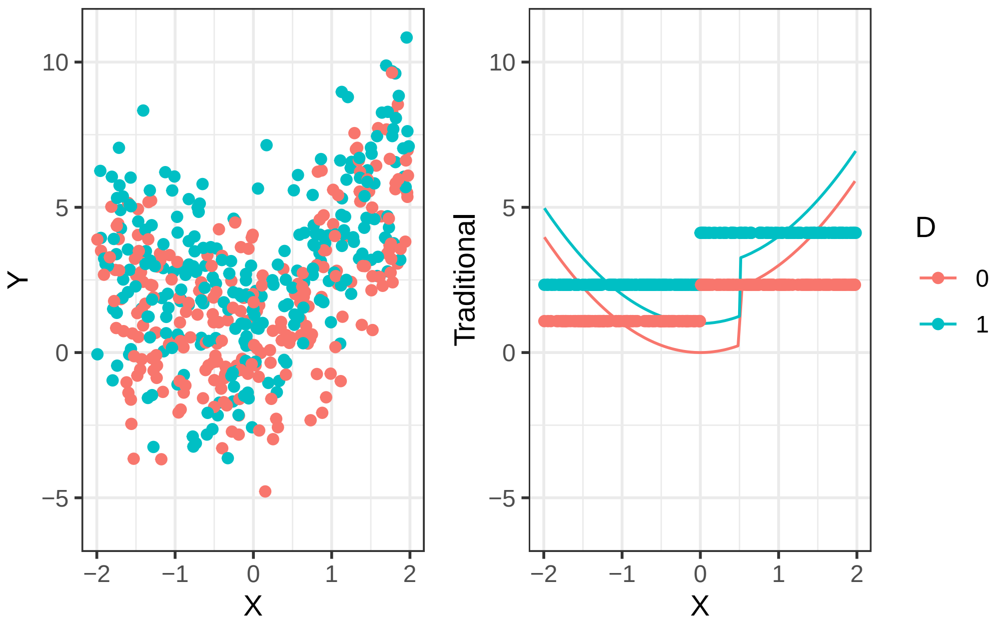
推定方法: Empirical Risk Minimization

- データ上の Loss を最小化するように推定: $L =$ 二乗誤差 であれば、

$$\beta = \arg\min_{\beta} \sum_i (Y_i - f(X_i))^2$$

- 伝統的アプローチでは、OLS | サブサンプル平均と一致

数値例



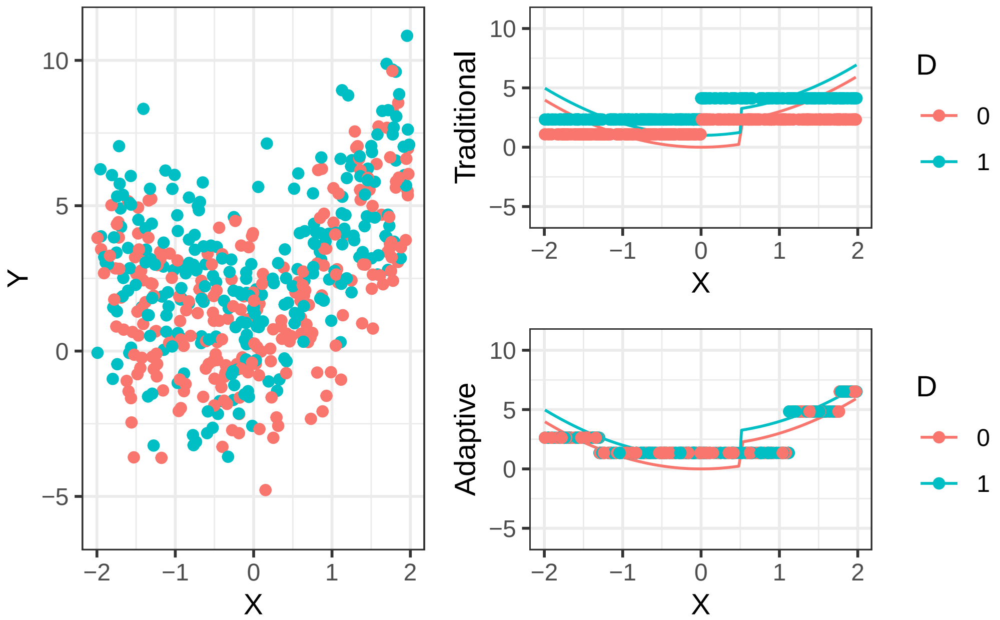
Adaptive Tree

- 伝統的方法: 分析者によるモデル設定 + Empirical Risk Minimization による推定
  - モデル設定にパフォーマンスが大きく依存
  - 適切なサブグループ分けは非常に困難
- Adaptive な推定: サブグループ分けにも、Empirical Risk Minimization を活用

Recursive Partition アルゴリズム

1. データ、**停止条件 (最大分割回数等)**
2. 第 1 分割: Empirical Risk を最小化するグループ分割 (通常 2 分割) を探索
3. 第 2 分割: 第 1 分割の結果を**所与**として、Empirical Risk を最小化するグループ分割を探索
4. 停止条件に達するまで、分割を繰り返す

