予測問題と予測木

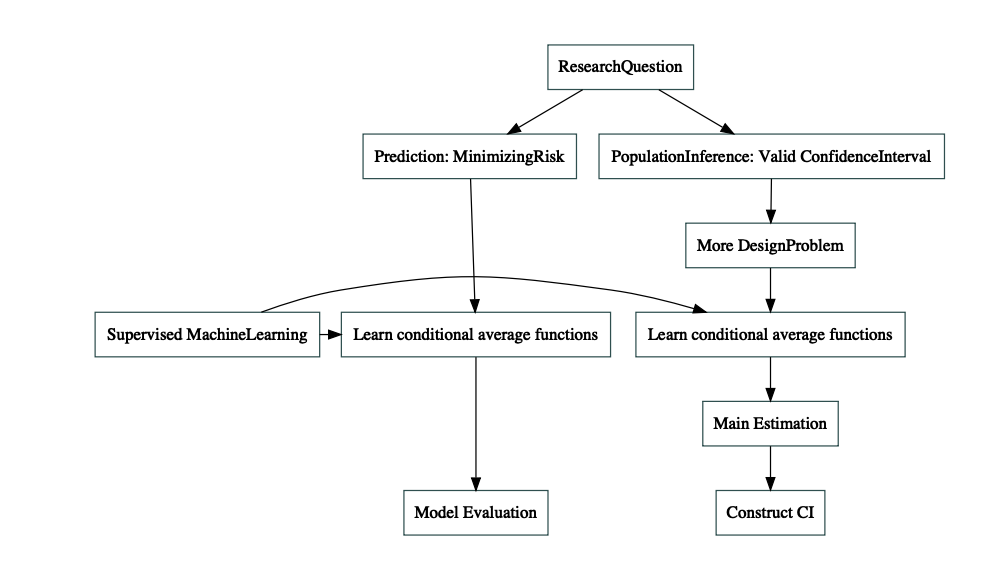
機械学習の経済学への応用

川田恵介

## 本スライドの内容

* 母平均関数に”適合する関数” を推定するアルゴリズムとして、予測木 (回帰木 | 分類木) アルゴリズムの紹介
* Motivationとして予測問題の概論を紹介
  + 母集団の推論問題は後日
* 比較対象として、Naiveなアルゴリズムも紹介

## 全体像



# 予測: 一般問題

* 教師付き学習の予測問題への応用を紹介

## 典型的問題設定

* データ が活用可能
  + ランダムサンプリング元の母集団を想定
  + **同じ**母集団から**新たに**抽出された事例について、 を予測
* データから の予測モデル を推定 (学習)

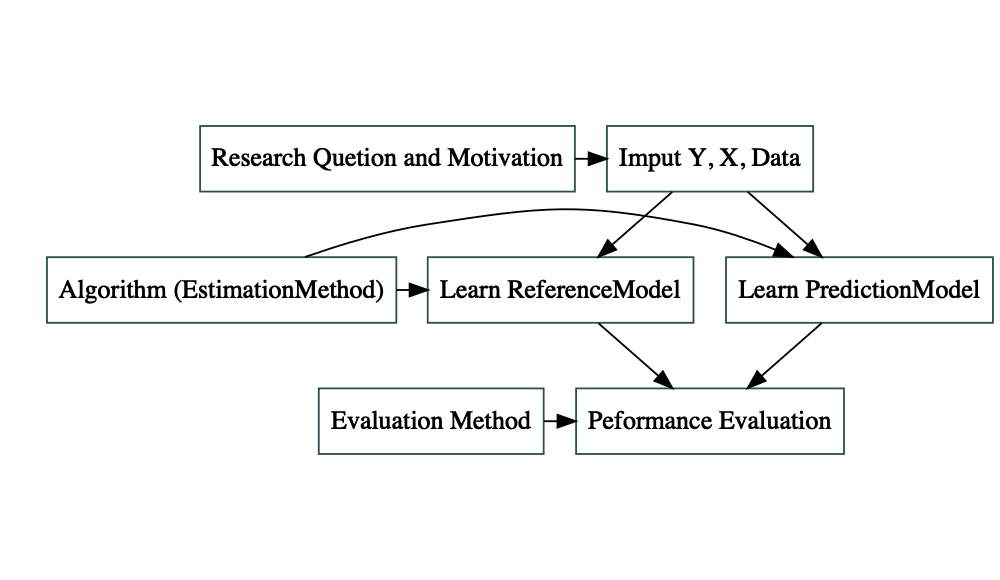
## 例

* 需要予測: 店舗の属性、気象予測、カレンダー, 販売量
* 皮膚癌: 写真、 犬 | 猫
* 滞納予測: 個人属性、 返済を滞納するかどうか
* キャッチーな議論: [予測するマシンの世紀](https://www.predictionmachines.ai/)

