Resampling: Model Averaging and Evaluation

機械学習の経済学への応用

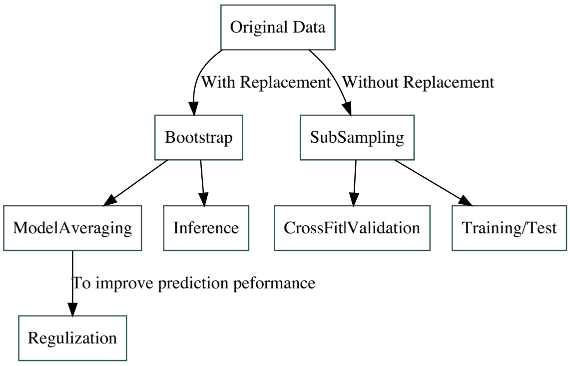
川田恵介

本スライドの内容

Resampling in ML

- データから再抽出を行う
  - Bootstrap, Subsampling, CrossValidaiton
- 現代的なデータ分析において、重要性が高まる
  - 機械学習において特に重宝されている印象

Concepts



Bagging | Random Forest

- ReSampling は、予測精度改善 (= 母平均への適合) に貢献できるか？
- どういう理屈で？

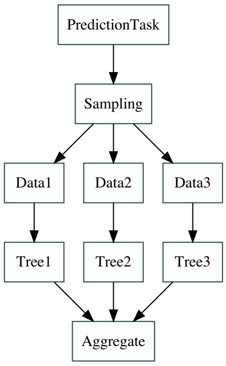
Regulization

- “複雑すぎるモデルを、適切に単純化する”
- 推定された” 複雑” すぎる予測モデルは、分散が大きすぎる
  - 分散の削減 (バイアスを導入)
- 予測木での実践: 深すぎる予測木を適切に単純化
  - **Bootstrap Model Aggregation (Bagging)**
  - Pruning (後日)

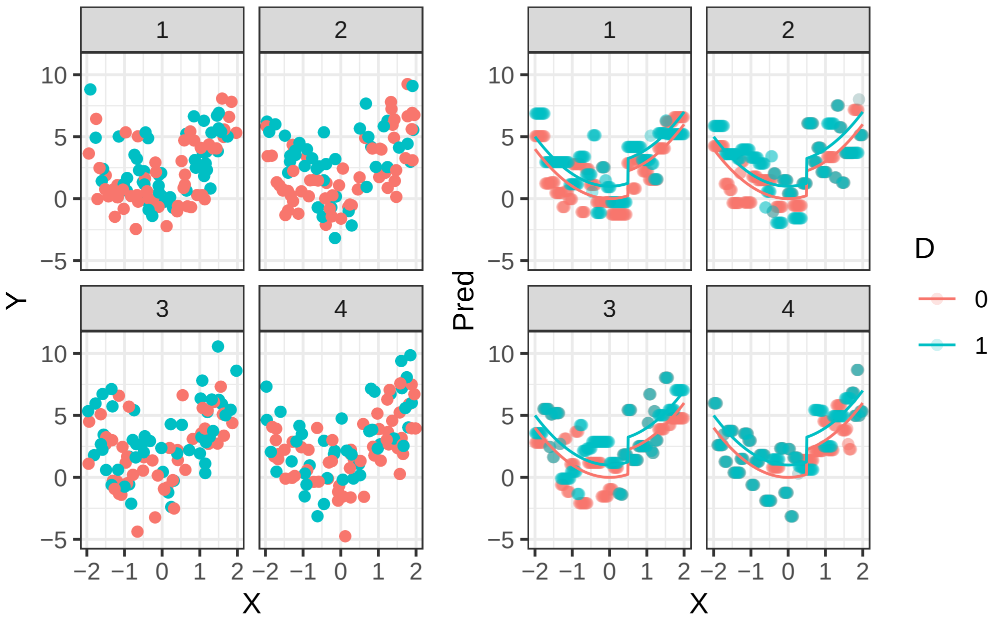
Bagging

- 予測精度を向上させるために、非常に”実用的”な手法
  - Pruning よりも有効なケースが多い
- アイディア: 大量の予測値の平均を取ることで、安定させる
  - 例: ranger 関数の Default 設定では、500 本の予測木を学習し、その平均値を最終予測モデルとする。