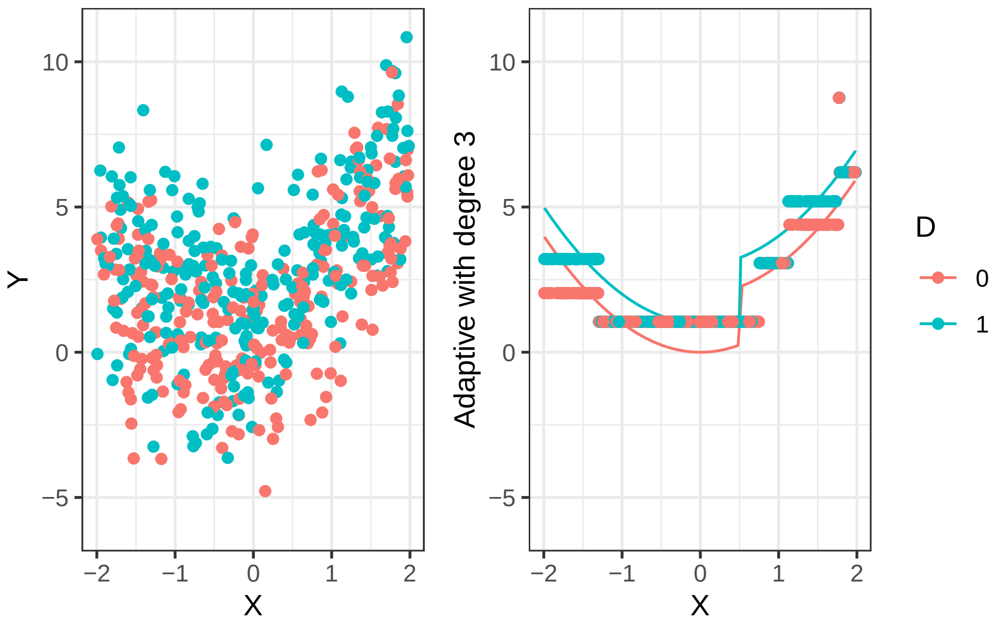
数値例



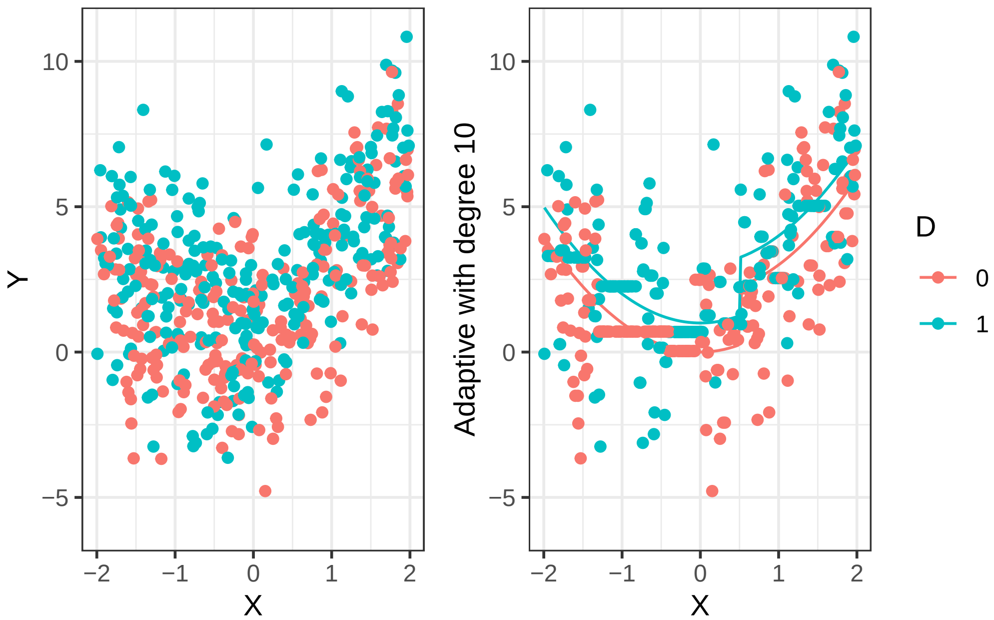
停止条件

- 停止条件をどう決める？
  - rpart 関数での初期値: 最小サンプルサイズ = 20, 最大分割数 = 30 など
- 推定されたモデルやパフォーマンスが決定的に左右される
- Naive な Idea
  - 現実は複雑なので、単純なモデルはよくない
  - Empirical Risk Minimization を適用

