まとめ

機械学習

川田恵介

機械学習の分類

研究計画

- あらゆる研究においては、明確なゴール設定が必須
- 不適切なゴール：「計量経済学 | 統計学 | 機械学習を使って、日本社会を明らかにする」
  - あまりにも不明確 | 野心的
- まだ不適切なゴール：　「$Y =$ 賃金の決定構造を、$X =$ { 性別、年齢、居住地 } から明らかにする」
  - 依然として難しい
- 自身の使える分析道具と紐付けたゴール設定が現実的

機械学習の分類

- 様々な分析ツールを提供
- 大別すると
  - 教師付き学習: $\{Y, X\}$ が観察できるデータから、$Y \sim X$ を推定
  - 教師無し学習: Y が観察できないデータから、$Y \sim X$ を推定
  - 他にも, 強化学習など
- 例: Chat Bot GPT = 教師付き学習 + 強化学習

教師無し学習: 例

教師無し学習による都市圏の同定

- $X =$ 座標, $Y =$ 都市圏 (観察できない)

Figure1: Elephant

- アイディア: 光の塊 = 都市圏?

教師無し学習: 問題点

- 幅広い状況に” 自信を持って” 応用するのは難しい?
  - どのように評価するのか？
  - 定式化に依存していないか?

教師付き学習の復習

教師付き学習 = 母平均の推定

- 教師付き学習 := 母集団の特徴を柔軟に推定するツール
- 典型的には、母平均関数 $E[Y|X]$ を”近似する”関数 $g(X)$ を推定
  - 伝統的な推定方法は、$g(X)$ の”形状”をある程度決め内
  - 機械学習は、より多くをデータに決めされることができる

ポイント: データへの適応

- $Y_i = E[Y|X] +$ 個人差
  - データ上のパターン $\{Y_i, X_i\}$ は、$E[Y|X]$ について一定の情報をもつ
  - データに適合されることで、情報を引き出せる

ポイント: 過剰適合

- $Y_i = E[Y|X] +$ 個人差
  - データに適合しすぎると、　個人差が $g(X)$ に反映され過ぎてしまう
  - データに偶然含まれた”ハズレ値”に致命的な影響を受けてしまう

工夫

- 複雑なモデルを適度に単純化する
  - 剪定
- 複雑なモデルの集計値を予測値とする
  - Bagging/RandomForest

応用: 予測研究

- $g(X)$ は、優れた予測モデルになりうる
- $\iff E[Y|X]$ をうまく近似できていたとしても、予測性能は保証しない

- テストすべき

予測問題

- 新しい事例 i について、$Y_i \sim g(X_i)$ を達成する
- 予測誤差
- $$Y_i - g(X_i) = \underbrace{Y_i - E[Y_i|X_i]}_{\text{予測不可能}} + \underbrace{E[Y_i|X_i] - g(X_i)}_{\text{予測可能}}$$
- 予測可能な部分を 0 にできたとしても、予測不可能分が大きい可能性がある
  - 個人差が大きい社会データでは、特に深刻な恐れ

独立したデータによる評価

- 予測不可能分の大きさを” 理論的” に予測することは不可能
- データを分割して評価する必要がある
  - 訓練データ: モデルの推定
  - テストデータ: モデルの評価

応用: 母集団の推論

- 母平均全体ではなく、母集団の” 特定の特徴” であれば、より高い精度で推定できる
  - 区間として推定可能
  - $\iff$ 予測では困難

特徴

- 出発点は分布: 各 $\{Y, D, X\}$ 組み合わせについて、事例数 (割合)
- 知りたい分布の特徴: θ
  - 平均, 分散
  - 本講義では、$\tau = E[Y|D+1, X] - E[Y|D=0, X]$

