決定木アルゴリズム

事例の集計

川田恵介

丸暗記の問題

- データへの適合は、データ分析の大部分で活用されている戦略だが、注意も必要
- 非常に複雑なモデルを作ると、データへの適合度は非常に高くなる
- $\iff$ **予測性能** が改善するとは限らない
  - データ固有の特徴を強く反映してしまう
  - = 事例集計が十分にできない

事例集計

- “すべての” 推定方法で、事例の集計を行う
- 集計を行わない方法も、“日常” 的に使われている
  - 例: 丸暗記法

丸暗記法

- 過去の全く同じ事例を、予測値とする
  - 全く同じ事例がなければ、最も近い事例を予測値とする
  - 巨大な決定木を推定しても、通常 OK
- 日々活用されている
  - 判例, 歴史, スポーツ選手の将来
- わかりやすいが、、、、多くの応用例で劣悪な予測性能
  - データから観察できない個人差がある場合、深刻なトラブル

例: 賃金予測

```
# A tibble: 4 x 2
   賃金　年齢
  <dbl> <dbl>
1   100    25
2    20    28
3    40    20
4    15    40
```

- 21 歳の賃金予測は?

例: 賃金予測

```
# A tibble: 4 x 6
   賃金　年齢 丸暗記法 平均法 `データへの適合度: 丸暗記` `データへの適合度: 平均法~1`
  <dbl> <dbl>    <dbl>  <dbl>                    <dbl>                  <dbl>
1   100    25      100   43.8                        0                   3164
2    20    28       20   43.8                        0                    564
3    40    20       40   43.8                        0                     14
4    15    40       15   43.8                        0                    827
# i abbreviated name: 1: `データへの適合度: 平均法記`
```

丸暗記法 VS 平均

- 丸暗記法の方が一般に
  - 複雑な予測モデル
  - データへの適合度が高い
- 現実はおそらく複雑 X の値が異なれば、Y の値も異なる
  - 丸暗記の方がいいのでは?

論点先取り

- 丸暗記のようなデータに適合したモデルは、観察できる要因 X の情報を有効活用できる**可能性**を高める
- 事例集計が不十分に終わり、**観察できない要因の偏り**が弊害をもたらしうる

  - 収集した事例によっては、質の極めて悪い予測モデルが生成される
- 観察できない要素について議論する必要があり、**抽象的な**枠組みが必要
  - 母集団とサンプリング

理想の予測モデル

- 完璧な予測は不可能
- 理想の予測モデルは母平均値関数

母集団とサンプリング

- 分析チームは、事例収集から始める
- 直接観察できない巨大な集団 (**母集団**) から、事例 (**データ**) が収集 (**サンプリング**) される
  - 大学前ローソンの潜在的な顧客 (母集団) から、ある日の利用者 (サンプル) の購入履歴
- **同じ母集団**から**新たに**ランダム抽出する事例を予測するモデル $g(X)$ を構築
- ポイント: 母集団を調べたいが、“何があっても” 不可能!!!
  - 不完全な推測 (推定) しかできない

予測モデル問題

- 新たな事例について、二乗誤差 $(Y - g(X))^2$ を最小化する
  - 特定の事例についてのみ、予測が上手くいく可能性
- 母集団について、平均的に上手くいくかどうかで、性能を評価
- 母集団における平均二乗誤差

$$:= E_P[(Y - g(X))^2]$$

予測モデルの評価

- 評価用の新しい事例は、通常存在しない
- よく似た状況を作り出せる
  - データをランダムに 2 分割する

– 一方のデータでモデルを作り、もう一方で平均二乗誤差を計算する

予測モデルのポイント

- ある X が分かれば、Y の予測値を自動的に計算してくれる
- X が同じであれば、予測値は必ず同じ
  - 母集団において X 以外の要素も Y に影響を与えるのであれば、完璧な予測は不可能

理想の予測モデル

- 最もマシな予測は?
- 二乗誤差であれば、“大外れ” を減らす必要がある
  - 最善の予測値は、$g(x) = E_P[Y|X]$
  - := ある X 集団における平均値
- X の持つ情報を完全に活用できているモデル

予測モデルのポイント

$$Y - g(X) = \underbrace{Y - E_P[Y|X]}_{\text{削減不可能}} + \underbrace{E_P[Y|X] - g(X)}_{\text{削減可能}}$$

- 個人差はどうしようもない
- どのように母平均を推定するか?
  - 事例集計を活用

推定問題

例: 顧客の就学状態予測

- $Y =$ 大学生/その他, $X =$ 休日/平日
- 母集団: 潜在顧客
  - 曜日だけでコンビニにいくかどうかは決まっていない
  - 休日であれば 25%, 金曜日以外の平日であれば 60% が大学生
  - 週全体では、50% が大学生

