決定木アルゴリズム: モデル集計

経済学のための機械学習入門

川田恵介

モデル集計

- データ分析の基本的発想: 事例を集計することで、観察できない要因の偏りの影響を緩和
  - 予測モデルへの影響はどうしても残る
- 新しい発想: 予測モデル自体を” 集計” する

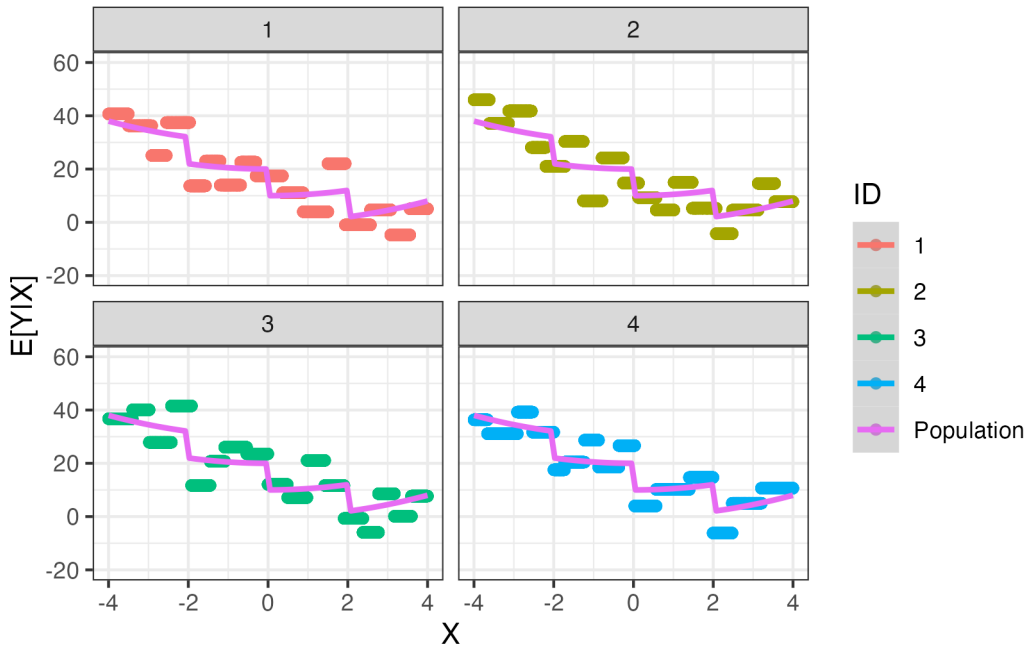
比喩: 予測屋会議

- 複数の” 専門家” の予測を集計して最終予測モデルとする
  - “エコノミスト” の見通しの平均値
  - 専門家委員会
- 一人の予測に頼るよりも、ましでは？
  - 教師付き学習にも応用可能な発想

