概要
機械学習

川田恵介
東京大学
keisukekawata@iss.u-tokyo.ac.jp

2025-09-09

1 概念整理

1.1 データ分析法

- 過去の経験や事例、歴史（データ）から学ぶ方法
- 意思決定に役立つ情報提供を目指す
  - 各顧客は、どのようなサービスを好みのか?
  - 事業全体で価格を上げると、どの程度需要が下がるのか?
  - どのような事業領域が伸びているのか?

1.2 現状

- 学術/実務機関から高い関心[1]
- 参考文献: 事例紹介 + 考察
  - 予測マシンの世紀　AI が駆動する新たな経済

1.3 近年の発展

- 機械学習: 予測研究への応用がメイン
  - 機械学習 + 計量経済学/医療統計等を身につけることで、因果効果や格差研究等への応用も可能に
    - 学部/大学院レベルの教科書[2]が書かれる段階になっている

1.4 到達目標

- 自身でデータ分析を行うための入門

[1] Amazon, Cyber Agent, Microsoft, Mizuho, Netflix, Uber
[2] Applied Causal Inference Powered by ML and AI

- データ分析の不完全性への対処に焦点

- データ分析の結果を意思決定に活用するための必要知識の取得
- エントリーシートなどに、“AI/機械学習を用いた不動産市場についての、予測・因果モデルを作成経験した経験あり”と書けるようにする

2 Get started

2.1 一般的課題

- データ分析を含む事例分析の共通課題
  - 独立した分析チームに**同じ市場/社会**について、事例収集とデータ分析を依頼したとしても、同じ結論に到達し得ない
    - 誰がやっても、「水は 100 度で沸騰する」という結論に到達する理科の実験とは対照的

2.2 データ

- 現代の分析では、大量の事例が含まれるデータを利用する機会が多い
  - 実習では、11311 の中古マンション取引事例
    - 全ての事例を人間が直接認識することは不可能

2.3 “個別事例分析”

- 特徴的な事例を研究者が選別する
- 最高価格で取引されている物件は

Price	District	Size
1,400	杉並	105

- “杉並区の 105 平米の物件の取引価格は 1 億 4000 万円” と主張できるか?

2.4 “個別事例分析”の問題点

- 全く同じ属性を持つ物件は、他にも存在する

Price	District	Size

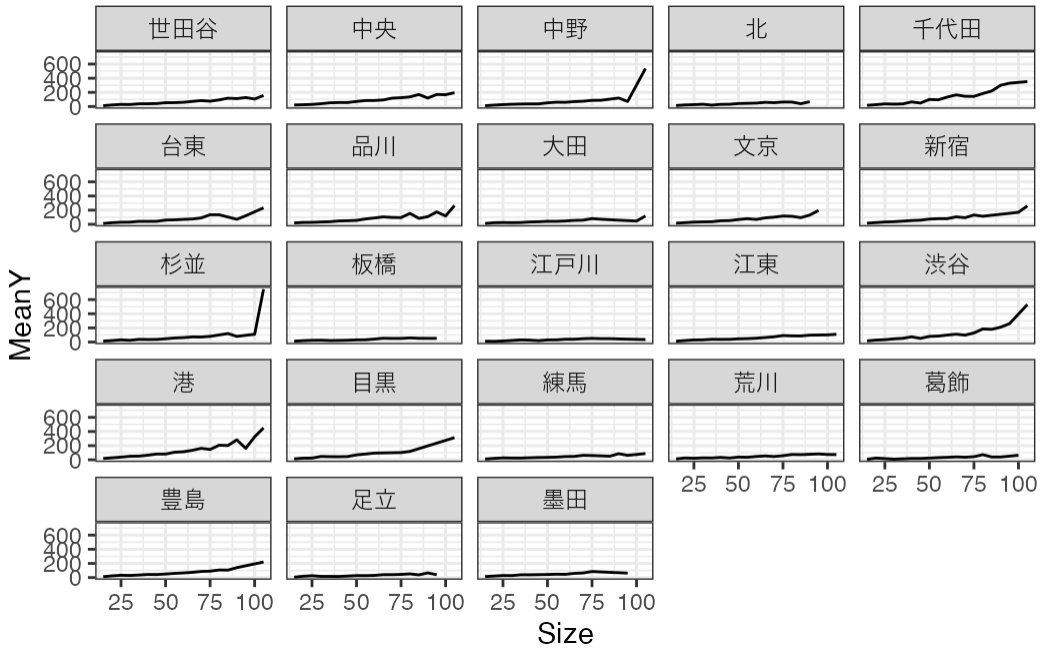
- データから観察できない要因で、先の事例は大幅に上振れたのでは?

