決定木アルゴリズム

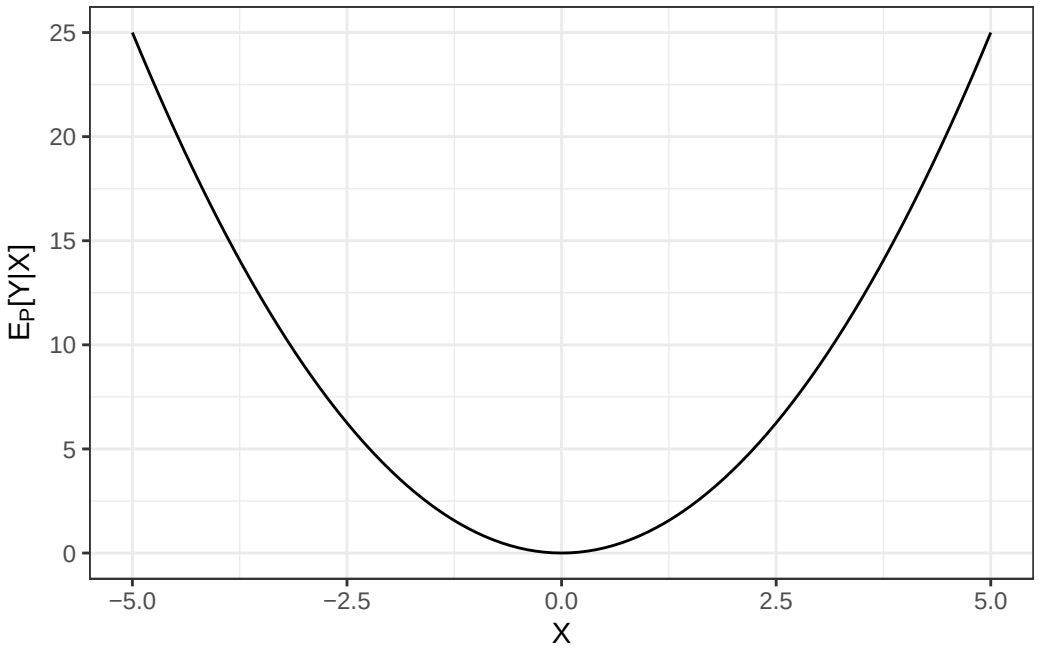
最適化

川田恵介

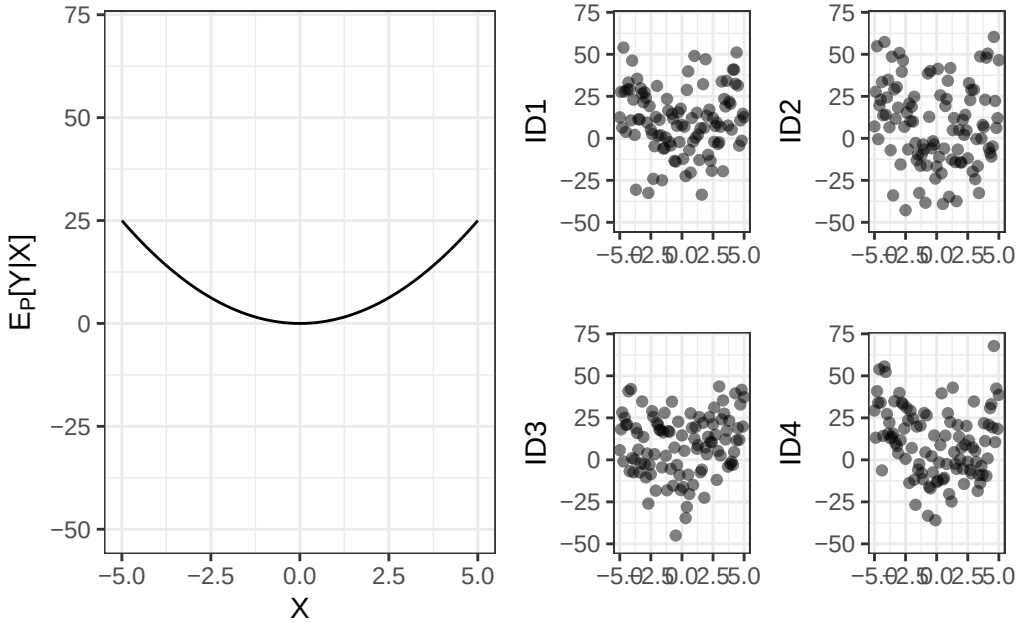
ここまでのまとめ

- 母平均関数 $E_P[Y|X]$ が理想的な予測モデル
- 理想的な予測モデルに近づけるには、“適度に” 複雑なモデルが必要

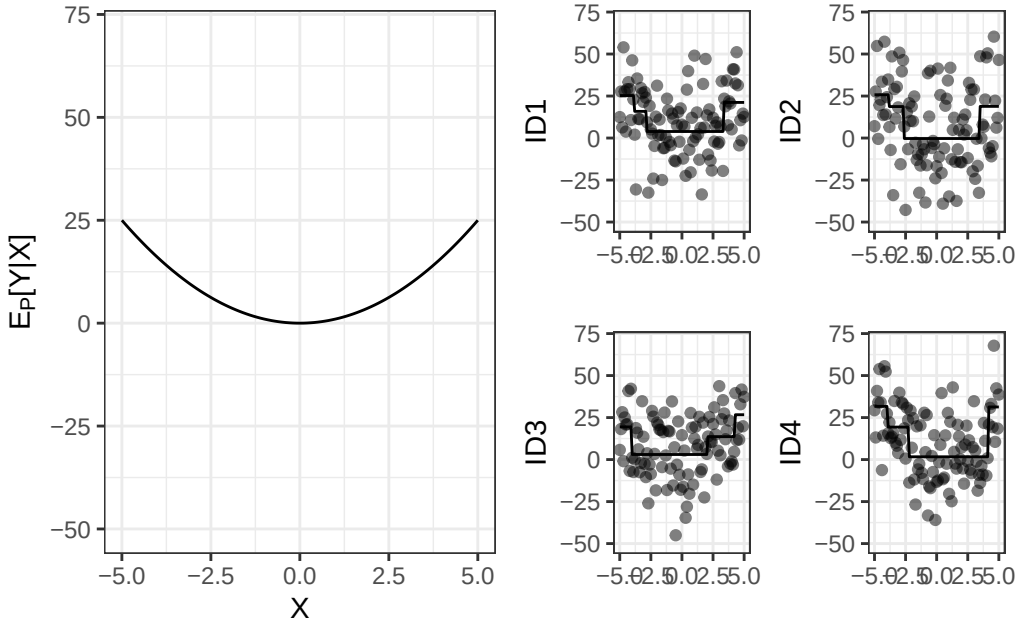
数値例: 母平均関数



