決定木アルゴリズム

データへの適合

川田恵介

概念: 予測モデル

- 予測モデル (関数) : 指定された X について、Y の予測値を自動回答する” 機械 (AI)”
  - 江古田駅から徒歩 5 分、1LDK の中古マンションの予想販売価格
- データを用いて、作成

コア・アイディア

- 全ての事例は、他の事例にも適用できる一般的な特徴 (シグナル) とその事例だけが持つ一般化できない特徴 (ノイズ) をもつ

1. モデルをデータに適合されることで、データの持つ” シグナル” を抽出
2. 事例を集約することで、データの持つ” ノイズ” を除去
3. 1 と 2 をバランスさせるモデルの複雑さを、独立したデータへの適合で決める

例

- 学園祭出店の仕入担当: 去年の事例から、仕入れ量を決定
  - X : 記録されている属性 (販材、規模)、Y : 販売量
- 今年出展する出店 $X = \{$ わたあめ, 2 名 $\}$ の販売量を予測
- チャレンジ: 一般に $Y =$ "X"共通要因 + 観察できない要因
  - 可能な限り” 共通要因” の特徴を捉えたい!!!

データ主導のモデリング

- アルゴリズム $\simeq$ 推定手法
- 決定木 $=$ 優れた出発点
  - **データに適合させる**モデリング法を学ぶ

予測モデル (関数)

- X が決まれば、Y の予測値が決まる
- $g(X)$ と表現
- 分析者による決定 $+$ データによる決定
  - 現代的アプローチ: より多くをデータが決定

サブグループ分析

- “もっとも” よく使われるデータ活用方法
- 属性 X についてサブグループ A を定義し、予測モデル $g(X)$ をサブグループ平均として定義

例

```
# A tibble: 6 x 4
   Size Location DistanceStation Price
  <int> <chr>              <int> <int>
1    87 練馬                  18     7
2    58 練馬                  19     9
3    20 練馬                   1     5
4    53 板橋                  10     5
5    62 板橋                  14     9
6    33 板橋                  10     9
```

- $\{$ 練馬 & 広さ $= 30$ & 駅からの距離 $= 2$ 分 $\} \simeq$ 予測販売価格?

例: 立地

```
# A tibble: 6 x 6
```

	Size	Location	DistanceStation	Price	SubGroup	Prediction
	<int>	<chr>	<int>	<int>	<chr>	<dbl>
1	87	練馬	18	7	練馬	7
2	58	練馬	19	9	練馬	7
3	20	練馬	1	5	練馬	7
4	53	板橋	10	5	板橋	8
5	62	板橋	14	9	板橋	8
6	33	板橋	10	9	板橋	8

- { 練馬 & 広さ $= 30$ & 駅からの距離 $= 2$ 分 } $\simeq 7$

例: 立地 & 広さ

```
# A tibble: 6 x 6
```

	Size	Location	DistanceStation	Price	SubGroup	Prediction
	<int>	<chr>	<int>	<int>	<chr>	<dbl>
1	87	練馬	18	7	練馬 & Size > 50	8
2	58	練馬	19	9	練馬 & Size > 50	8
3	20	練馬	1	5	練馬 & Size <= 50	5
4	53	板橋	10	5	板橋 & Size > 50	7
5	62	板橋	14	9	板橋 & Size > 50	7
6	33	板橋	10	9	板橋 & Size <= 50	9

- { 練馬 & 広さ $= 30$ & 駅からの距離 $= 2$ 分 } $\simeq 5$

二乗誤差最小化

- サブグループ平均 $=$ 最もサブグループに” 適合する” 値
- 各サブグループ内の平均二乗誤差 (Mean Squared Error) を最小化する

$$E[(Y - g(X_i))^2|A]$$

例

```
# A tibble: 6 x 9
```

	Size	Location	DistanceStation	Price	SubGroup	Mean	Max	MSE_Mean	MSE_Max
	<int>	<chr>	<int>	<int>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	87	練馬	18	7	練馬	7	9	0	4
2	58	練馬	19	9	練馬	7	9	4	0