2.5 要約計画

- データにおいて観察できない要因の影響を減じるために、
  - 何らかの要約を用いて、集団の”傾向”を論じる

  - 典型的な事例の報告、研究者の所見/印象、平均値、中央値、分散
- 恣意的分析を避けるために、データを見る前に、要約計画を立てることが重要
  - 平均などの”自動計算”できる方法が有効

2.6 実例: 平均取引価格



2.7 実例

- 1 事例のみで平均値を計算している組み合わせが 24 組存在 (例: 千代田区 95 平米)
  - 上振れ/下振れの恐れが高い
- さらなる要約が有益

2.8 伝統的アプローチ

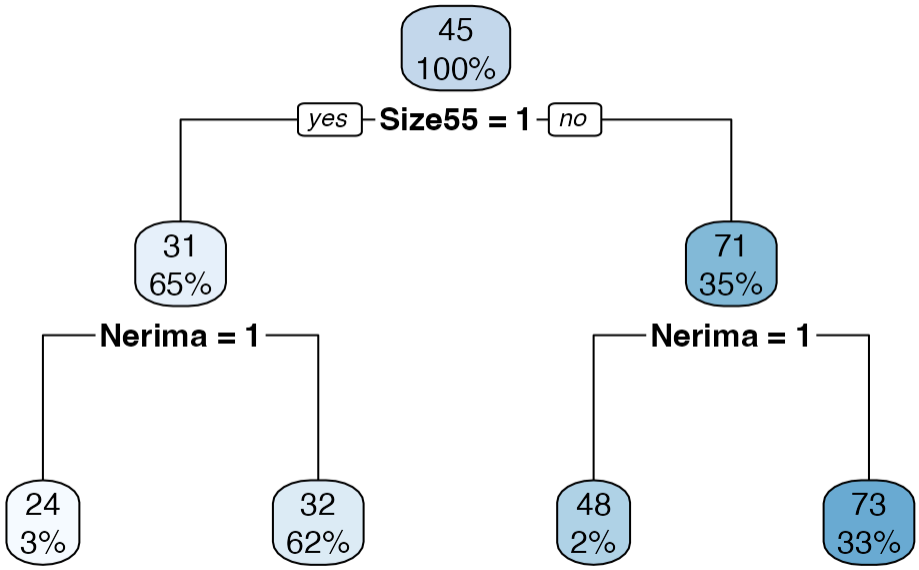
1. **分析者**が、サブグループ (モデル) を定義する

- 例: 練馬区かどうか × 部屋の広さが 55 平米以下かどうか

2. サブグループの平均値を計算

- 集計により偶然の上振れ/下振れの影響を軽減する

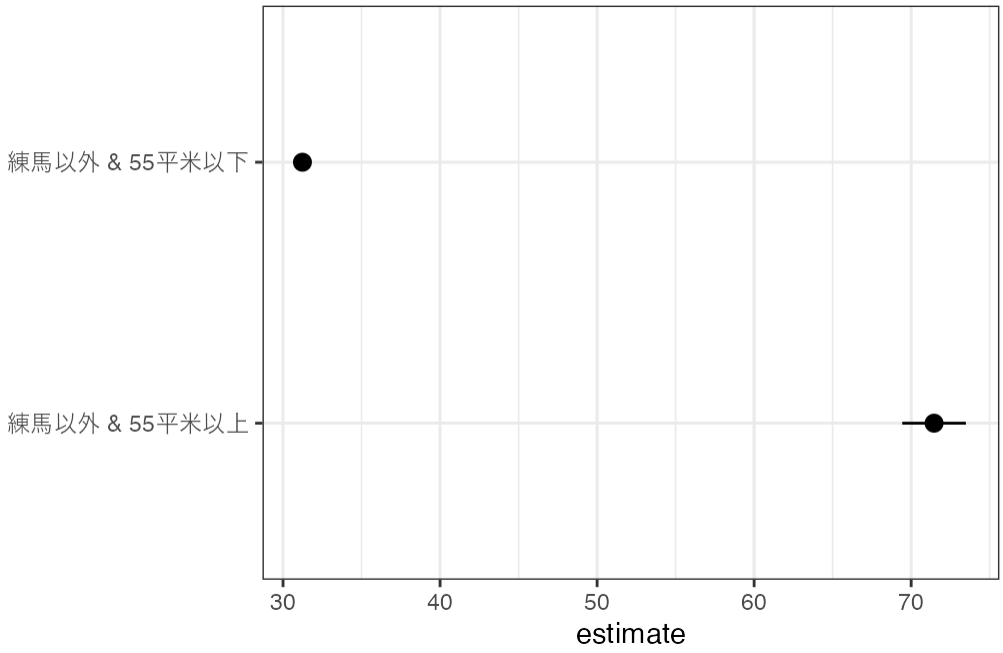
2.9 伝統的アプローチ



2.10 伝統的アプローチの利点

- データの偶然の上振れ、下振れを定量化できる (信頼区間)
  - 一定の確率で、“無限大”の事例数で計算された平均値(母平均)を含む区間
  - 偶然生じた傾向か否かを区別でき、重大な意思決定への活用において、特に望ましい性質

2.11 例: 信頼区間