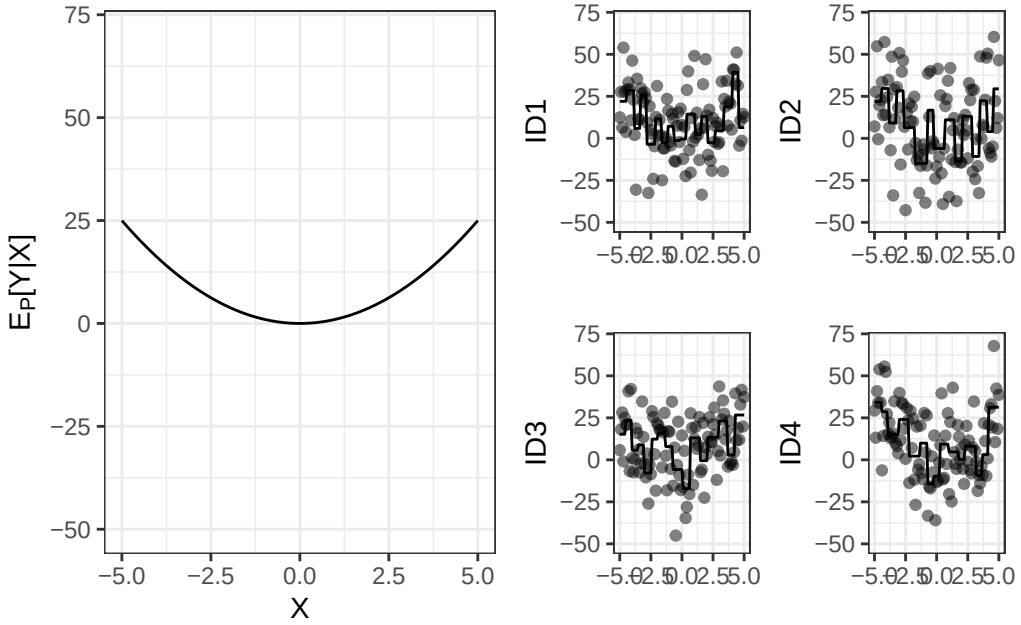
数値例: データ



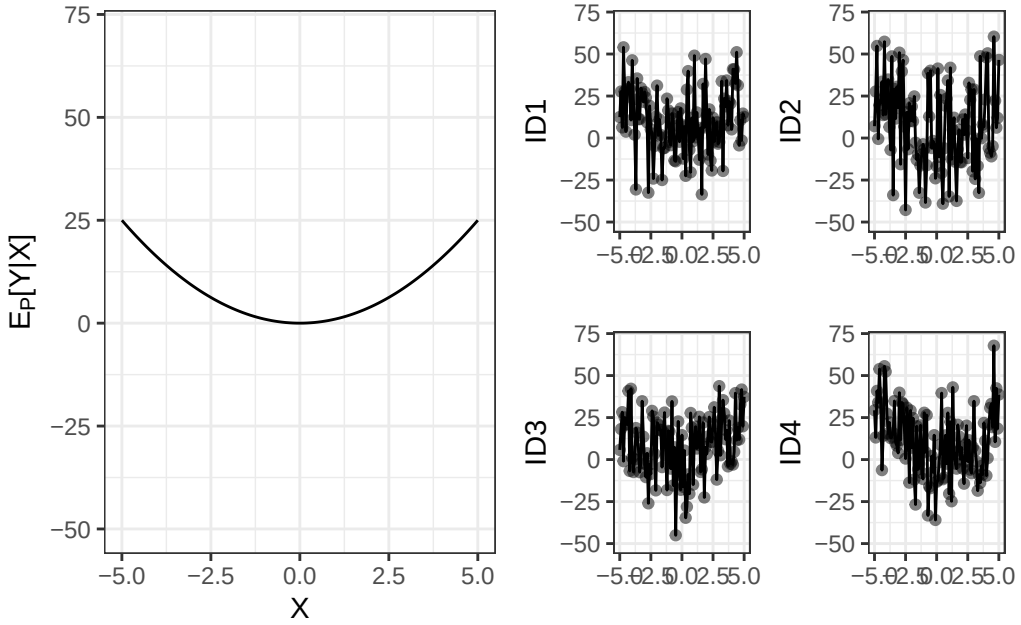
数値例: 浅い決定木



数値例: 深い決定木



数値例: 丸暗記



剪定

- 剪定: 一旦非常に深い木を推定した後に、単純化を行う

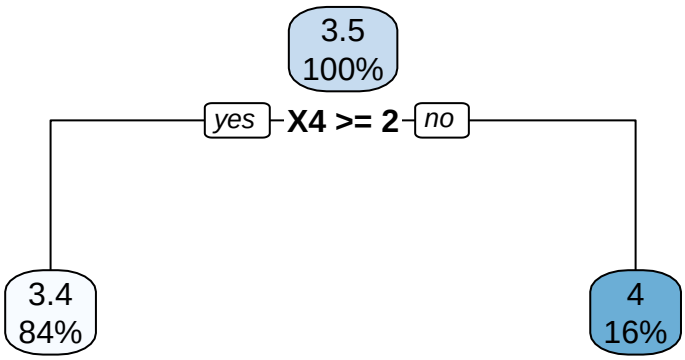
Step 1. 深い木の推定

- 停止条件を緩めると、一般にどこまでもサブサンプル分割が行われる
