中間まとめ: 予測と把握

機械学習

川田恵介 (keisukekawata@iss.u-tokyo.ac.jp)

Table of contents

1 意思決定

- “予測モデルの推定” 以外を目標とするデータ分析も一般的
  - 大量の事例の特徴を人間が理解できるようにする (記述モデルの推定)
- 活用したい意思決定問題に応じて、しっかり区別することが重要
  - しばしば研究者も混同してきた

1.1 機械学習による予測モデル推定

- データを構成する事例単位について、個別予測を提供できる
  - 自身の転職後の賃金、物件ごとの中古価格、店舗ごとの需要予測
- 伝統的な方法 (比較的単純なモデルを推定する) に比べて、複雑なモデルを適切に推定できる
  - 過剰適合を緩和できる

1.2 予測例

Simple	LASSO	Size	Tenure	Distance	District
76.8	70.7	60	23	7	CBD
84.5	78.7	65	26	4	CBD
38.4	29.9	50	49	12	目黒区
38.0	41.2	60	38	3	大田区
64.5	61.0	65	5	7	大田区
58.0	61.8	55	26	1	世田谷区
79.8	75.0	75	21	7	豊島区
54.0	56.1	60	9	5	荒川区
38.7	35.6	70	38	8	板橋区
43.0	44.9	70	12	13	江戸川区

1.3 適した意思決定

- 活用の前提: 十分な過去事例をデータとして活用できる
- 予測したい” 事例数” が少ない、“日常的” な意思決定に有効 (“ミクロな意思決定”)
- 例: 個別物件の不動産取引

  - 売り手/買い手ともに、取引対象の物件の取引価格を予測できることが有益
    * 予測モデル自体を理解する必要性が低い
  - 大量の事例取引がされており、事例数も十分

1.4 予測の実務への応用

- 以下が重要
  - 応用したい重要な意思決定問題は何か?
  - どのような Y を予測したいのか?
  - どのようなデータが活用できるのか?
- 現状、“AI” では判断不可能

1.5 練習問題

- 問題: あるサブスク制動画配信サービスにおける「User の視聴履歴やいいね数を X」とした予測モデルを構築したい。
  - サービスの持続的発展のために、どのような Y を予測すべきか?
    * hint: ある動画をクリックし、視聴を開始するかどうか**ではない**
- 実例: Netflix のデータによるデータ分析コンペ
  - kaggle

2 “マクロ” な意思決定

- 大量の事例に影響を与えるような意思決定
  - 企業の経営戦略の決定、政府の政策決定、投票決定
- 「事例ごとの大量の予測値」ではなく、「事例集団の” 大雑把な” 特徴を捉える集計情報」(記述モデル) の方が有益なケースが多い

2.1 例

- 中期経営計画: 就業者や株主、世間に伝えやすい、集計情報に基づいて、説明を行なっている
  - セブン&アイ

  - 資生堂
- 白書: 有権者等に向けて、集計情報に基づいた、現状分析結果を説明している
  - 経済財政白書
- “統計学の母”

2.2 Recap: 予測モデル

- “事例ごと” に X から欠損情報 Y を予測する
  - 極力多くの事例を使い、$X-Y$ の” 過去の” パターンを抽出し、予測に活用
- 事例ごとに大量の数字 (予測値) が出力される

2.3 “マクロな” 意思決定への応用

- 影響の範囲が広い (“マクロな”) 意思決定に対しては、個別事例の予測値” そのもの” の便益は限定的
  - 複数の情報を組み合わせた、人間による意思決定が要求されがち
    - 影響を与える (大量の) 事例の特徴について、意思決定者が理解できる情報提供が必要
    - 大量の予測値を示されても、理解できない
- 幅広い合意形成には、事実の共有が重要

2.4 例

- ミクロな意思決定: ある物件をどの程度で買い取るか?
  - 予測モデルによる価格予測
    - 目の前の物件について、予測値を活用すれば良い
- マクロな意思決定: 支店網の再編戦略
  - **全**物件について、**各々**の予想取引価格が提供できたとしても、意思決定者が理解できない
  - 市場の現状を把握できる、人間が理解可能な集計情報が有益