理想の Bagging



数値例

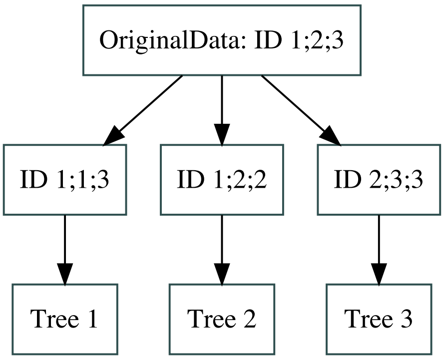


実現可能な Bagging

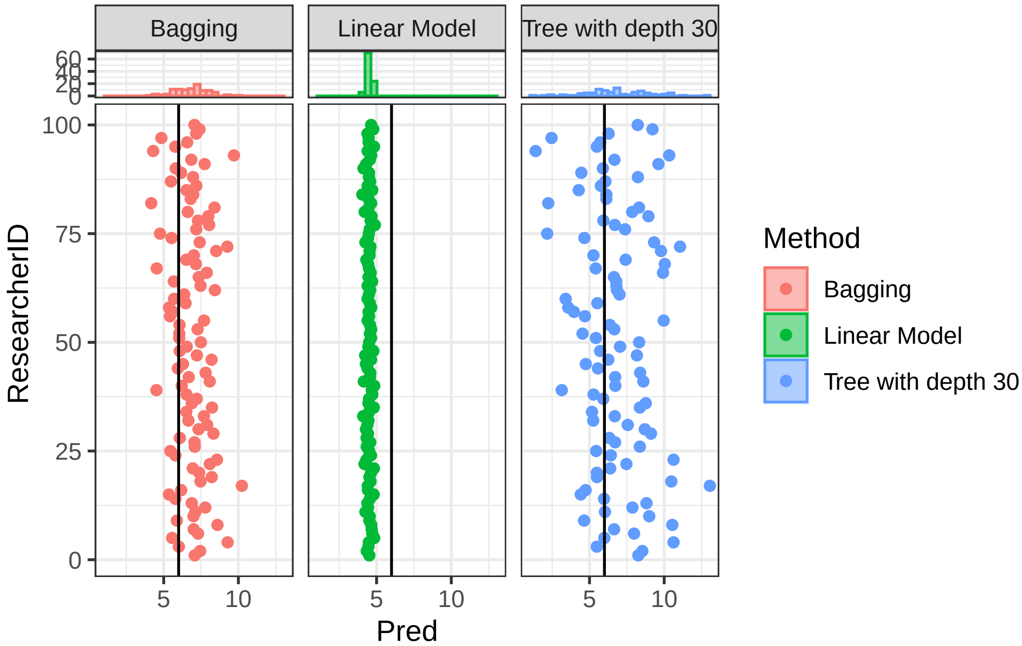
- 母集団から複数のデータを再抽出することは、非現実的
  - (Nonparametric) Bootstrap データで代替