  - 平均値が異なるサブグループが見つかる限り止まらない

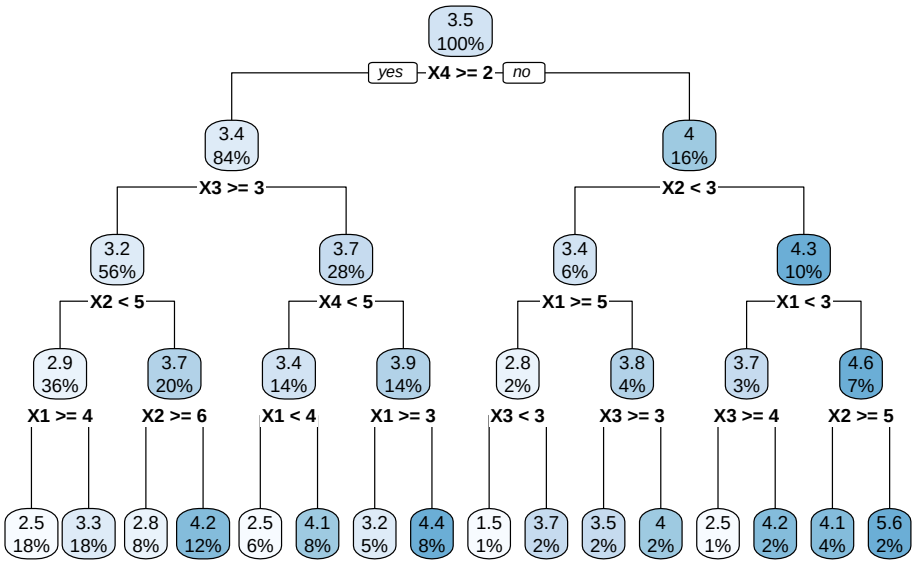
数値例: サイコロゲーム

- ディーラーは、サイコロを 5 つふり、4 つ $(X_1, .., X_4)$ プレイヤーに見せる
  - プレイヤーは残り一つの出目 Y を予測
- サイコロの出目は、uniform 分布 (完全無相関) に決定
  - 理想の予測モデル $g(X_1, .., X_4)$
- “見” を 200 回行いデータ収集

