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

- 母平均を” 学習した関数” $f(X)$ 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
- 比較対象として、Naive なアルゴリズムも紹介

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

- 母集団を想定
  - ランダムサンプリングされたデータ $\{Y, X_1, .., X_L\}$ が活用可能
- データから Y の予測モデル $f(X_1, .., X_L)$ を推定 (学習)
  - **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

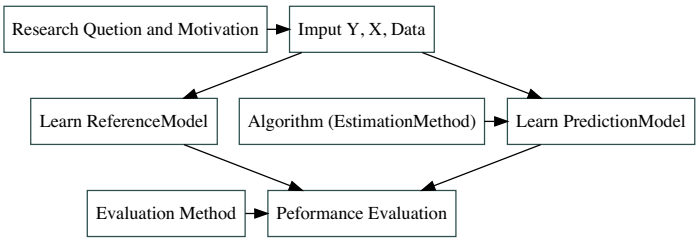
例

- 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 Research RoadMap



理想的だが非現実的な評価

- 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

$$E[(Y, f(X_1, .., X_L))^2] = \underbrace{E[(Y - \mu_Y(X_1, .., X_L))^2]}_{Irreducibel=個人差}$$

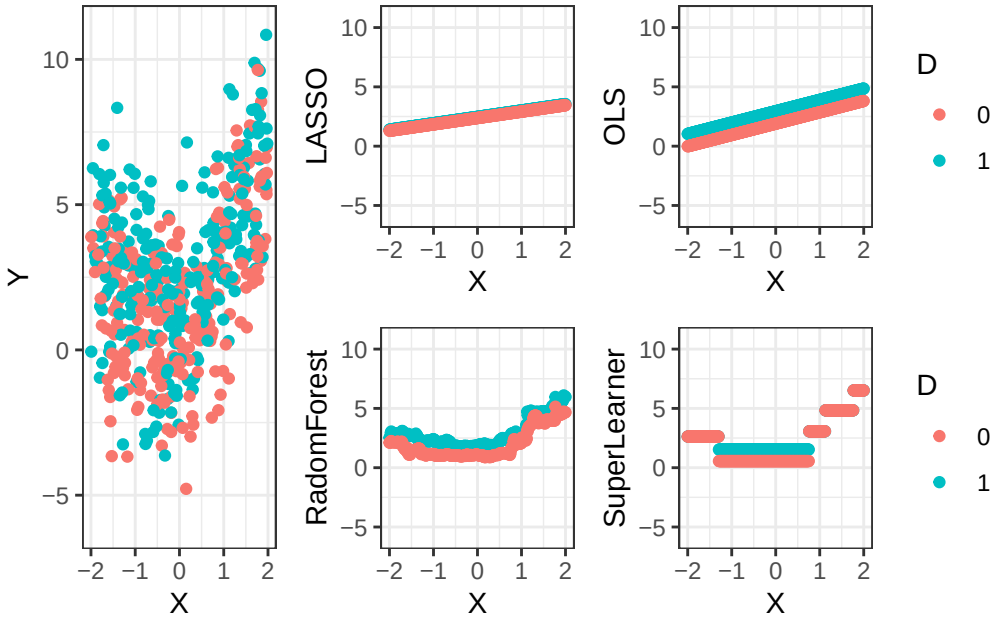
$$+ \underbrace{E[(\mu_Y(X_1, .., X_L) - f(X_1, .., X_L))^2]}_{Reducible}$$

- ただし $\mu_Y(X_1, .., X_L) = E[Y|X_1, .., X_L]$
- 最善の予測モデル: $f(X_1, .., X_L) = \mu_Y(X_1, .., X_L)$

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**
- 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