1. Bootstrap で大量の" 複製" データを生成 (2000 個など)
2. 各複製データについて、予測木を生成
3. 各予測値を集計（平均値）

補論: Bootstrap



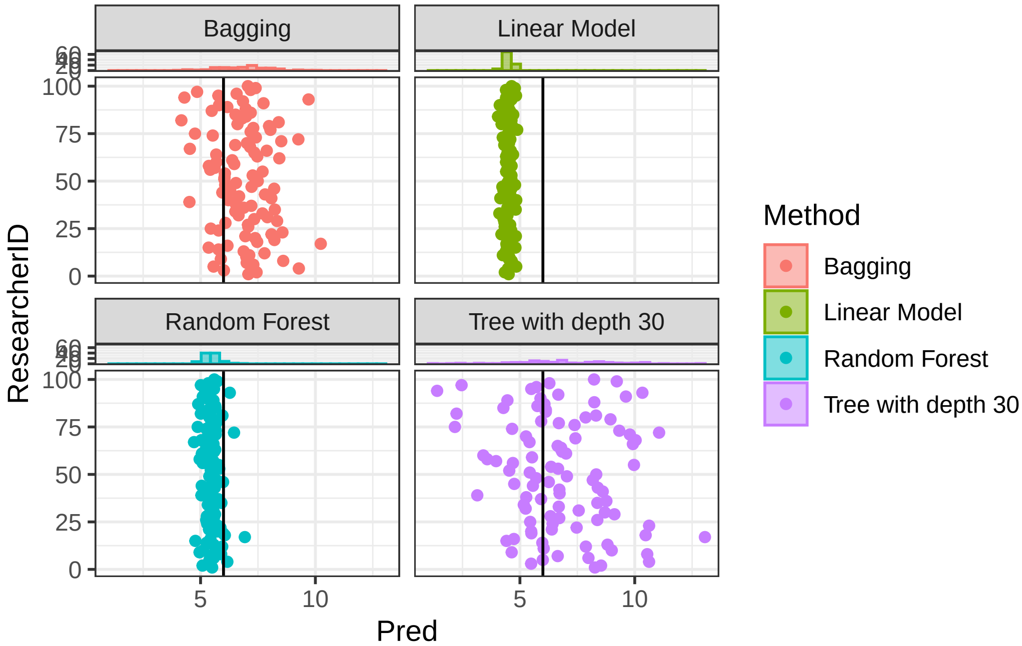
数値例



Random Forest

- データ分割に用いることができる変数群をランダムに選ぶ
- 例：ある予測木の第 n 分割を行う際に
  - Bagging: { 年齢、性別、学歴 } から選ぶ
  - Random Forest: { 年齢、性別 } から選ぶ
- Bagging をさらに改善可能

数値例



確率変数としての予測値

- Data が確率変数なので、そこから生成される予測値 x_b も確率変数
- 同じ分布から抽出される複数の予測値の平均として、新しい予測値を生成

$$x_{ave} = \frac{\sum_b x_b}{B}$$

分散削減

- B を無限に増やすと、分散は

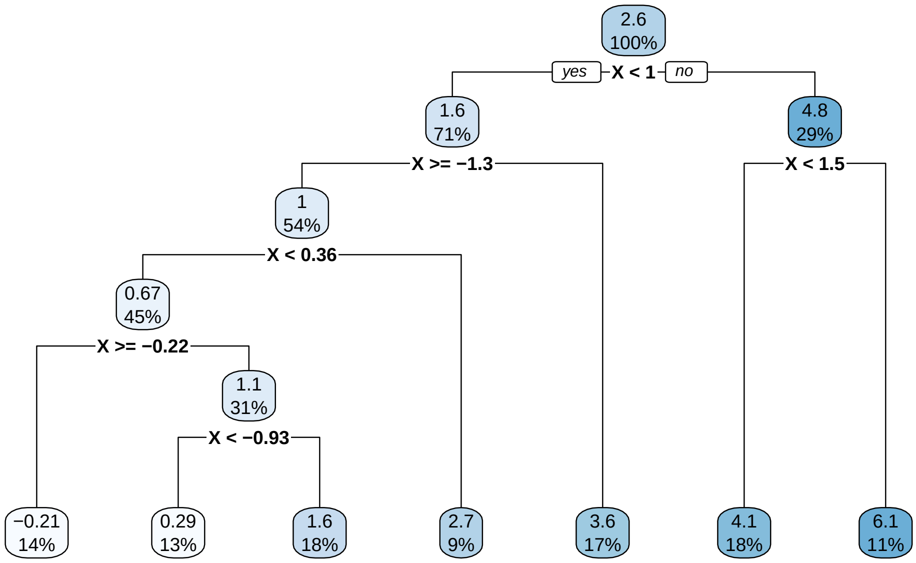
$$E[(x_{ave} - E(x_{ave}))^2] = \underbrace{\frac{var(x_b)}{B}}_{\rightarrow 0}$$

$$+ \underbrace{\frac{B-1}{B} \times corr(x_b, x_{b'}) \times var(x_b)}_{\rightarrow corr(x_b, x_{b'}) \times var(x_b)}$$