例



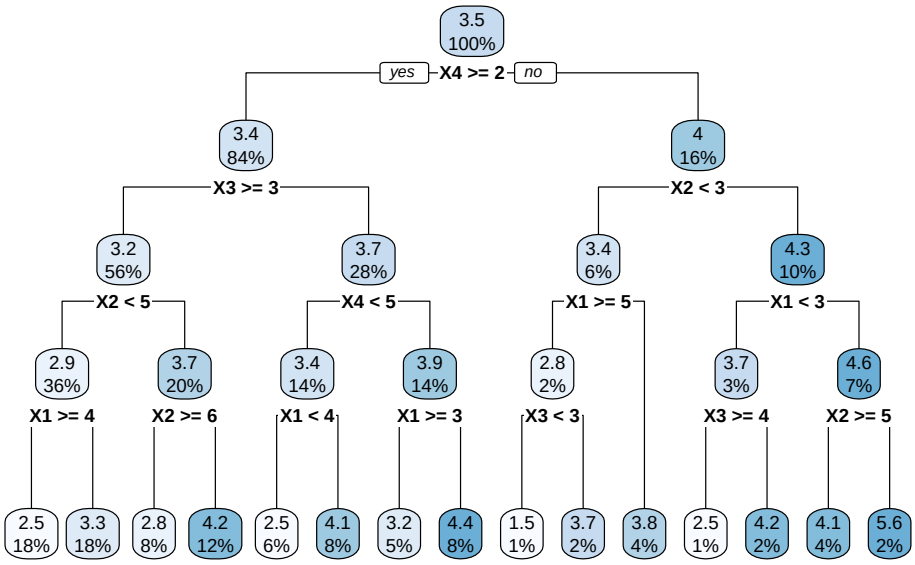
例



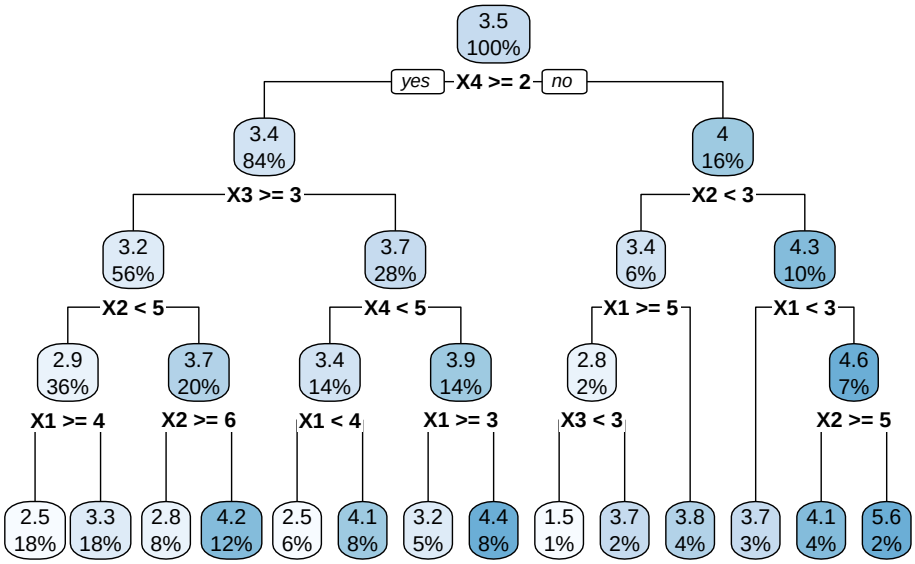
Setp 2. 剪定

- 分割しても、データへの適合が悪化しずらいサブグループから再結合していく
  - 小規模なサブグループを分割している
  - 分割しても予測値があまり変化しない

例: 剪定


3.5
100%
yes
X4 >= 2
no
3.4
84%
4
16%
X3 >= 3
X2 < 3
3.2
56%
3.7
28%
3.4
6%
4.3
10%
X2 < 5
X4 < 5
X1 >= 5
X1 < 3
2.9
36%
3.7
20%
3.4
14%
3.9
14%
2.8
2%
3.7
3%
4.6
7%
X1 >= 4
X2 >= 6
X1 < 4
X1 >= 3
X3 < 3
X3 >= 4
X2 >= 5
2.5
18%
3.3
18%
2.8
8%
4.2
12%
2.5
6%
4.1
8%
3.2
5%
4.4
8%
1.5
1%
3.7
2%
3.8
4%
2.5
1%
4.2
2%
4.1
4%
5.6
2%

例: 剪定


3.5
100%
yes
X4 >= 2
no
3.4
84%
4
16%
X3 >= 3
X2 < 3
3.2
56%
3.7
28%
3.4
6%
4.3
10%
X2 < 5
X4 < 5
X1 >= 5
X1 < 3
2.9
36%
3.7
20%
3.4
14%
3.9
14%
2.8
2%
4.6
7%
X1 >= 4
X2 >= 6
X1 < 4
X1 >= 3
X3 < 3
X2 >= 5
2.5
18%
3.3
18%
2.8
8%
4.2
12%
2.5
6%
4.1
8%
3.2
5%
4.4
8%
1.5
1%
3.7
2%
3.8
4%
3.7
3%