3	20	練馬	1	5	練馬	7	9	4	16
4	53	板橋	10	5	板橋	8	9	9	16
5	62	板橋	14	9	板橋	8	9	1	0
6	33	板橋	10	9	板橋	8	9	1	0

“伝統的” VS Data adaptive modelling

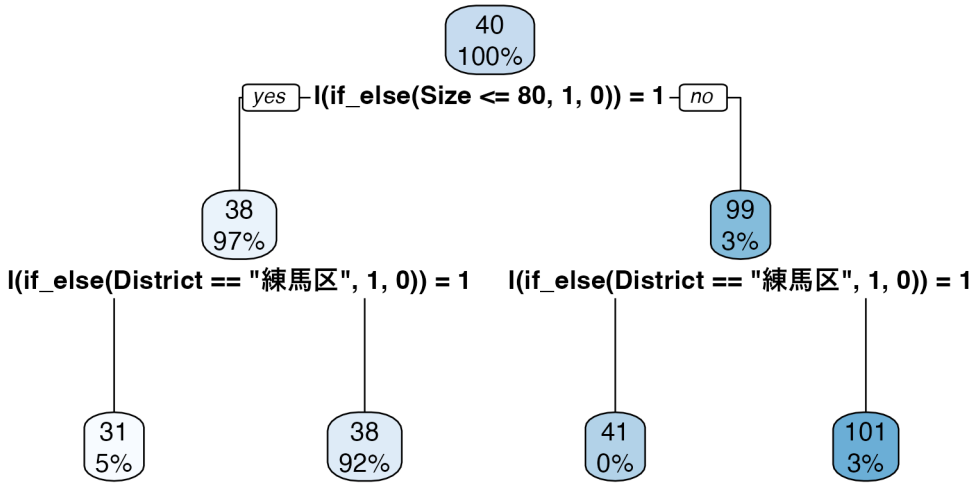
- 伝統的アプローチ: 研究者が事前 (データを見る前) にサブグループを定義
  - サブグループごとの予測値のみ、データに適合するように決定
- Data adaptive: グループ分けもデータに適合するように決定
- (注) “Bad practice” : 研究者がデータ $\{Y, X\}$ を見ながらサブグループを決定

伝統的アプローチ

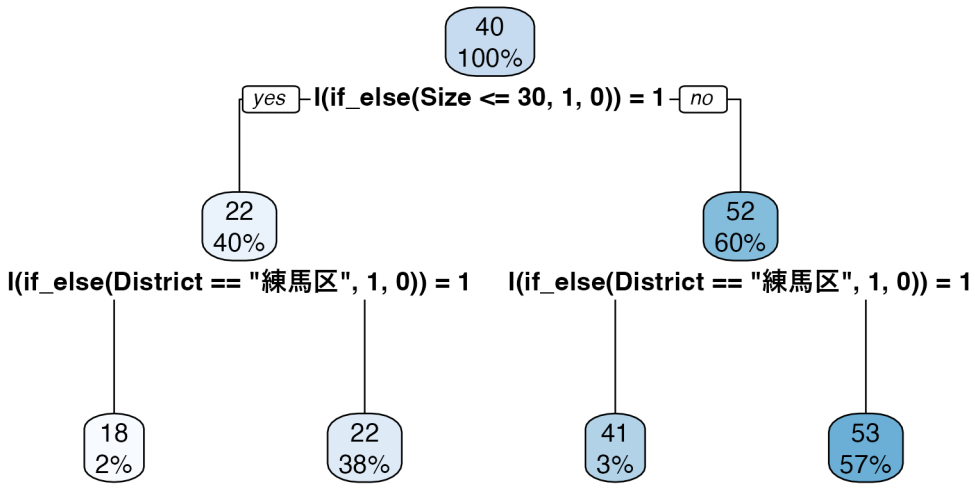
1. 研究者が事前にサブグループを定義
2. 各サブグループについて、サンプル平均を計算し、予測モデルを構築