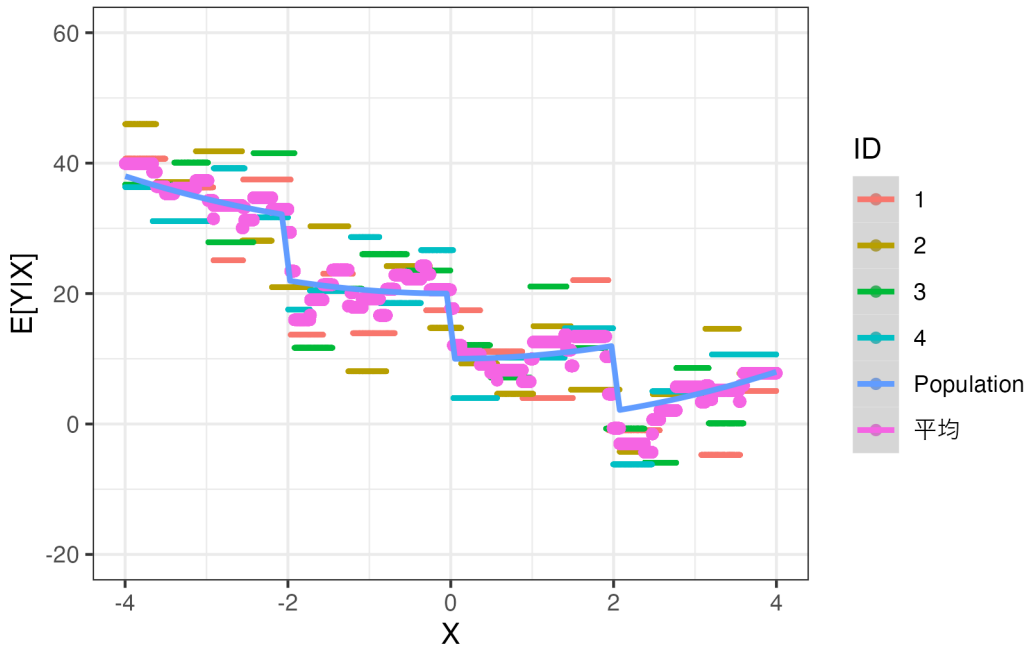
数値例

- 独立して収集したデータについて、深い予測木 (剪定なし) を推定
- 各予測値と、予測値の平均を比較

予測結果



集計



チャレンジ

- 「独立して抽出された」有限個データから生成された予測モデル
- 「独立して抽出した複数のデータから得た」予測モデルの集計は通常不可能
  - 推定に使ったサンプルサイズが実質的に増えているので、性能改善は” 当たり前”
- 近似的に行う
  - (Nonparametric) bootstrap の活用

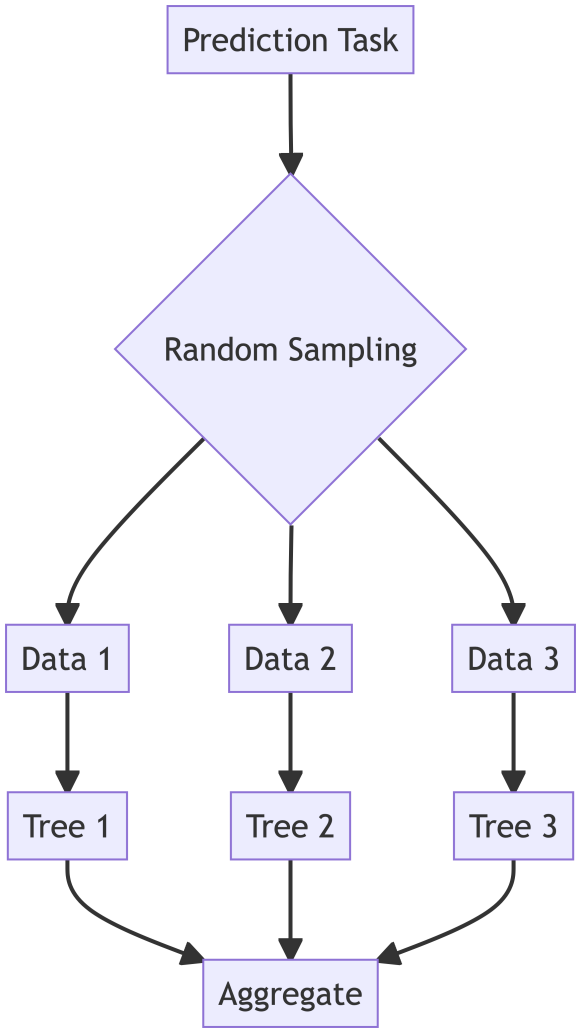
Bootstrap Aggregating

- Bagging

決定木の不安定性

- 現実は複雑なので、複雑なモデル (巨大な木) が本来は望ましい
  - 事例数が限られている場合、データ固有の特徴を強く反映してしまう
  - 適度に単純化する必要があるが、、、、
- 多くの実践で、決定木推定の不安定性 (= データ固有の特徴を強く反映してしまう) は、剪定を行っても十分に緩和できない
  - Bootstrap でデータを複製して、モデル集計