  - 実際には未知

- データ: 一部をランダムに抽出
  - 偶然偏る

例: 4 事例を抽出

- A さんがランダムに 4 名を抽出

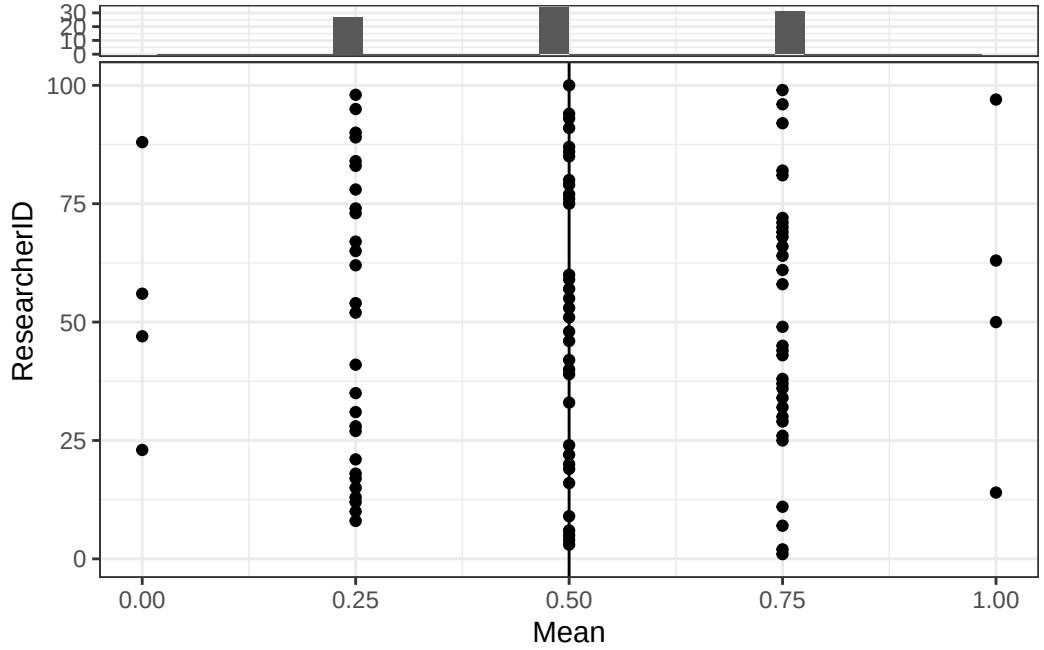
```
# A tibble: 4 x 2
  Holiday University
  <chr>   <chr>
1 Not     No
2 Not     Yes
3 Not     Yes
4 Yes     Yes
```

- B さんがランダムに 4 名を抽出

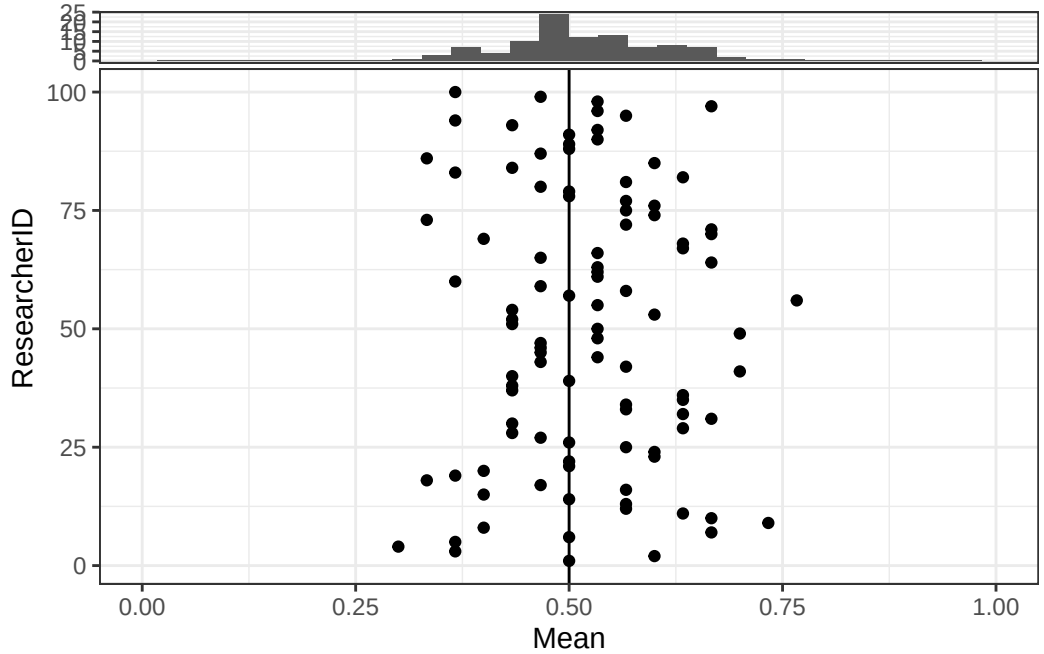
```
# A tibble: 4 x 2
  Holiday University
  <chr>   <chr>
1 Not     Yes
2 Not     Yes
3 Not     No
4 Not     Yes
```