4.1
4%
5.6
2%

最適化問題の活用

- 残された問題は、どこまで剪定するか?
  - 剪定の水準をどのようにコントロールするか?
- 最適化問題に落とし込む
  - 何をやっているか (慣れれば) わかりやすく、自由度が高く、PC にも優しい枠組み

最適化問題

- “ある指標を最大化/最小化するように、決定する” 枠組み
  - 経済学: 効用最大化問題の結果として財の購入、利潤最大化問題として生産計画、社会厚生最大化として政策

練習問題

- 以下を **最小化** するようにサブグループを再結合できない

$$\text{データにおける二乗誤差}$$

- 剪定が行われず、複雑になりすぎる
  - 経済学: 公害物質が排出されすぎる、騒音がですぎる

罰則付き最適化

- 以下を **最小化** するようにサブグループを再結合

$$\text{データにおける二乗誤差} + \underbrace{\lambda \times \left|\text{サブグループの数}\right|}_{\text{罰則項}}$$

- λ : Hyper Parameter (rpart 関数では cp)
  - 罰則項 = 複雑性への” 課税”

まとめ

- 「データへの当てはまり改善」は活用
  - そのままでは複雑になりすぎるので、複雑さへの課税でコントロール

- 経済学でもお馴染みのアイディア
  - “市場” を一切活用していない” 都市” は” 存在しない”
  - 完全に” 市場” 任せにすると問題が生じるので、政策介入 (課税/補助金など) をする

モデルの試作と評価

- 最適な課税水準をどのように決めるのか
- 社会政策とは異なり
  - 目標が明確 (予測性能の改善)
  - 実験の費用が安い
- 「特定の課税水準のもとでモデルを試作し、中間評価する」を繰り返すことで最適な水準を探り出す