## 経済学における応用例

* 「新しいアルゴリズムを用いると、予測性能がこのくらい改善する」的な研究は少ない
  + 研究動機を工夫したものが多い
* [1年後生存の予測](https://doi.org/10.1126/science.aar5045) (Einav et al. 2018)
  + 「終末期医療論争」の前提条件は成り立っているのか？
* [経済モデルの評価](https://doi.org/10.1086/718371) (Fudenberg et al. 2022)
  + 「構造モデル」の評価

## Standard Prediction RoadMap



## 理想的かつ実現不可能な評価

* 論点整理に有益
* 理想的な評価は、既知の損失関数 についての母平均
* よく用いられるのは、二乗誤差

## 含意

* ただし
* 最善の予測モデル:

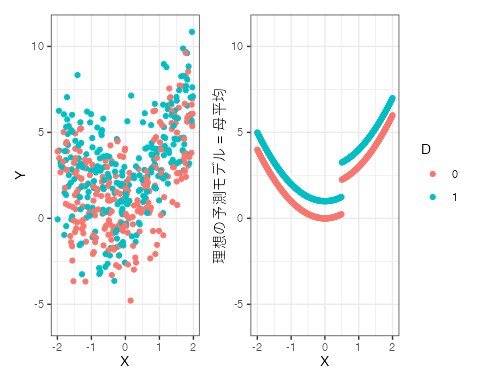
## 含意

* 母集団上で定義される評価を、データ上でどのように行うか？
  + AIC|BICなどの活用, **サンプル分割**
  + 後日
* 予想誤差 をどのように削減するか？
  + Irreducible: 有効な予測変数 が活用できるデータの探索
  + Reducible: Algorithmの改善

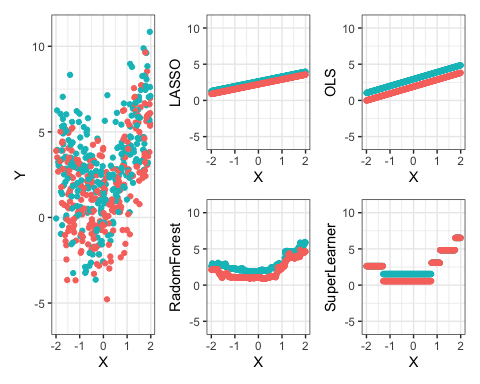
## Algorithm

* 推定手順の大枠
* データを予測モデルに変換
* 母平均に近い予測モデルを”得やすい”アルゴリズム 優れたアルゴリズム

## 数値例: 理想のアルゴリズム



## 数値例: 実際のアルゴリズム



## まとめ

* 母平均が最善の予測モデル
* 頑張って母平均を推定する

# Naive algorithm

* 単純平均法と丸暗記法

## 単純平均法

* 全データについての平均値
* は完全無視だが、大量の事例について平均を取れる

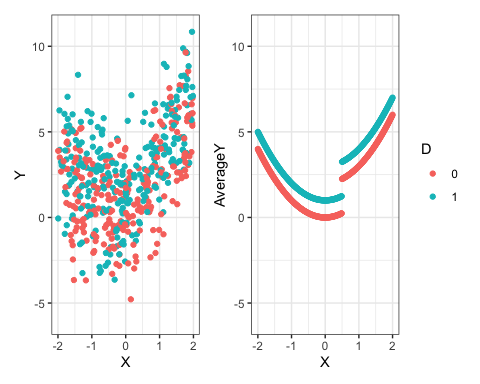
## 丸暗記法

* 全く同じ の値を持つ事例についての平均値
  + “最も近い” の事例について平均値
* 一般に、少数事例について平均

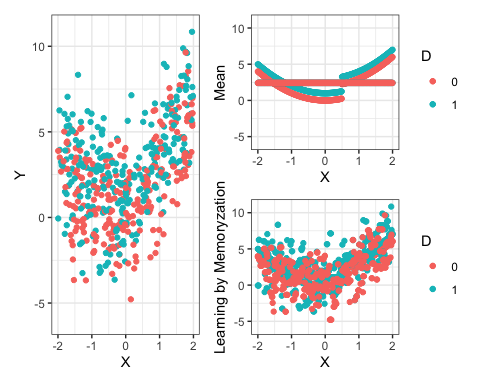
## 数値例

* から を予測
* データ生成プロセス
* , ,
* 理想の予測モデル:

## 数値例



## 数値例



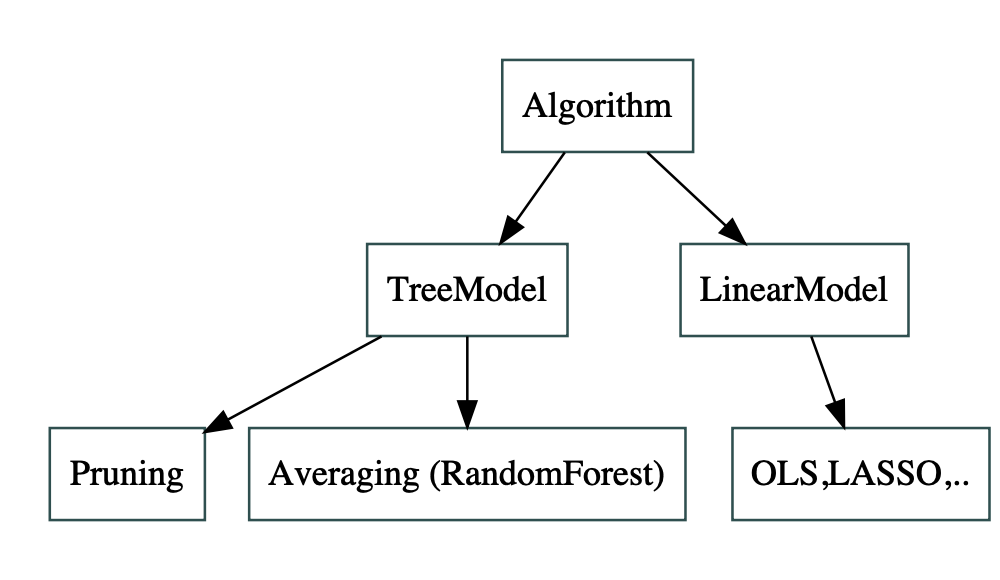
## まとめ

* 単純平均法の問題点: “一定”の予測値を決めうつ、荒い近似
  + と との関係性を完全無視
* 丸暗記法の問題点: 平均値の推定に、“個人差”が強く反映
  + の予測年収は?
  + データにおける最も近い事例が、大谷翔平だと？？？
* 予想: 中間的Algorithmが良さそう

# 予測木アルゴリズム

* 非常に”透明性が高く”教育的なアルゴリズム
  + コンセプトが明快、モデルが可視化できる場合も
  + 重要な論点を抑えられる

## 全体像



## 予測木アルゴリズム

* サブグループの”平均値”を予測値とする
  + 伝統的方法: 人間がサブグループを決定
  + 本講義: データがサブグループを決定
* トリビア: であれば回帰木、 であれば分類木|決定木 と呼ばれる

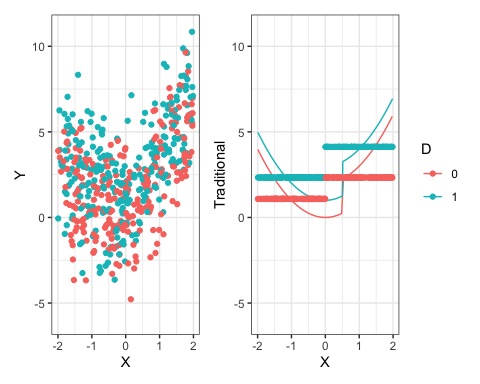
## 伝統的方法

* データを見る前に推定する(有限個のパラメータからなる)予測(母平均)モデルを設定
  + パラメータのみをデータによって決める
* 例:

## 推定方法: Empirical Risk Minimization

* データ上のLossを最小化するように推定: であれば、
* 伝統的アプローチでは、OLS | サブサンプル平均と一致

## 数値例



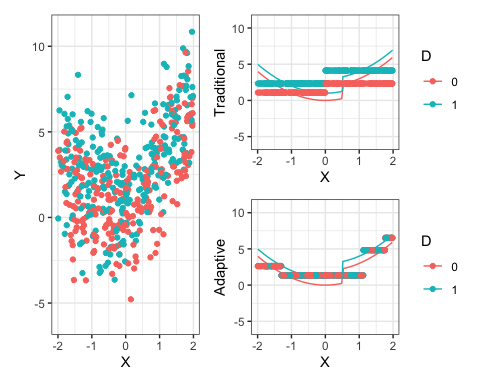
## Adaptive Tree

* 伝統的方法: 分析者によるモデル設定 Empirical Risk Minimizationによる推定
  + モデル設定にパフォーマンスが大きく依存
  + 適切なサブグループ分けは非常に困難
* Adaptiveな推定: サブグループ分けにも、 Empirical Risk Minimizationを活用