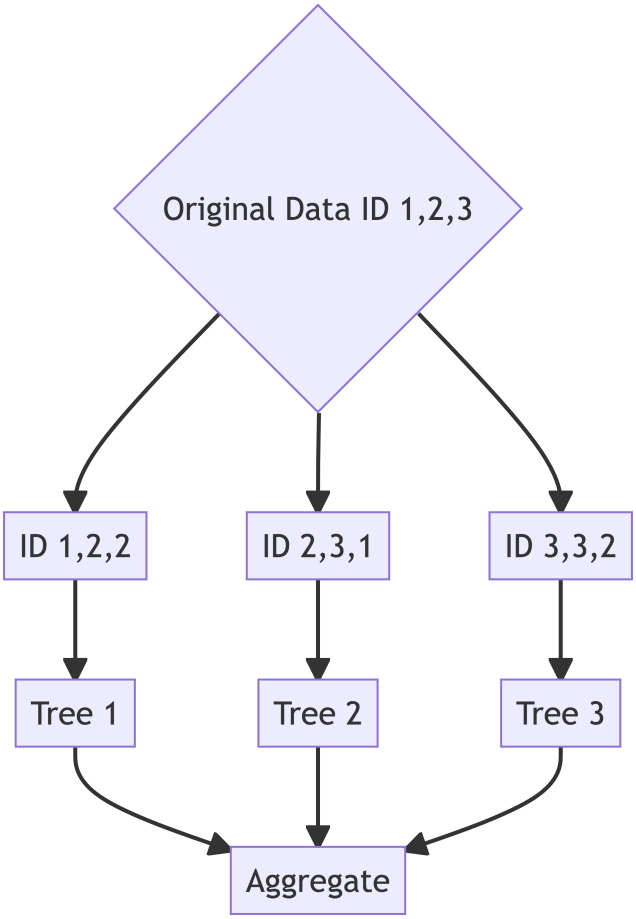
理想の Bagging



アルゴリズム

1. Nonparametric bootstrap で、データの複製を行う (500,1000,2000 など)
2. 各複製データについて、“深い” 決定木を推定
3. 各 X についての予測値の平均を最終予測値とする

補論: Bootstrap



Bagging の発想

- 基本アイディア: 非常に深い木を生成すれば、予測結果が不安定になるが、
- 平均を取れば、安定する
  - 独立・無相関であれば、無限個の**複製データ**から予測モデルを作れば、分散を 0 にできる
  - 今の PC であれば、大量の予測モデルの生成は可能

Bagging の限界

- リーマンショック、地震保険が (火災保険などと比べて) 難しい理由は？
  - 事象間での相関

- よく似た予測結果ばかりであれば、平均をとってもあまり意味がない
  - 予測結果を十分に” 散らばらせる” 必要がある
  - 金融ポートフォリオであれば、平均的なリターンが低く買ったとしても、” 海外” の商品も組み込むなど

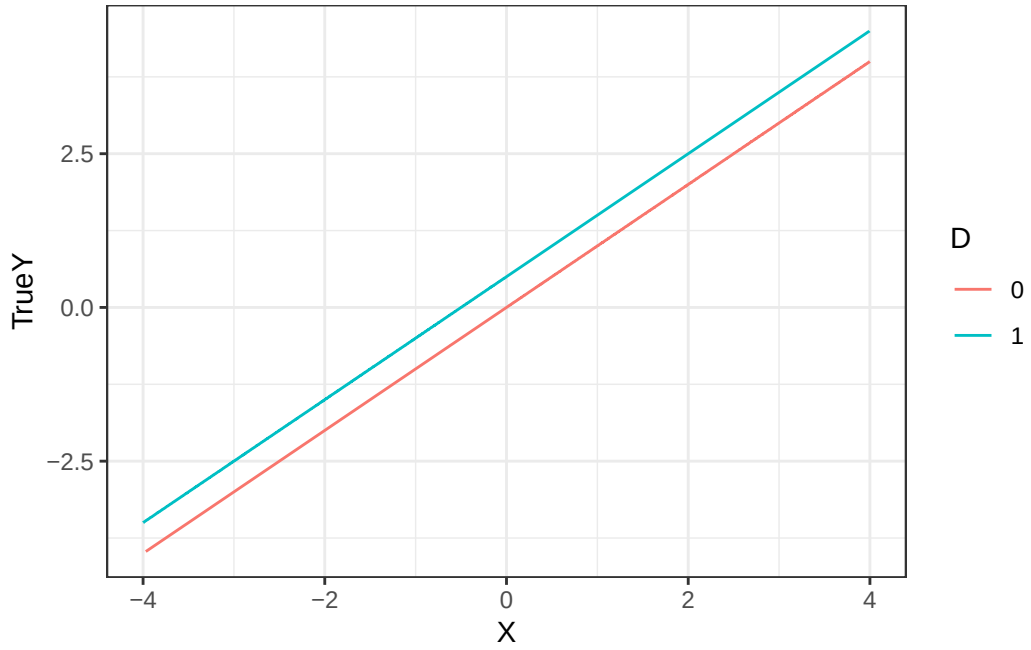
RandomForest

- 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
    - 第 $n+1$ 分割を行う際には、{ 学歴、性別 }