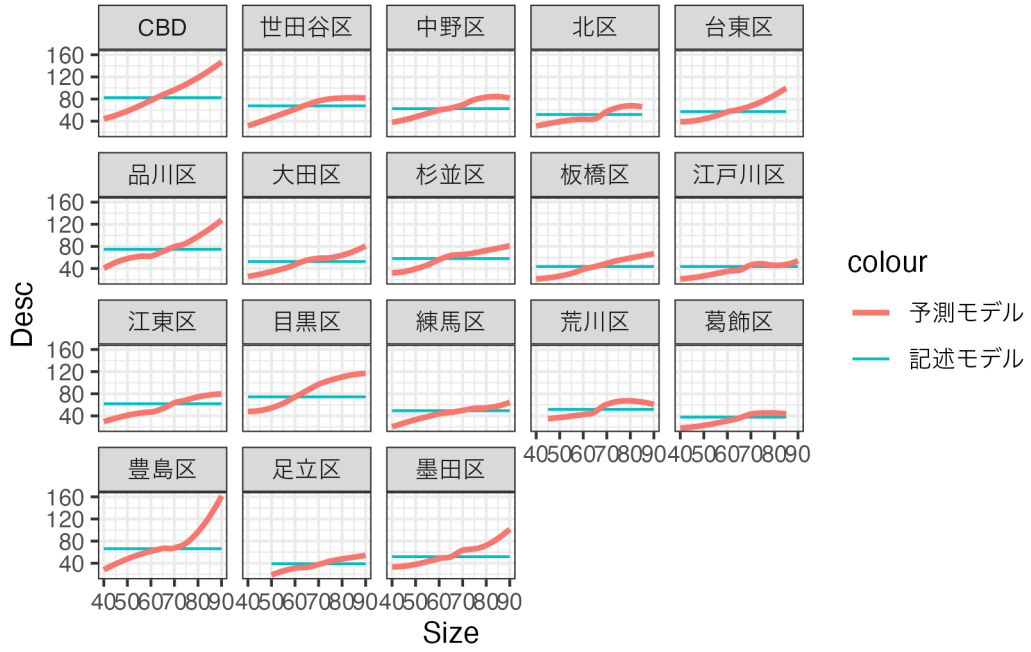
2.5 用語: 記述モデル

- 地区ごとの平均値など、人間が把握できる程度に簡単なモデルは記述モデルと呼ばれる
  - 人間の把握を手助けするモデル

    - よく似た動機のモデル: ビジネスモデル

- 人間が理解できる程度に単純なので、良い理論的性質を目指す

2.6 記述モデル VS 予測モデル



2.7 実例

- 支店網再編計画を議論するために、各地域の平均取引価格を把握したい
- いくつかの推定方法が考えられる
    - 地域ごとに平均取引価格を計算
        - 伝統的な方法
    - 予測モデルを推定し、予測値の平均を計算
- 各地域について事例数が十分 (200 事例以上) あれば、前者がおすすめ

2.8 実例

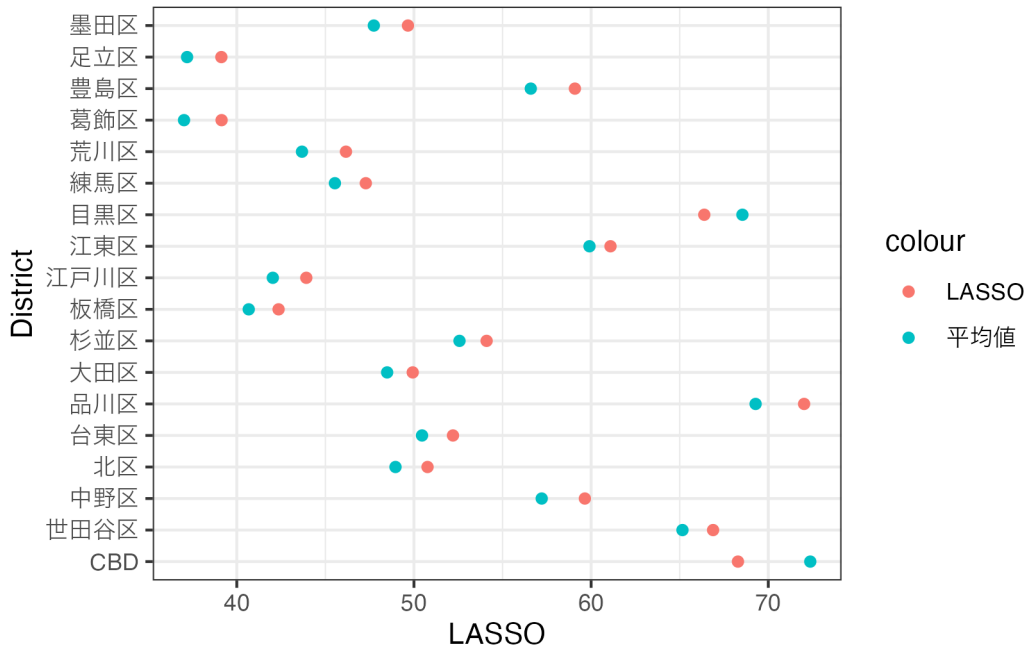
Price	LASSO	District
59	52	墨田区
44	47	墨田区

84	89	CBD
79	89	CBD
32	51	CBD
16	10	墨田区
70	76	墨田区
48	65	CBD
75	71	CBD
90	84	CBD

- Price と LASSO、どちらの平均値を用いるべき?