## Recursive Partition アルゴリズム

1. データ、 **停止条件(最大分割回数等)**
2. 第1分割: Empirical Riskを最小化するグループ分割 (通常2分割)を探索
3. 第2分割: 第1分割の結果を**所与**として、Empirical Riskを最小化するグループ分割を探索
4. 停止条件に達するまで、分割を繰り返す

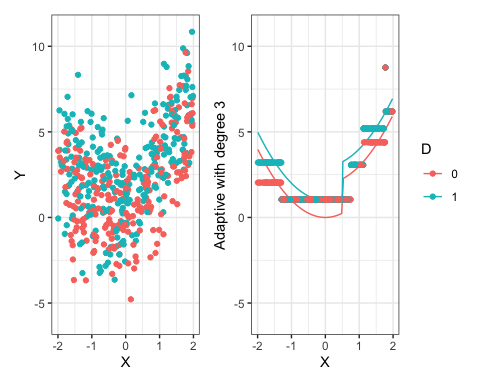
## 数値例



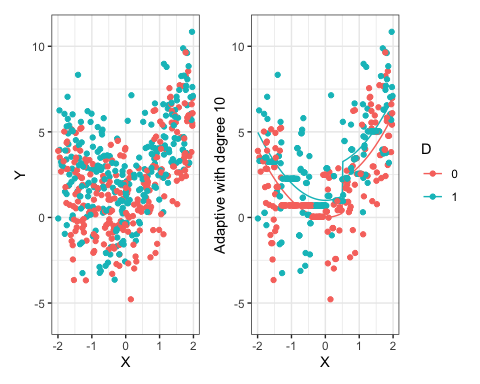
## 停止条件

* 停止条件をどう決める？
  + rpart関数での初期値: 最小サンプルサイズ = 20, 最大分割数 = 30など
* 推定されたモデルやパフォーマンスが決定的に左右される
* NaiveなIdea
  + 現実は複雑なので、単純なモデルはよくない
  + Empirical Risk Minimizationを適用

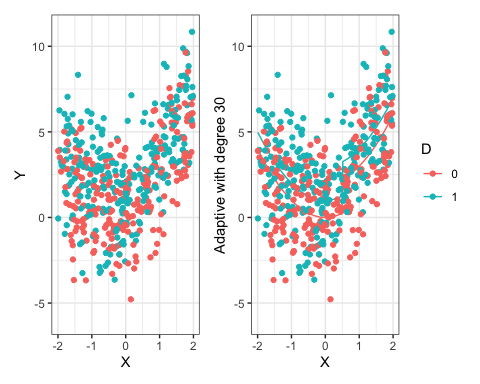
## 例



## 例



## 例



## まとめ

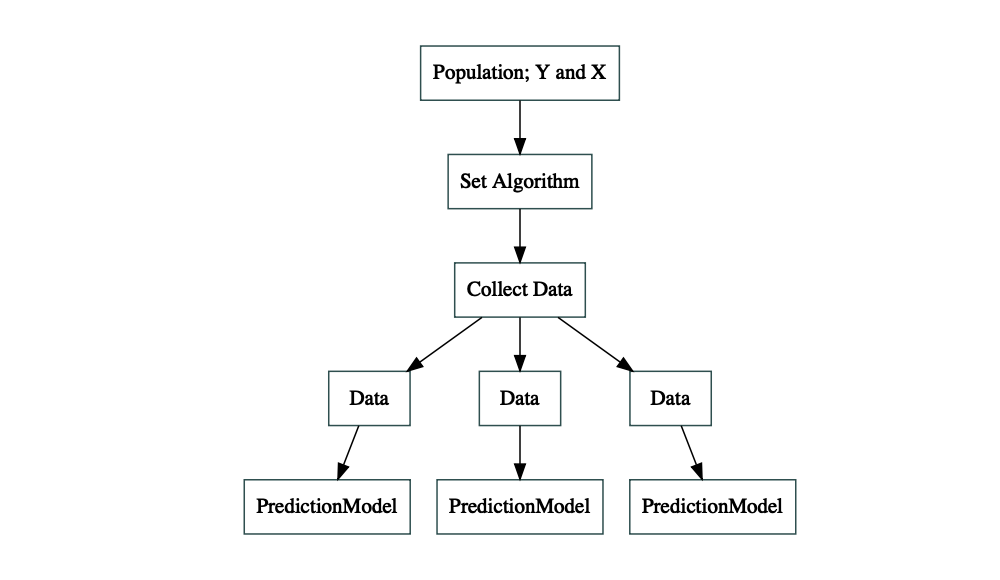
* 伝統的な予測木 有限個のパラメータを推定
  + 多くの潜在的パラメータを0と決めうち
* Adaptiveな予想木 サブグループをデータに合うように生成
  + 潜在的に無限個のパラメータを推定
* 停止条件に決定的な影響を受ける