例



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

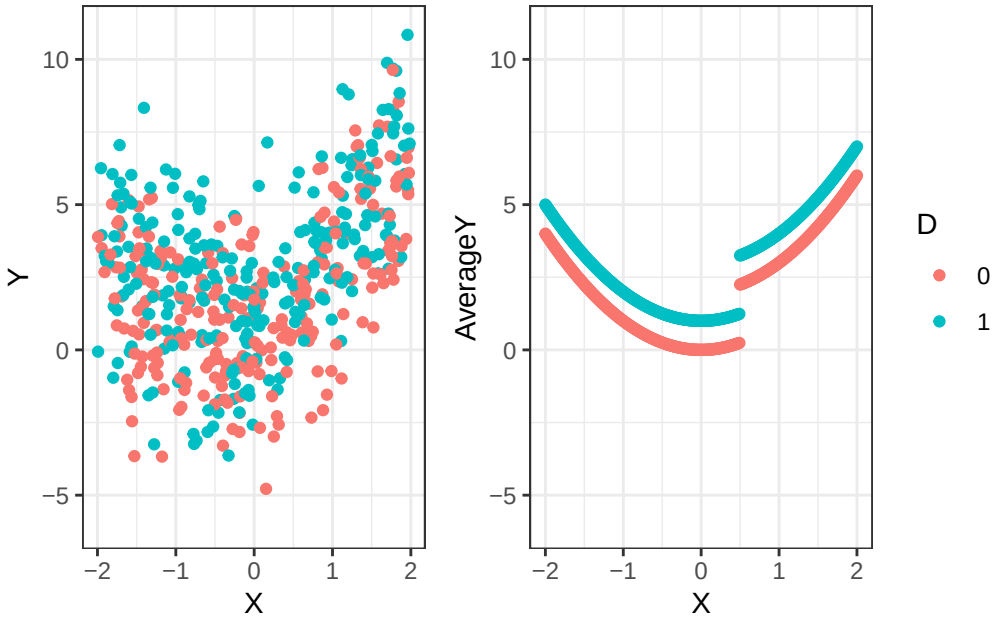
数値例

- $\{D, X\}$ から Y を予測
- データ生成プロセス

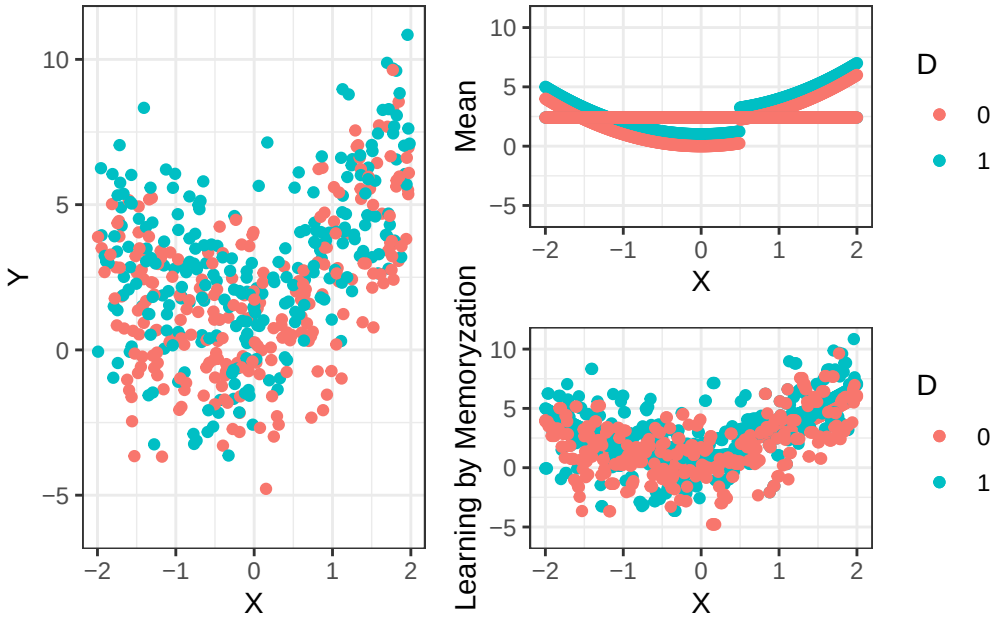
$$Y = D + 2 \times I(X >= 0.5) + X^2 + u$$

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

数値例



数値例

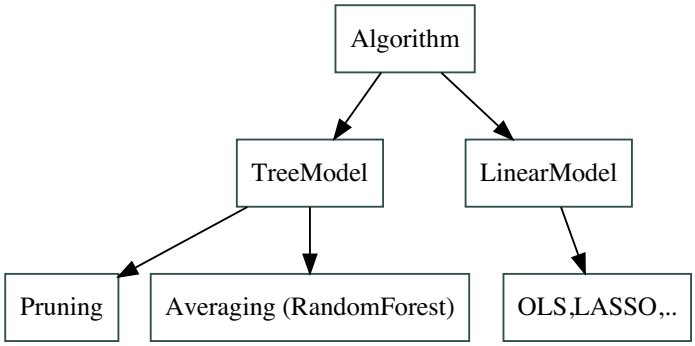


Misspecification VS Variance

- 単純平均法の問題点: “一定” の予測値を決めうち
  - $E[Y|X]$ 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” を強く反映
  - $X = \{1994$年7月$4 or 5$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