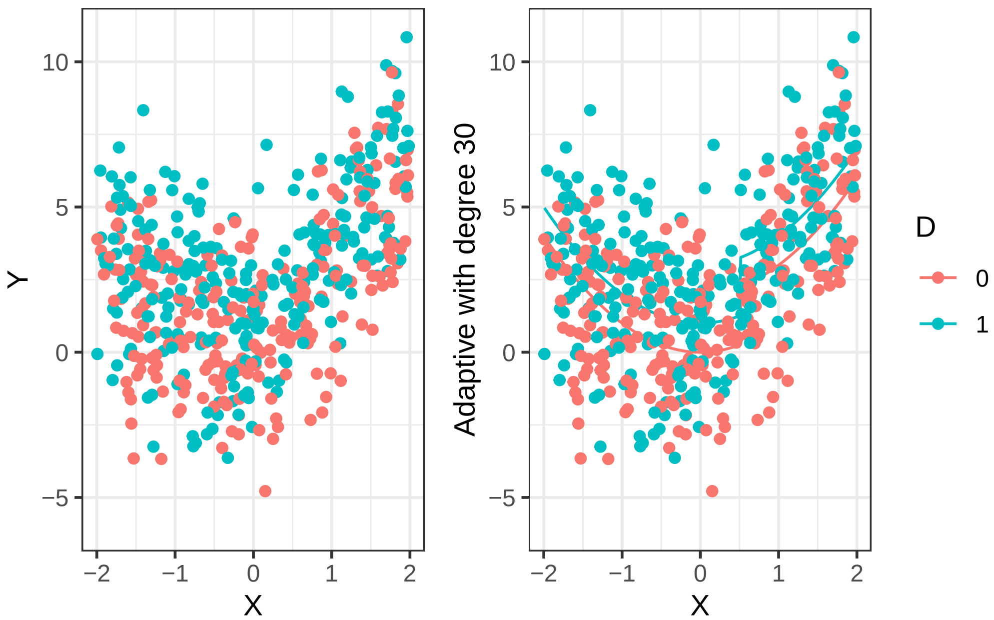
例


10
5
0
−5
Y
−2
−1
0
1
2
X
Adaptive with degree 3
D
0
1

例


10
5
0
−5
Y
−2
−1
0
1
2
X
Adaptive with degree 10
D
0
1

例



まとめ

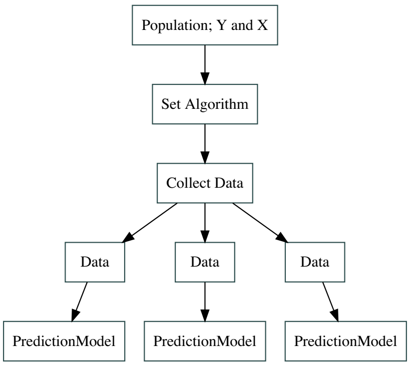
- 伝統的な予測木 = 有限個のパラメータを推定
  - 多くの潜在的パラメータを 0 と決めうち
- Adaptive な予想木 = サブグループをデータに合うように生成
  - 潜在的に無限個のパラメータを推定