- 樹形図で可視化可能

実例: 伝統的アプローチ



実例: 伝統的アプローチ



伝統的アプローチの問題点

- サブグループ分けに決定的に依存する
- 予測研究においては、サブグループを定義する際の、**Practical** guide line が限られている
  - 比較・因果研究であれば、研究課題により自動的に決まる部分がある (例: 大卒高卒間賃金格差 = 少なくとも大卒/高卒でグループ分け)
- 予測結果は、グループの定義に決定的な影響を受ける
  - 予測モデルに活用しない変数も指定する必要がある

Data adaptive アプローチ

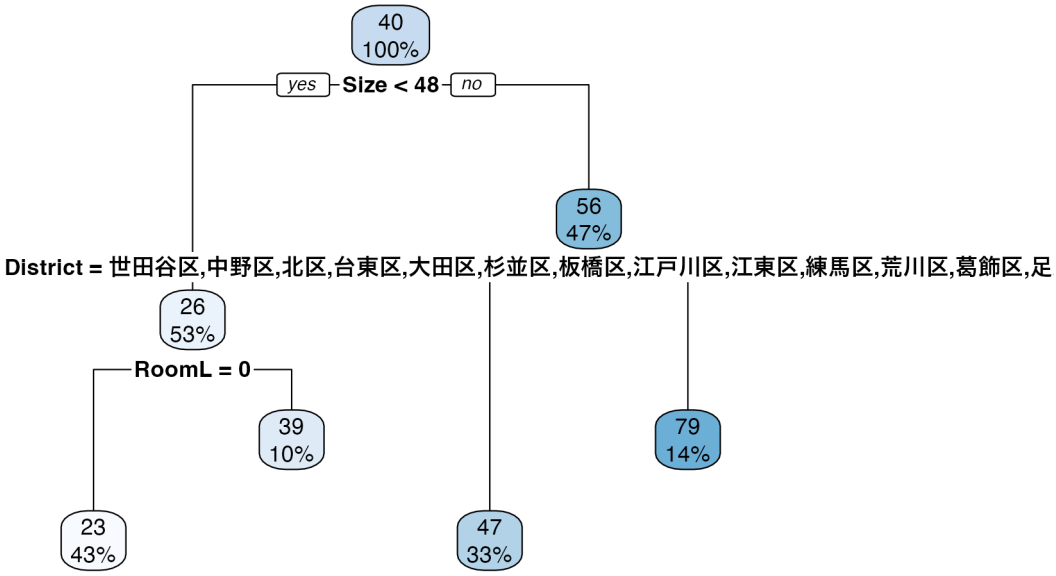
0. 停止条件 (最大いくつのサブグループを作るかなど) を設定
1. データに適合するように、サブグループを設定
2. 各サブグループについて、サンプル平均を計算し、予測モデルを構築

- 課題: 具体的には？

貪欲な (Greedy) アルゴリズム

0. 停止条件を設定
1. 2 分割する: データ内二乗誤差を最小化するように一つの変数、閾値を選ぶ
2. 1 度目の分割を” 所与” として、 2 度目の分割を行う
3. 停止条件に達するまで、繰り返す