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:

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

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

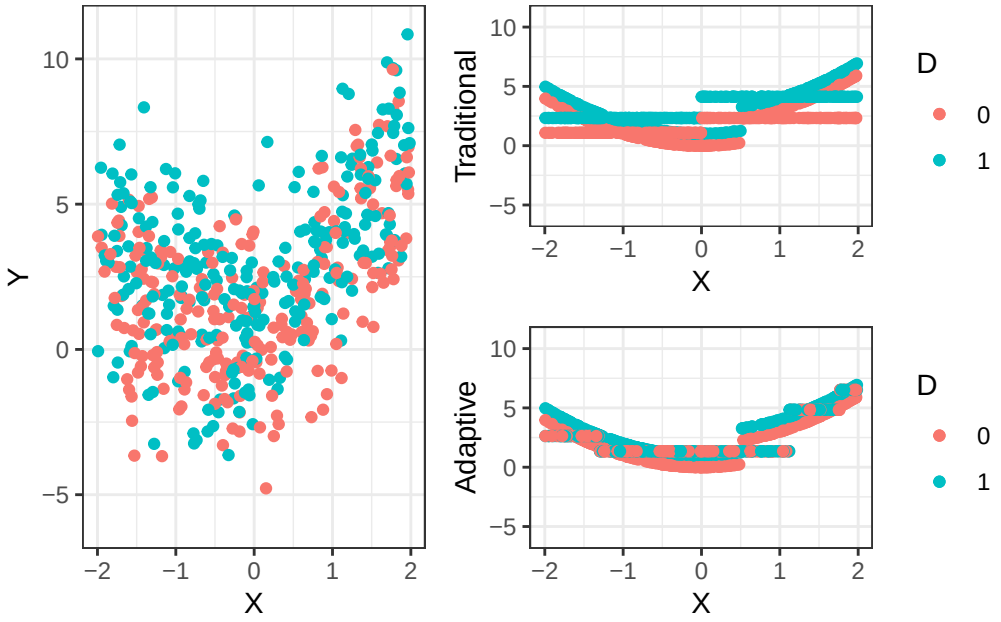
推定方法: Empirical Risk minimization

- 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、

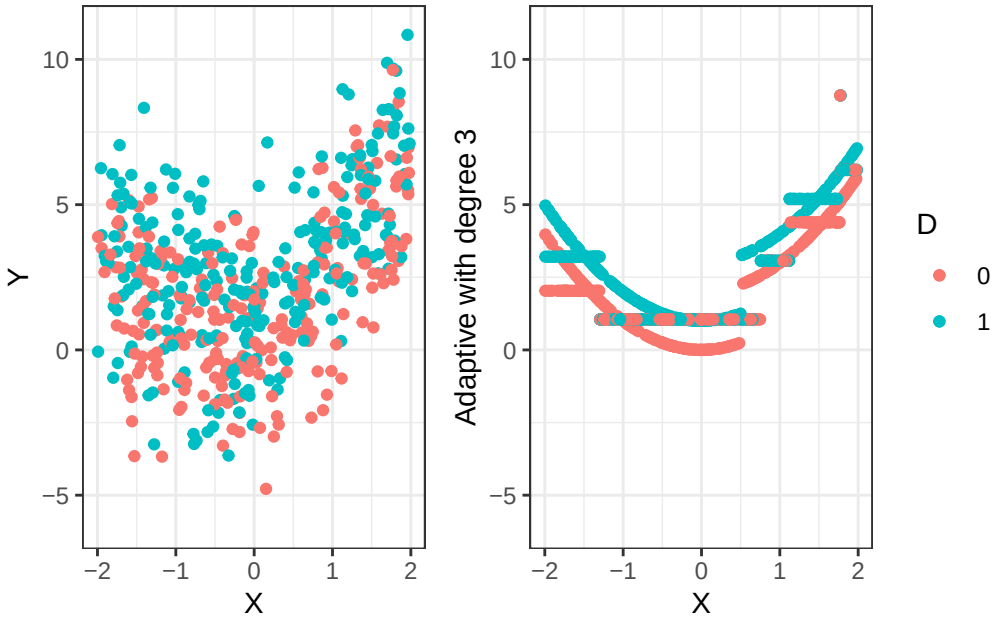
$$\beta = \arg\min_{\beta} \sum_i (Y_i - f(X_i))^2$$