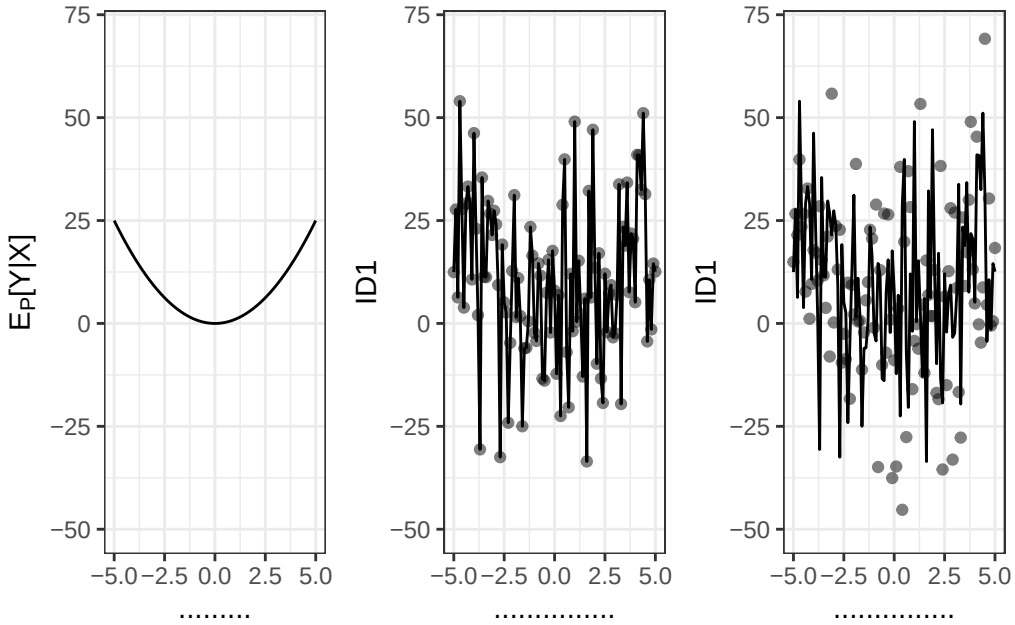
評価

- 現代 PC を使えば、モデルの試作は簡単
- 難しいのは適切な評価
  - モデルの試作に用いたデータは使用できない
  - 理論的指標は色々提案されているが (AIC,BIC など)、使える状況は” 限られている”

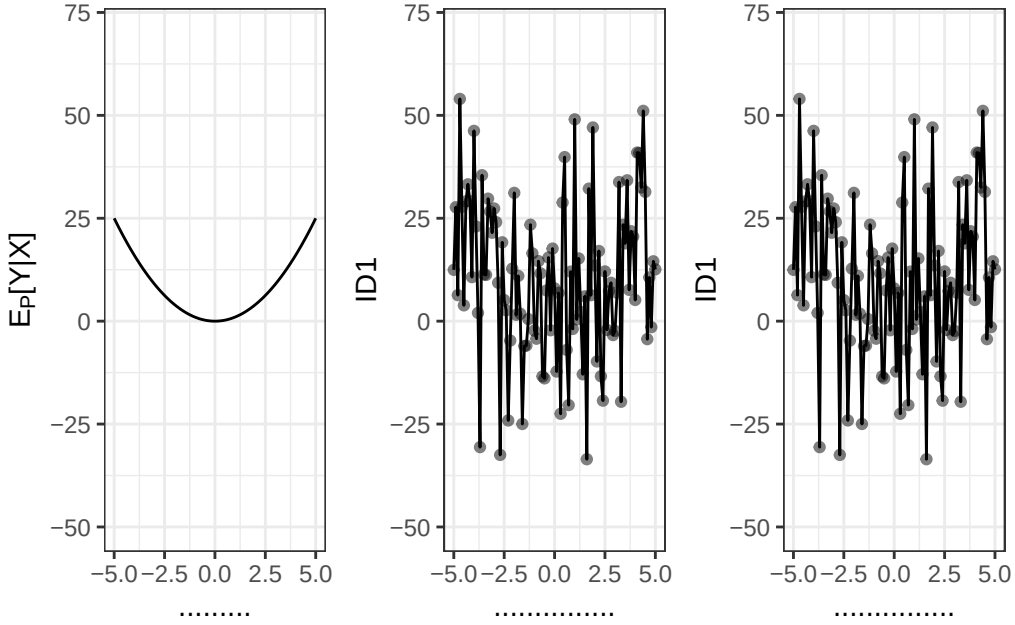
データ分割法

- 本来やりたいことは、“新しい” 事例を予測しやすいモデルを選ぶこと
- 元々のデータをランダムに 2 分割することで、擬似的に新しいデータを作り出す
  - モデル試作用データ := 訓練データ
  - 検証用データ := 検証データ