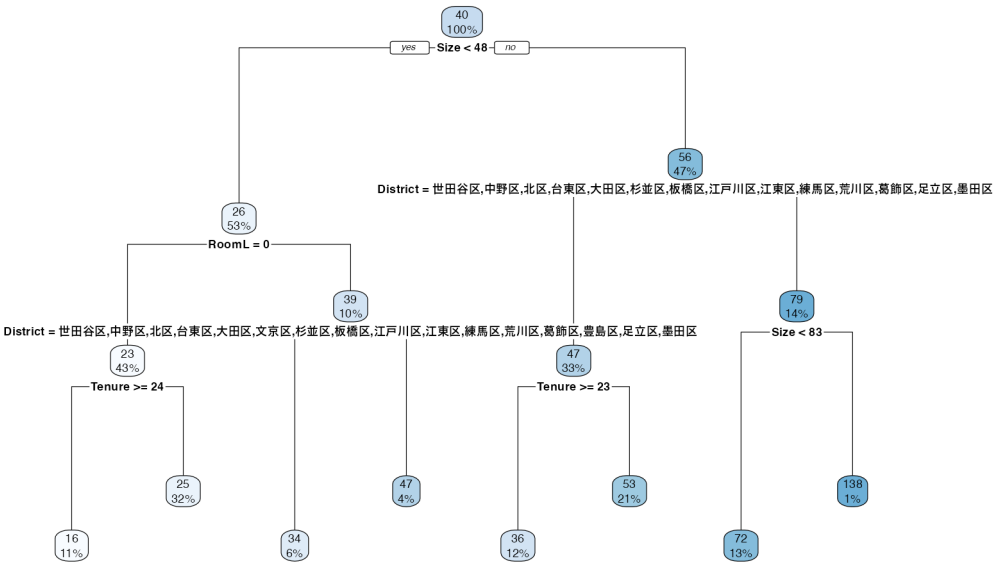
実例: 停止条件 $= 2$ 回分割



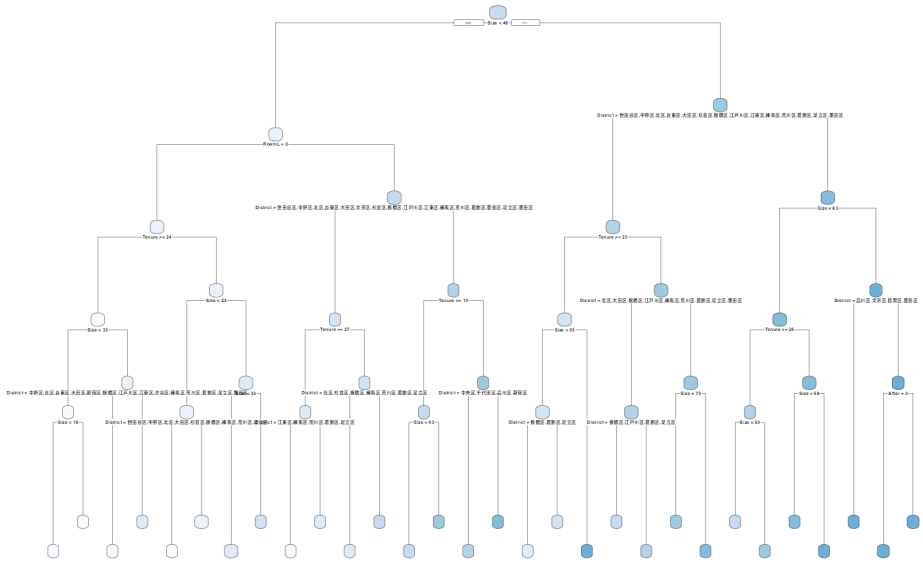
Data adaptive アプローチの課題

- モデルが停止条件に決定的な影響を受ける
- 停止条件を緩める (最大分割回数を増やす, 最小サンプルサイズを減らす) と巨大な (複雑な) 決定木が生成される

実例: 停止条件: 3 回分割



実例: 停止条件: 5 回分割



Data adaptive アプローチの課題

- “停止条件をどのように決める?”
- 異なる条件のもとで、モデルの試作と中間評価を繰り返し、最善の条件を探す
  - データへの当てはまり
  - 独立して抽出されたデータへの当てはまり
  - 理論的評価指標 (AIC など)

“データへの適合” の限界

- データに適合するように停止条件を決めると何が起きるか?
  - 最もデータに適合する予測木は?

複雑すぎる予測モデル

- データに適合するように停止条件を決めると、丸暗記方予測モデル (極めて複雑な予測モデル) が出来上がる
- X の数が十分あれば、全く同じ X をもつ事例はデータ内に一つしかない決定木 (巨大な決定木) が生成できる
  - データとの矛盾がなくなる
  - 予測値 $=$ 最も近い 1 事例の値

まとめ

- 「データに適合させる」は、データ分析の基本戦略
- 現代の PC を使えば、サブグループ分け自体もデータによって決められる
- **厳重注意**データに適合するように停止条件 (モデルの複雑さ) を選ぶと、極めて複雑・データに適合するが、予測性能が劣悪なモデルが出来上がる
  - 現実が複雑であったとしても、一般に劣悪なモデルが出来上がる
- なぜか?
  - 事例の集計ができなくなるため