- 伝統的アプローチでは、OLS | サブサンプル平均と一致

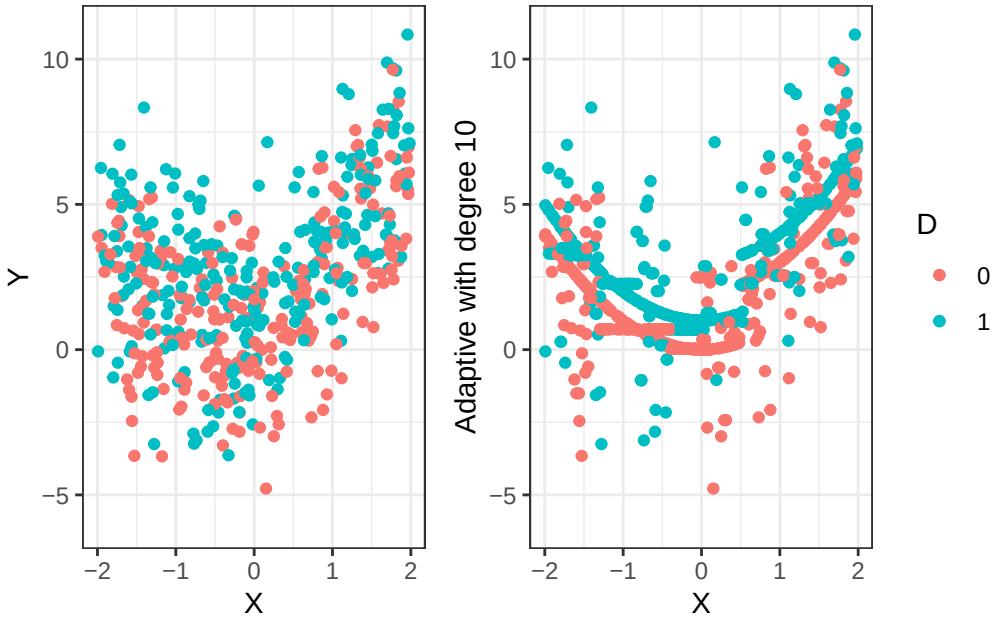
例



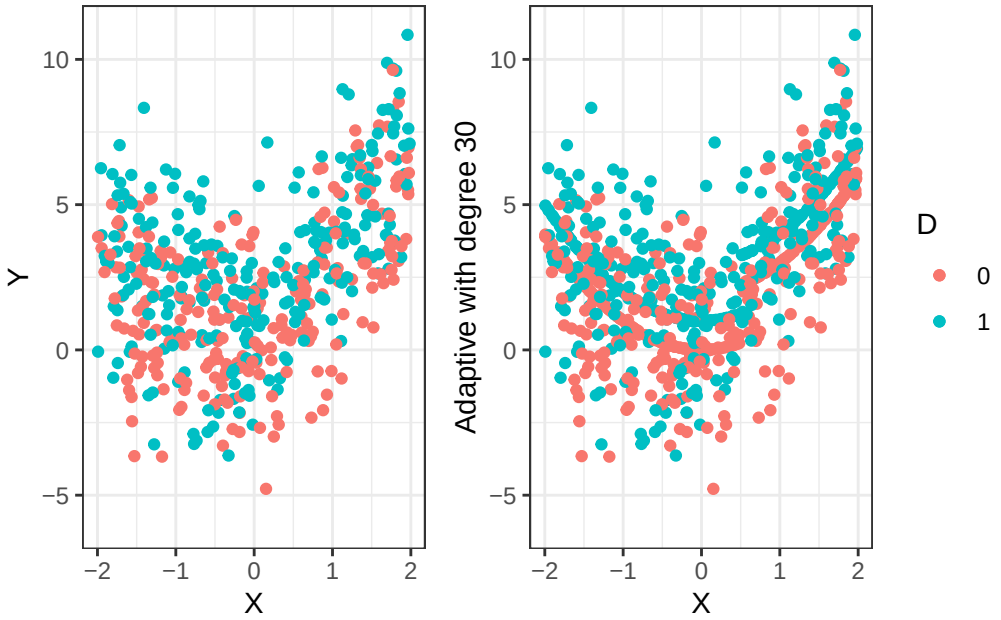
例



例



例



引用

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.