2.12 モデルの設定

- 事例の要約方法(モデル化)が属人的で、不透明
  - 推定結果を大きく左右しうる
    - 多くの教科書で直接的な言及を避けてきた問題
      - 「理論や背景をしっかり踏まえて、適切なモデル化を行うべし」以上の提案が難しい
    - 機械学習の適切な活用が有効

2.13 付論: モデルの設定

- モデルを色々試すことで、分析者にとって”望ましい結果” (例: 明確な因果効果、存在しない格差の存在”証明”など)を、“捏造”できる
  - Cherry-pick/p-huck などと呼ばれる
- 無実の証明が難しい
  - 疑われないようにするために、“複雑な分析”を避ける傾向
    - 例: どのような背景属性の組み合わさると、改装は大きな効果を持つか?

2.14 データ主導のモデリング

- 決定木アルゴリズム: **データ**に最も適合するように、サブグループを定義する

- よりデータ主導
- 明確な基準(“データへの適合”)のもとで、要約方法を決定

- 機械学習の得意分野

2.15 例

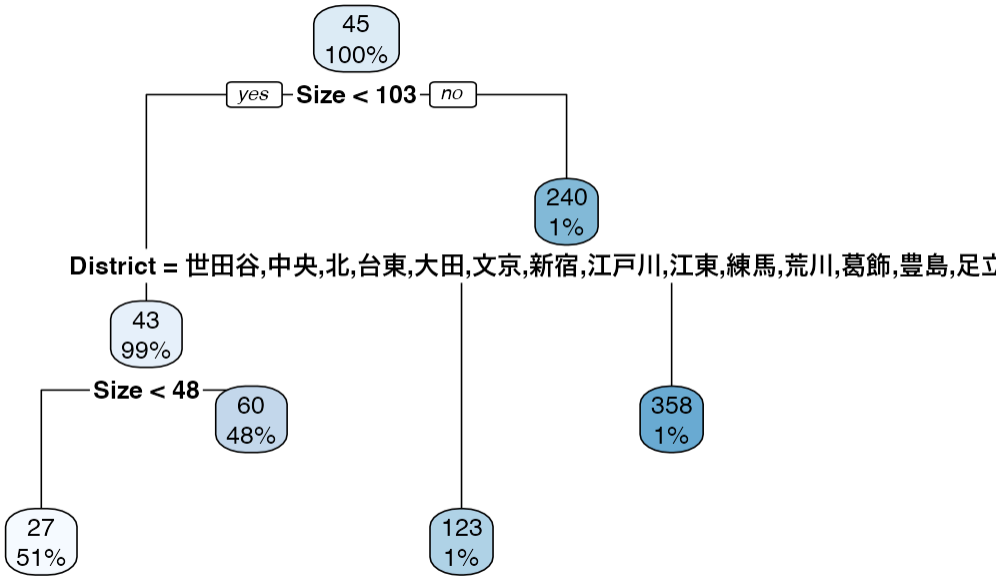
- 最大 4 グループに分けることは前提に、平均二乗誤差

$$(Y - \text{モデルの予測値})^2 \text{のデータ上での平均値}$$

を可能な限り削減するようにグループ分けを行う

- 近似的に削減する (Greedy-algorithm)