- データの性質 (平均値など) が大きく異なる

平均値: N = 4



平均値: N = 30



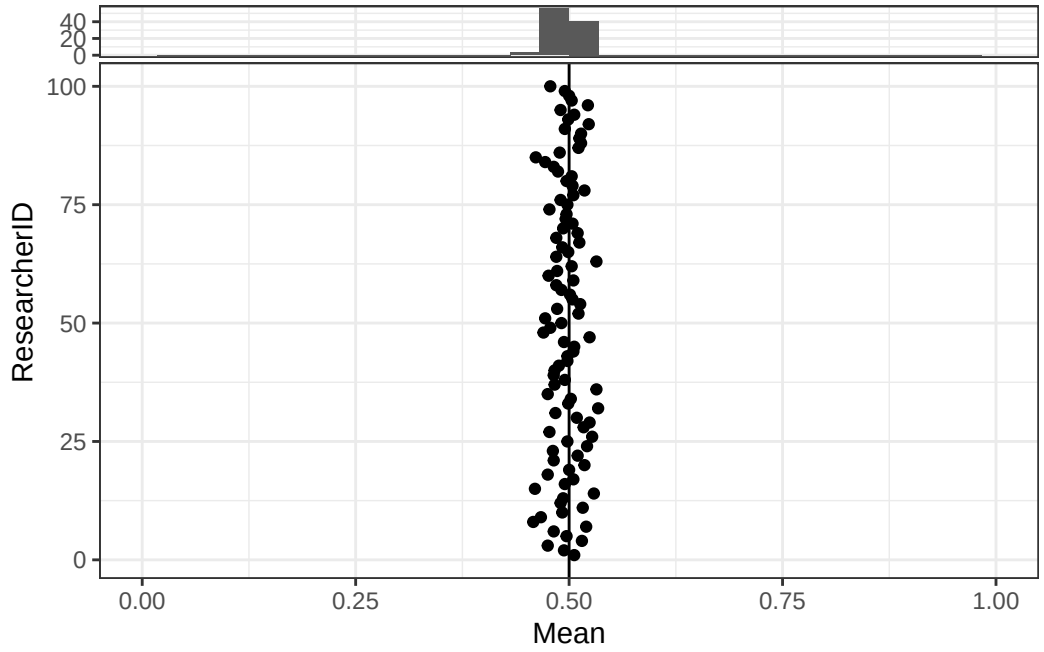
くじ引き

- どのようなデータを用いることができるのか？ ＝ 確率的 (くじ引き) で決定
  - データから生成されるモデルも確率的に決定
- 二乗誤差を評価指標として用いるのであれば、大はずしを減らしたい
  - 事例集計が有効

事例の集計

- ランダムに抽出されたデータであれば、事例数が増えれば増えるほど、サンプル平均値は母平均に近づいていく
  - **平均値**に観察されない要因が与える影響を緩和できる!!!
  - 運悪く的外れなモデルを用いてしまうリスクの緩和

平均値: N ＝ 1000



予測モデルの推定:

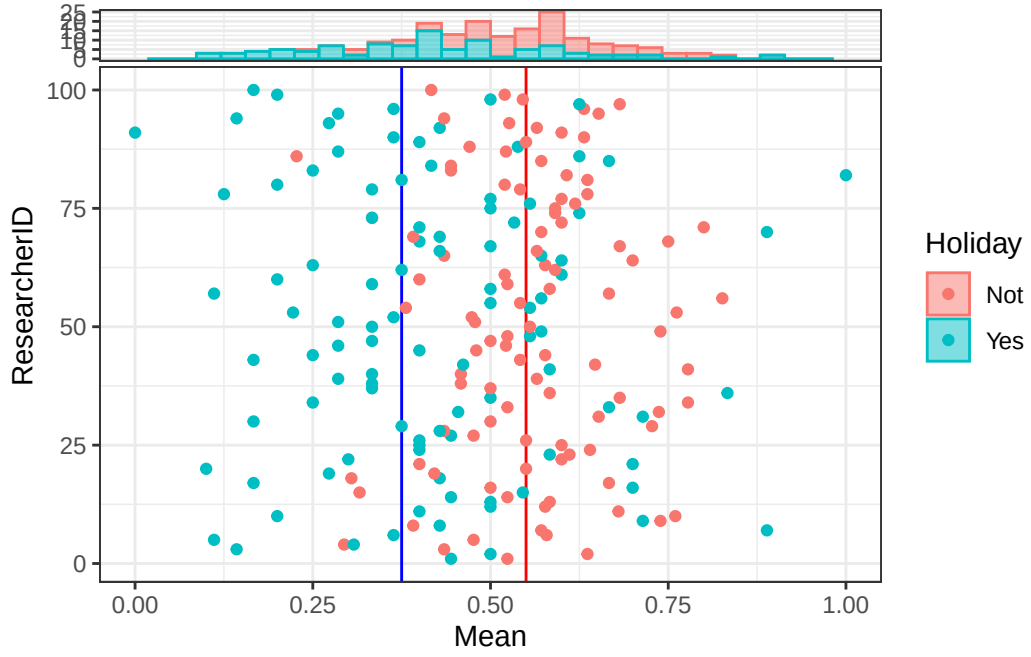
- ２つの選択肢

  - 全事例の単純平均
  - 平日/休日ごとの平均 (丸暗記 / 1 回のみ分割を行う決定木)

トレードオフ

- 単純平均は
  - 兵員多くの事例を集約でき、観察されない要素の偏りを緩和できる
  - 観察できる情報 (曜日) を未活用

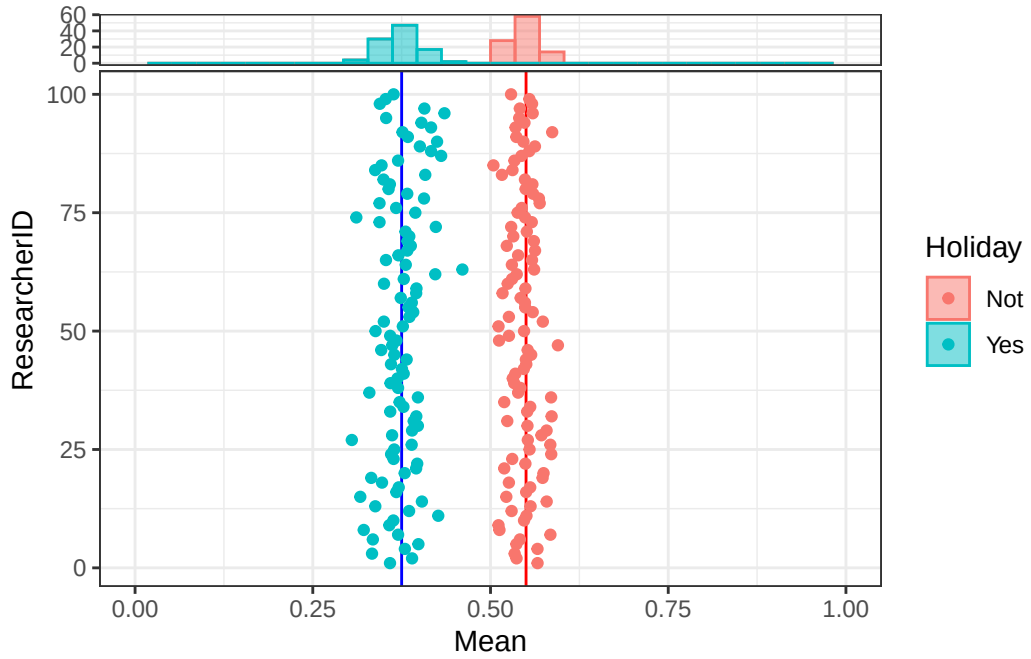
サブグループ平均値: N = 30



サンプルサイズの恩恵

- 一般に、サンプルサイズが大きければ、サンプル分割の弊害は減少する
  - データ分析版: 限界生産性低減
- より細かくサンプル分割 (複雑なモデル) しても、観察できない要因の偏りを軽減できる

サンプル増加: N = 1000



まとめ: **データへの適合 VS 事例集計**

- 一般化できる含意: X の持つ情報を完全活用するためには、
  - 事例集計により、データが偶然もった観察できない要因の偏りを緩和したい
- 複雑なモデル (X の情報をより活用したモデル) は、事例が少ないと
  - データへ適合が改善するが、データが偶然もった特徴も反映してしまう
  - 過剰適合/過学習