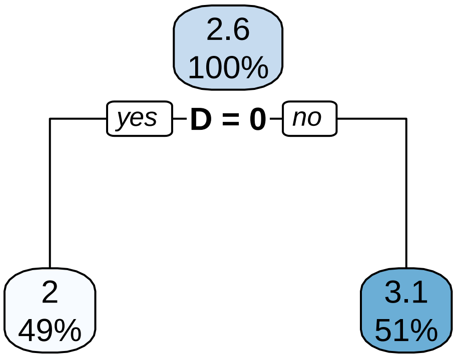
相関削減

- 理想の Bagging (完全に独立したデータから予測値を作る) では、分散は 0
- Bootstrap データから予測値を作ると、一般に $corr(x_b, x_{b'}) > 0$
  - ”上振れした” データから生成した Bootstrap データは、上振れしやすい
- 用いる変数もランダムに選ぶことで、予測値 x_b の相関を減らす
  - 予測力をもつが（より強力な変数のせいで）未活用な変数も用いることができる

例



例



まとめ

- モデル集計は非常に強力なアイディア
  - Bootstrap による平均化は、Tree モデルについて非常に有効
  - “OLS” には無意味
- “わざと一部の変数を利用不能にする” ことで、より性能向上が可能
  - 現実でも行われきた (縛りプレイなど）？
- デメリット: 計算時間、モデルが” 人間に理解できない” ほど複雑になる
  - 経済学研究への応用では、” 致命的” ではない？

Test データ

- SubSampling を用いて、モデルの評価を行う

Figure1: TwiceDip

２度付禁止!!!!

評価指標

- 理想の評価:

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

- 母集団上で定義されとおり、推定する必要がある

２度づけによる評価

- 予測モデルの推定に用いたデータで評価すると

$$\sum(Y_i - f(X))^2 = \sum(\mu_Y(X_i) + \underbrace{u_i - f(X_i)}_{Dependent})^2$$

- 丸暗記モデルが常に望ましい
  - X は完全一致するが、Y が異なる事例がなければ、0
- **一般的な統計ソフトが自動的に報告する MSE や R2 は、性能評価に用いるべきではない**

評価指標の推定

- ある予測モデルについて、評価指標を推定したい

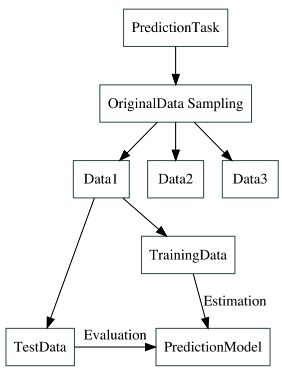
- Training & Test データアプローチが有力

1. 元データを Training | Test データに、ランダム分割 (0.8:0.2, 0.95:0.05)
2. Training データのみを用いて、予測モデル構築
3. Tsst データで評価

評価値

$$\sum_{i \in TestData}(Y_i - f(X))^2 = \sum_{i \in TestData}(\mu_Y(X_i) + \underbrace{u_i - f(X_i)}_{Independent})^2$$

RoadMap



解釈

- 推定された (非確率的な) 予測モデルの評価
- テストデータについての SamplingUncertainly は、考慮可能
  - テストデータ $\neq$ 母集団

- ただし通常の漸近理論に基づく信頼区間計算が可能

不確実性の評価

- ランダムサンプリングデータから、ランダム分割されたデータ = 母集団からランダムサンプリング

$$\underbrace{(Y_i - \underbrace{f(X_i)}_{Fix})^2}_{IID\ Error}$$

- 平均値について、漸近正規性が成り立つ

まとめ

- 推定と評価を同じデータで行うことは、一般に不適切
  - 学習に用いた過去問集について、完璧に答えられるようになったとて、、、
- Train | Test データへの分割は非常に一般的
  - モデル推定 | 評価に用いることができるサンプルサイズについて、トレードオフが発生
  - 目的に応じて交差検証を利用すべき (後日)