数値例: 深い決定木



数値例: 間違った評価法



データ分割法の手順

1. サンプルをランダムに 2 分割する
2. 検証対象とする λ を設定
    - 訓練データを用いて決定木を推定
    - 検証データでテスト
3. 2 を異なる λ について繰り返し、最も検証データへの当てはまりが良くなる λ を探す
4. 最善の λ と全データを用いて、決定木を推定

まとめ

- 同じデータで、モデル試作と評価はできない!!!
- 資格試験勉強の比喩
    - 過去問を繰り返しとき、答え合わせをすることで、試験対策を学習
    - 学習した方法の有効性を同じ過去問でテスト…?
    - 可能であればもしでテスト、不可能ならば過去問の一部は答えを見ずに残しておく

交差推定

- 訓練/評価 データが、ランダムに分割されていれば OK
    - “役割の固定” は本質ではない

交差推定

1. データをいくつか (2,5,10,20 など) に分割
2. 第 1 サブデータ **以外** を用いて予測モデルを試作
3. 第 1 サブデータに予測値を適用
4. 全てのサブデータに 2,3 を繰り返す

交差検証

- Cross validation

5. 交差推定で導出した予測値と実現値について、予測誤差を推定

数値例: 単純平均 VS 決定木 (深さ 2)

```
# A tibble: 6 x 3
  Group     Y     X
  <dbl> <dbl> <dbl>
1     1     6     3
2     1     7     1
3     2     4     3
4     2     5     2
5     3     4     1
6     3     4     1
```

数値例: 単純平均 VS 決定木 (深さ 2)

```
# A tibble: 6 x 5
  Group     Y     X PredMean PredTree
  <dbl> <dbl> <dbl>    <dbl>    <dbl>
1     1     6     3     4.25        4
2     1     7     1     4.25        4
3     2     4     3    NA          NA
4     2     5     2    NA          NA
5     3     4     1    NA          NA
6     3     4     1    NA          NA
```

数値例: 単純平均

```
# A tibble: 6 x 5
  Group     Y     X PredMean PredTree
  <dbl> <dbl> <dbl>    <dbl>    <dbl>
1     1     6     3     4.25        4
2     1     7     1     4.25        4
3     2     4     3     5.25        6
4     2     5     2     5.25        6
5     3     4     1    NA          NA
6     3     4     1    NA          NA
```

数値例: 単純平均

```
# A tibble: 6 x 5
  Group     Y     X PredMean PredTree
  <dbl> <dbl> <dbl>    <dbl>    <dbl>
1     1     6     3     4.25        4
2     1     7     1     4.25        4
3     2     4     3     5.25        6
4     2     5     2     5.25        6
5     3     4     1     5.5         7
6     3     4     1     5.5         7
```

数値例: 単純平均

```
# A tibble: 6 x 7
  Group     Y     X PredMean PredTree ErrorMean ErrorTree
  <dbl> <dbl> <dbl>    <dbl>    <dbl>     <dbl>     <dbl>
1     1     6     3     4.25        4    3.06           4
2     1     7     1     4.25        4    7.56           9
3     2     4     3     5.25        6    1.56           4
4     2     5     2     5.25        6    0.0625         1
5     3     4     1     5.5         7    2.25           9
6     3     4     1     5.5         7    2.25           9
```

- 平均二乗誤差 (Mean) 2.79
- 平均二乗誤差 (Tree) 6

トレードオフの緩和

- サンプル分割法では、訓練に多くの事例を割くと、評価に割ける事例が減り、評価の精度が下がる
- 交差検証では、すべての事例について予測値を計算し、その平均を取るので、評価の精度を確保できる

まとめ

- 複雑なモデルの推定は、現代的な PC + アルゴリズムであれば容易
- モデルを適切に単純化することに工夫が必要
- ２度漬け禁止の大原則

– モデルの推定に使ったデータは、評価に原則使わない