- 停止条件に決定的な影響を受ける

過剰適合 (過学習) 問題

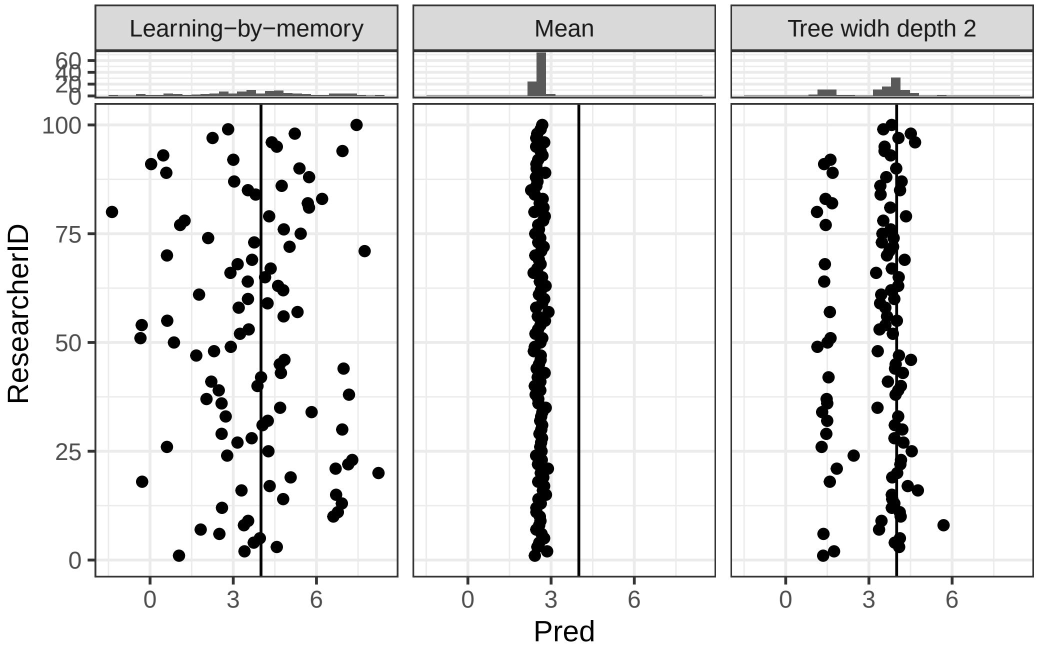
とりあえず頭に入れること

- ある母集団を予測する上で、優れたアルゴリズム (停止条件の設定を含む) を選びたい
  - 常にうまくいくアルゴリズムは存在しない
- 次善の策は、上手く**いきやすい**アルゴリズムを用いてる
- 通常の統計学と同じ脳内モデルが有益

脳内モデル


Population; Y and X
Set Algorithm
Collect Data
Data
Data
Data
PredictionModel
PredictionModel
PredictionModel