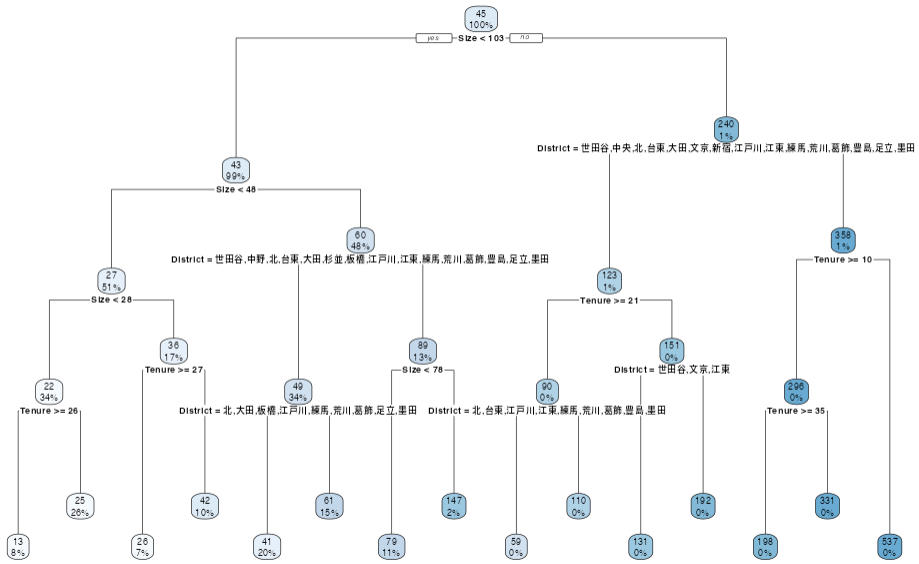
2.16 実例



2.17 例. 複雑なモデル

- 機械学習を用いれば、より複雑なモデルも容易に生成できる
- 例. 最大 32 グループに分ける

2.18 複雑なモデル



3 Takeaway

3.1 伝統的アプローチの利点と課題

- **利点:** 研究者が設定するシンプルなモデルについて、データの偏りの影響を評価しながら、推定できる
  - 無限大の事例数で推定した場合について、概ね正しい推測ができる(信頼区間)
- **課題:** 予測性能は、研究者が設定したモデルに強く依存する

3.2 機械学習の利点と課題

- **利点:** データ主導でモデルを推定できる
- **課題:** 予測性能は、研究者が設定したモデルに強く依存する
- “何らかの予測モデル”は推定できるが、社会/市場分析への活用が難しい
  - 人間行動など個人差が大きい変数を予測することは難しい
  - データとの関係性が複雑なため、信頼区間などを計算することも難しい

3.3 本講義の提案

- 機械学習と伝統的な統計学/計量経済学/医療統計等の手法と”併用”する

3.4 予習資料

- 入門的な統計学 + R

  Statistical Inference via Data Science

- ガッチリ機械学習
  - Introduction to Statistical Learning (Trevor Hastie, Robert Tibshirani, Jerome Friedman)

4 R

4.1 R VS Python

- Python と並ぶ、データ分析の人気言語
  - 高い透明性と拡張性、再現可能性、無料、AI 活用
  - 多様な統合開発環境(IDE)

4.2 準備

- Posit cloud への登録
  - クラウド環境で R を使用できる
  - ただし時間制限あり
- 関心がある人は、自身の PC へ Rstudio をインストール
  - 時間無制限

4.3 Example code

- コードを実行する際には、(慣れるまでは)、以下の手順を徹底

1. ctr + a を押し、全ての行を選択する
2. ctr + enter を押し、実行する

4.4 Error が出たら

- 「error は必ず起きる」、という心構えをもつ
- 再現性の確認： 全ての行を再度実行
  - コード実行しわすれ、がエラーの原因となることが多い
- よくあるミス (大文字/小文字の区別、 コンマ)を確認
  - 極力予測変換を活用し、タイポを減らす
- 解決できない場合は、コード全体をチェット欄にコピペしてください