部分線形モデル

- $$E[Y|D,X] = \underbrace{\tau}_{\text{知りたい特徴}} \times D + f(X)$$

1. $$E[Y|X] \sim g_Y(X), E[D|X] \sim g_D(X)$$

   を機械学習で推定

2. $$Y - g_Y(X) \sim D - g_D(X)$$

   を OLS で推定

- 信頼区間が計算可能
- 効果の異質性推定へ拡張可能

注意点

- 母集団の理解に向けた、機械学習の単純な応用には注意が必要

単純化の弊害

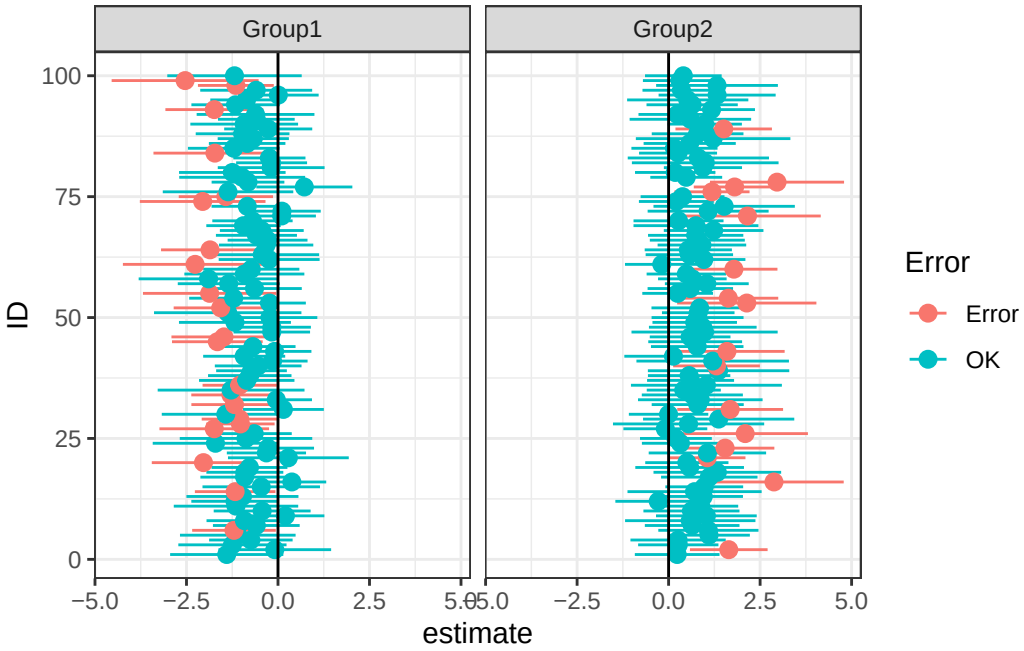
- 機械学習で $E[Y|D,X] \sim g(D,X)$ を直接推定し、$g(1,X) - g(0,X)$ を推定値とすればいいのでは?
- 例：世代間格差移転問題を機械学習により明らかにする
  - $Y =$ 子供世代の所得 , $X =$ 父親の学歴, 母親の学歴
- 父親の学歴のみを用いた決定木が推定された
  - 母親の学歴は関係ない？？？？

信頼区間計算の失敗

- 教師付き学習の**予測性能**は、データ分割で評価できる
- 母平均の特徴についての含意は限定的
  - (この範囲内のどこかにあります!!) とは言えない
- 信頼区間を計算すればいいのではないか
  - うまくいかない

数値例

- $E[Y|X] = 0$



なぜ

- 信頼区間の大前提は、

$$Y_i = E[Y|X] + \underbrace{\text{個人差}}_{\text{平均}=0}$$

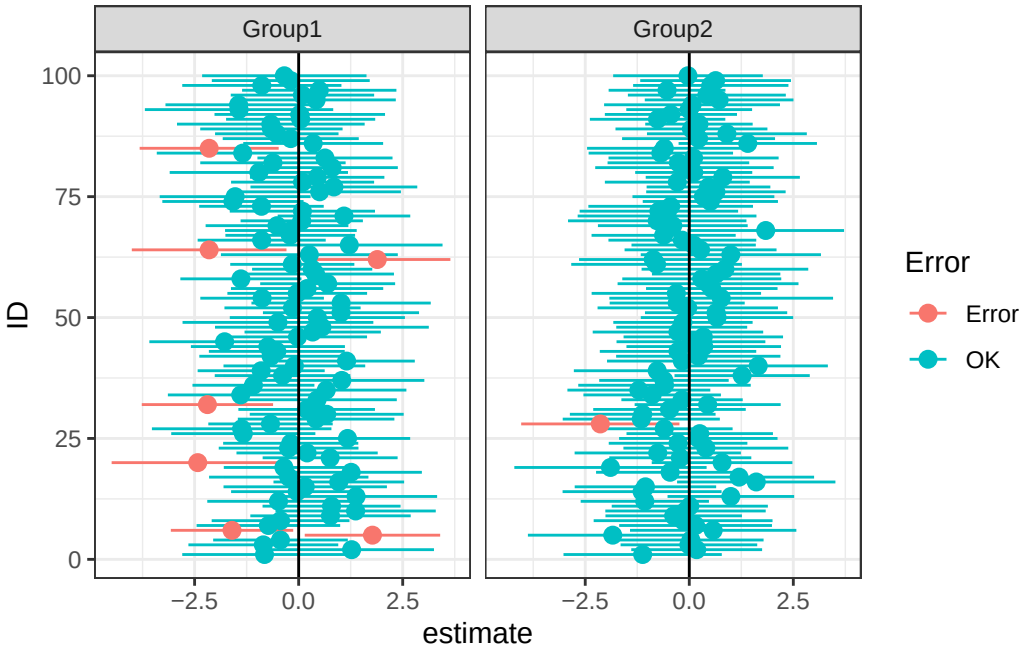
- 機械学習で生成されるモデルは一般に、個人差も反映している
  - 決定木: 高い値を予測されたグループにおいて、個人差も高め
- 誤った信頼区間が計算されがち

Honest Tree

- サンプル分割を用いて解決
  - 訓練データ: サブグループを形成
  - テストデータ: 信頼区間を計算

数値例

- $E[Y|X] = 0$



まとめ

- 機械学習をさまざまな局面で有益
- ただし、安直な応用は極めて危険