# 過剰適合(過学習)問題

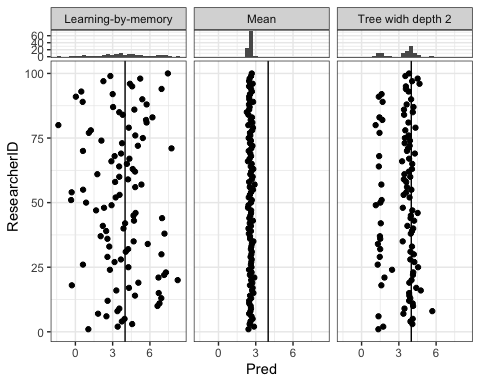
## とりあえず頭に入れること

* ある母集団を予測する上で、優れたアルゴリズム (停止条件の設定を含む) を選びたい
  + 常にうまくいくアルゴリズムは存在しない
* 次善の策は、上手く**いきやすい**アルゴリズムを用いてる
* 通常の統計学と同じ脳内モデルが有益

## 脳内モデル



## 数値例



## Decomposition

* 無限大のサンプルサイズで学習した結果得られる “仮想的な”予測モデル

## トレードオフ

* モデルを複雑化 (より多くの分割)を行うと、現実は極めて複雑なので

## Estimation Errorの源泉

* Signalのみを取り出せる人がいれば、全て解決
  + 目の前の香川出身38歳男性の所得を聞き、香川出身38歳男性の平均所得と個人差を分割できる？
* 伝統的戦略は、大量の事例の平均を取る
  + 漸近性質の活用
* 複雑なモデルは、
  + の影響を強く受け、Estimation Errorが上昇する

## 例

* 単純なモデル
* 複雑なモデル

## Empirical Risk Minimizationの問題

* 理想は母集団上でRiskを最小化する
* できないのでEmpirical Risk Minimization

## 過剰適合 (過学習)

* が の影響を**受けるほど**、小さくなる
  + 丸暗記モデルでは0!!!!
* 一般にデータと無矛盾なモデルを推定することは難しくない
  + データに過剰に適合する(から過剰に学んだ)モデルであり、予測性能は悪い
* 新しい論点: Benign Overfiting (Hastie et al. 2022; Bartlett et al. 2020)

## 過剰適合 (過学習)の弊害

* データの”偶然の偏り” (平均値からの大きな乖離があるサンプル) に影響されてしまう
* Empirical Riskでは正しく評価できない
  + 丸暗記でも勉強することはいいことだ！！！

## まとめ

* 通常の統計学と同様にSampling Uncetainlyを頭に入れる必要がある。
* 矛盾する戦略
  + 大量の事例の平均をとる (Estimation errorの削減)
  + 多くのサブグループを作る (Approximation errorの削減)

## 引用

Bartlett, Peter L., Philip M. Long, Gábor Lugosi, and Alexander Tsigler. 2020. “Benign Overfitting in Linear Regression.” *Proceedings of the National Academy of Sciences* 117 (48): 30063–70. <https://doi.org/10.1073/pnas.1907378117>.

Einav, Liran, Amy Finkelstein, Sendhil Mullainathan, and Ziad Obermeyer. 2018. “Predictive Modeling of u.s. Health Care Spending in Late Life.” *Science* 360 (6396): 1462–65. <https://doi.org/10.1126/science.aar5045>.

Fudenberg, Drew, Jon Kleinberg, Annie Liang, and Sendhil Mullainathan. 2022. “Measuring the Completeness of Economic Models.” *Journal of Political Economy* 130 (4): 956–90. <https://doi.org/10.1086/718371>.

Hastie, Trevor, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. 2022. “Surprises in High-Dimensional Ridgeless Least Squares Interpolation.” *The Annals of Statistics* 50 (2): 949–86. <https://doi.org/10.1214/21-AOS2133>.