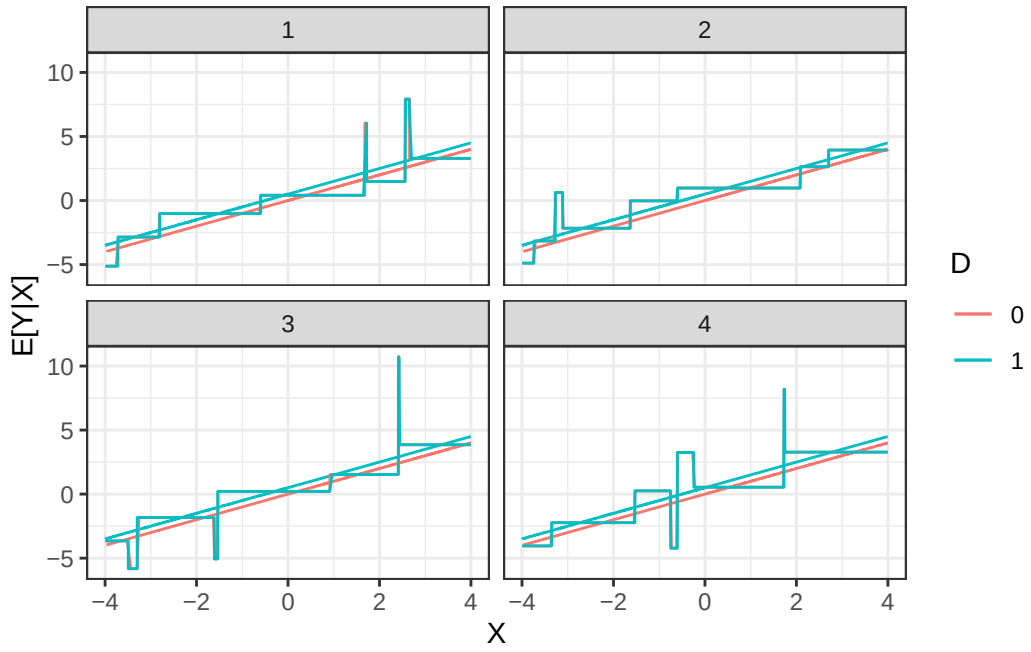
RandomForest: **動機**

- 動機: 予測値同士の相関を弱める
  - 相関を強める要因 (データが多少変わっても、同じような変数を活用する) を排除
  - そこそこの予測力を持つ変数が、強力な予測力を持つ変数の陰に隠れてしまうことを避けられる

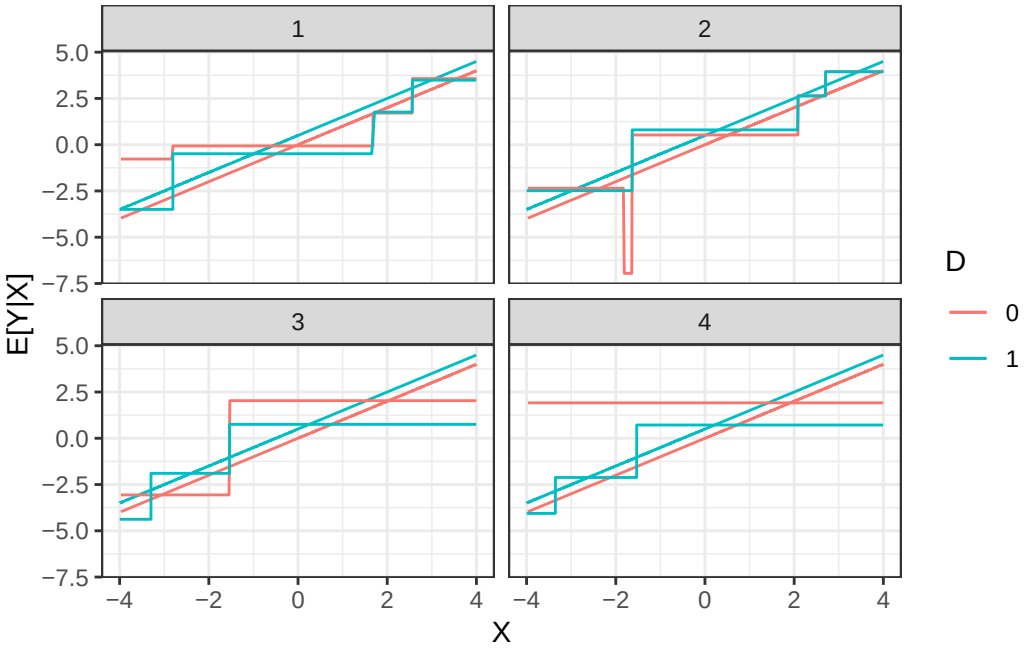
数値例



数値例



数値例



実例

```
X <- select(Data, -Price)

Y <- Data$Price

Fit <- SuperLearner(
  X = X,
  Y = Y,
  SL.library = c(
    "SL.mean",
    "SL.rpart",
    "SL.rpartPrune",
    "SL.ranger"
  )
  )

Fit$cvRisk
```

```
SL.mean_All      SL.rpart_All SL.rpartPrune_All     SL.ranger_All
   696.8034          307.2621          269.4826          218.9182
```

まとめ

- Resampling = 元々のデータから、“新しい” データを作り出す
- Resampling は現代のデータ分析において、強力な手法
  - 交差検証 (事例の被りは許さない): モデル評価など
  - Bootstrap: モデルの改善 (Bagging/RandomForest)
  - 伝統的な Inference への Bootstrap の応用も、もちろん重要