2.9 実例

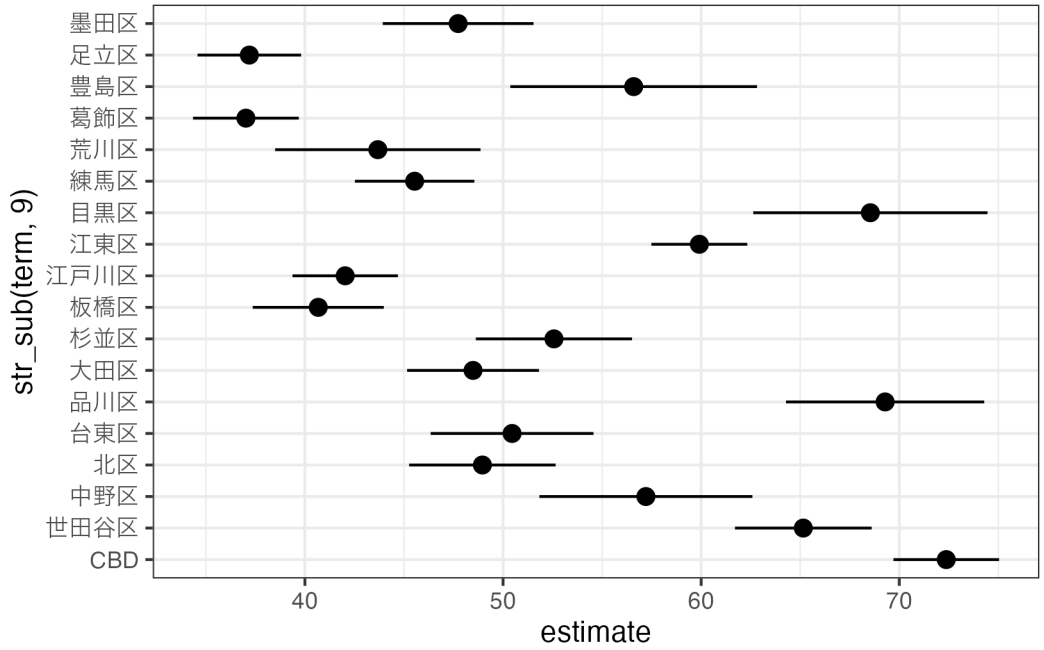
- 実際のデータで計算すると、かなり乖離がある



2.10 Yの平均値の利点

- ある程度の事例数を用いて計算された平均値は、“誤差の範囲”(信頼区間) を示すことができる
  - 母平均が含まれているであろう範囲
- 一般に影響が甚大なマクロな意思決定においては、判断の根拠となる数字は慎重に取り扱う必要がある
  - 誤差の範囲を示すことは重要
    - (機械学習を用いて算出した) 予測値の平均では難しい

2.11 信頼区間



3 まとめ

- 意思決定に応じて、必要な情報の” 細かさ” が異なる
  - “細かさ” が異なれば、最適な手法が異なる可能性がある

3.1 対比例

- Price 以外に、Size, District, Tenure, Distance が活用可能なデータ
- 予測モデル: 全ての変数と機械学習を活用した複雑な Price の予測モデル
- 記述モデル: 各 District の平均価格
  - $X = District$ であり、各 X の組み合わせについて、十分な事例数を確保できる

3.2 事例ごとの予測

- 大量の X と機械学習を用いた複雑な予測モデルが有効
  - モデル全体の予測性能は評価できる

- 限られた X のみを使用した平均値は不十分
  - 例: 立地している区のみでは、予測性能が低い
    - 同じ区であったとしても、大きな価格差があるため

3.3 事例全体の特徴把握

- 限られた X と OLS や平均値を用いた単純な記述モデルが有効
- 大量の X と機械学習を用いた複雑な予測モデルはあまり有効ではない
  - 大量の値があり、人間が理解できない
  - 予測値を集計しても、信頼区間が計算できない
    - 予測値と” 母平均” との乖離を評価することが難しい
  - 目的と比べて、過剰に複雑なモデルを推定している