数値例



Decomposition

- $f_n(X)$ 無限大のサンプルサイズで学習した結果得られる "仮想的な" 予測モデル

$$Y - f(X) = \underbrace{Y - \mu_Y(X)}_{IrreducibleError} + \underbrace{\mu_Y(X) - f(X)}_{ReducibleError}$$

$$= \underbrace{Y - \mu_Y(X)}_{IrreducibleError} + \underbrace{\mu_Y(X) - f_\infty(X)}_{ApproximationError} + \underbrace{f_\infty(X) - f(X)}_{EstimationError}$$

トレードオフ

- モデルを複雑化 (より多くの分割) を行うと、現実は極めて複雑なので

$$Y - f(X) = \underbrace{Y - \mu_Y(X)}_{IrreducibleError} + \underbrace{\mu_Y(X) - f_\infty(X)}_{ApproximationError\downarrow} + \underbrace{f_\infty(X) - f(X)}_{EstimationError\uparrow}$$

Estimation Error の源泉

- $Y = \underbrace{\mu(X)}_{Signal} + \underbrace{u}_{Y-\mu(X):Noise}$

- Signal のみを取り出せる人がいれば、全て解決
  - 目の前の香川出身 38 歳男性の所得を聞き、香川出身 38 歳男性の平均所得と個人差を分割できる？
- 伝統的戦略は、大量の事例の平均を取る
  - 漸近性質の活用
- 複雑なモデルは、
  - u の影響を強く受け、Estimation Error が上昇する

例

- 単純なモデル

$$\sum_{i|D_i=1} Y_i/N_{i|D_i=1} = \underbrace{\mu_Y(D_i=1)}_{Larger\ Approximation\ Error} + \sum_{i|D_i=1} u_i/N_{i|D_i=1}$$

- 複雑なモデル

$$\sum_{i|D_i=1;X_i=1} Y_i/N_{i|D_i=1\&X_i=1} = \mu_Y(D_i=1;X_i=1) + \underbrace{\sum_{i|D_i=1;X_i=1} u_i/N_{i|D_i=1;X_i=1}}_{Larger\ Estimation\ Error}$$

Empirical Risk Minimization の問題

- 理想は母集団上で Risk を最小化する

$$\min E[(Y_i - f(X_i))^2] = E[(\mu_Y(X_i) + \underbrace{u_i - f(X_i)}_{Indepdent})^2]$$

- できないので Empirical Risk Minimization

$$\sum_i (Y_i - f(X_i))^2 = \sum_i (\mu_Y(X_i) + \underbrace{u_i - f(X_i)}_{GenerallyCorrelated})^2$$

過剰適合 (過学習)

- $f(X_i)$ が u_i の影響を**受けるほど**、小さくなる
  - 丸暗記モデルでは 0!!!!
- 一般にデータと無矛盾なモデルを推定することは難しくない
  - データに過剰に適合する (から過剰に学んだ) モデルであり、予測性能は悪い

- 新しい論点: Benign Overfiting (Hastie et al. 2022; Bartlett et al. 2020)

過剰適合 (過学習) の弊害

- データの" 偶然の偏り" (平均値からの大きな乖離があるサンプル) に影響されてしまう
- Empirical Risk では正しく評価できない
  - 丸暗記でも勉強することはいいことだ！！！

まとめ

- 通常の統計学と同様に Sampling Uncetainly を頭に入れる必要がある。
- 矛盾する戦略
  - 大量の事例の平均をとる (Estimation error の削減)
  - 多くのサブグループを作る (Approximation error の削減)

引用

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. https://doi.org/10.1073/pnas.1907378117.

Einav, Liran, Amy Finkelstein, Sendhil Mullainathan, and Ziad Obermeyer. 2018. "Predictive Modeling of u.s. Health Care Spending in Late Life." *Science* 360 (6396): 1462–65. https://doi.org/10.1126/science.aar5045.

Fudenberg, Drew, Jon Kleinberg, Annie Liang, and Sendhil Mullainathan. 2022. "Measuring the Completeness of Economic Models." *Journal of Political Economy* 130 (4): 956–90. https://doi.org/10.1086/718371.

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. https://doi.org/10.1214/21-AOS2133.