数理統計 補助資料
～非線形回帰・非線形分類～
（多項式回帰と決定木）

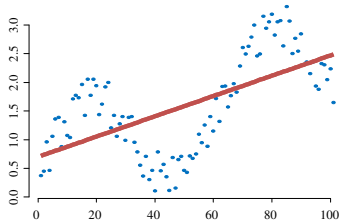
2024年度2学期：　月曜1限, 水曜3限

担当教員：　石垣　司

非線形回帰・非線形分類

非線形構造を持つデータからの情報抽出や予測のための回帰分析や分類

- **得意：予測・分類の精度を上げる。ルールに基づいた回帰**
- **対象外：実証分析**
- **キーワード：スプライン回帰, 動径基底関数を用いた回帰, カーネル法を用いたSVM（サポートベクターマシン）, ランダムフォレスト, ガウス過程回帰, ニューラルネットワーク, 深層学習（Deep Learning）, などなど**



例：非線形構造を持つデータ

線形回帰モデルではデータの傾向をとらえきれていない

非線形モデルによる回帰とは？ #1

線形モデルによる回帰分析

- **データ（説明変数）とパラメータ（回帰係数）の線形結合の関数でモデル化される線形モデルを用いた回帰分析**
- **例：線形回帰モデル**

$$y = f(\boldsymbol{x};\boldsymbol{\theta}) = b_0 + b_1 x_1 + , \cdots , + b_P x_P$$

- **例：ロジスティック回帰モデル**

$$\Pr(y = 1|\boldsymbol{x}) = f(\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + , \cdots , + b_P x_P)}}$$

線形結合

ロジスティック回帰モデルやポアソン回帰モデルの出力は非線形関数であるが, 説明変数と回帰係数の関係は線形結合

#メモ1 線形モデル、線形回帰モデル、線形モデルによる回帰の用語はややこしいので要注意。
#メモ2 本授業では取り扱わないが非線形モデルを用いた判別は非線形判別と呼ばれる。

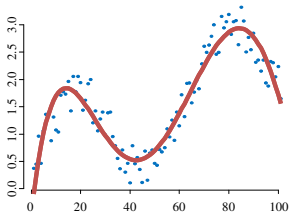
非線形モデルによる回帰とは？ #2

非線形モデルによる回帰分析

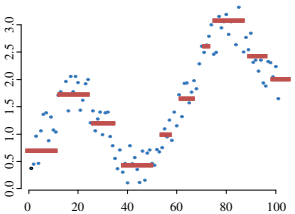
- **線形モデルによる回帰以外の回帰分析**
- **例：多項式回帰モデル**

$$y = f(\boldsymbol{x};\boldsymbol{\theta}) = b_0 + b_1 x + b_1 x^2 + , \cdots , + b_M x^M$$

- **例：決定木**
  説明変数 x に対するIF-THENルールで目的変数 y を説明



多項式回帰モデルによる回帰



決定木による回帰

補足：ロジスティック回帰モデルは非線形回帰の手法？

答え：非線形回帰の定義次第で異なる

非線形回帰の定義1：出力が非線形関数

～線形モデルによる回帰分析～
- 線形回帰モデル

～非線形回帰モデル～
- ロジスティック回帰モデル
- 多項式回帰モデル
- 決定木
- 深層学習モデル

非線形回帰の定義2：説明変数と回帰係数の線形結合の関数による回帰モデル

～線形モデルによる回帰分析～
- 線形回帰モデル
- ロジスティック回帰モデル

～非線形回帰モデル～
- 多項式回帰モデル
- 決定木
- 深層学習モデル

#メモ ただの呼び名の話なので、データ分析においてあまり本質な議論ではない

多項式回帰

```
Rのコード
x = seq(0,10,0.1)
set.seed(1111)
y = sin(x) + 0.2*x +
0.8*runif(length(x))
data = data.frame(rbind(y,x))

reg = glm(y~poly(x, degree = 50, raw
= TRUE),data,family="gaussian")
reg$aic

plot(y)
lines(fitted(reg))
```

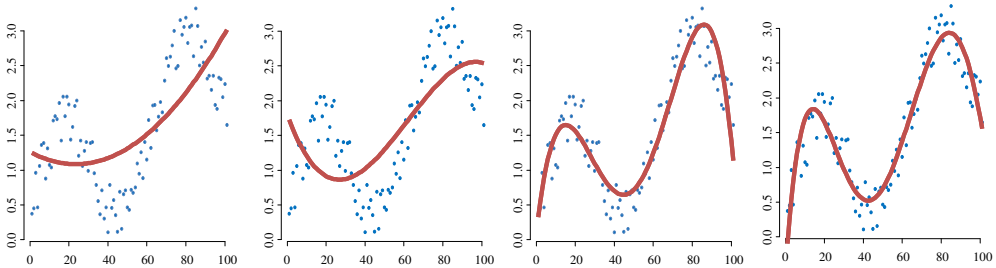
目的変数の多項式で表現される回帰モデル

– M 次の多項式回帰モデル

$$y = f(x, \boldsymbol{\theta}) = b_0 + b_1 x + b_2 x^2 + b_3 x^3 +, \cdots, +b_M x^M$$

回帰係数は最小2乗法で推定可能

非線形構造をもつデータの傾向に合わせた関数で回帰できる

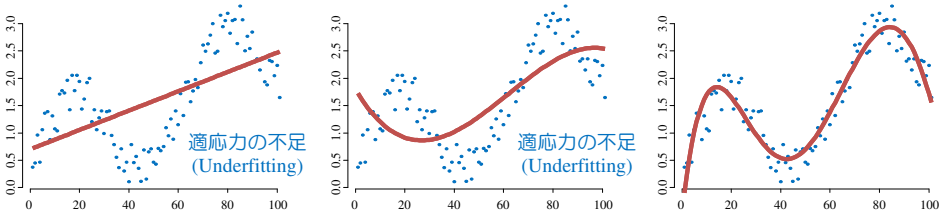


2次の多項式モデル　3次の多項式モデル　4次の多項式モデル　5次の多項式モデル

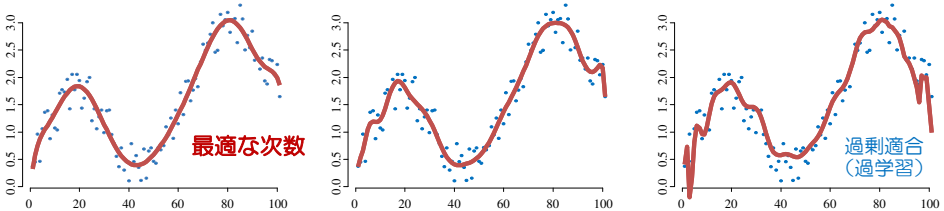
多項式回帰モデルのモデル選択

AICを用いることで最適な多項式の次数を選択可能

– **複雑な関数による過剰適合（過学習）を回避できる**



1次の多項式 (AIC = 215.5)　3次の多項式 (AIC = 196.0)　5次の多項式 (AIC = 47.1)



10次の多項式 (AIC = 7.8)　20次の多項式 (AIC = 19.0)　50次の多項式 (AIC = 99.9)

決定木

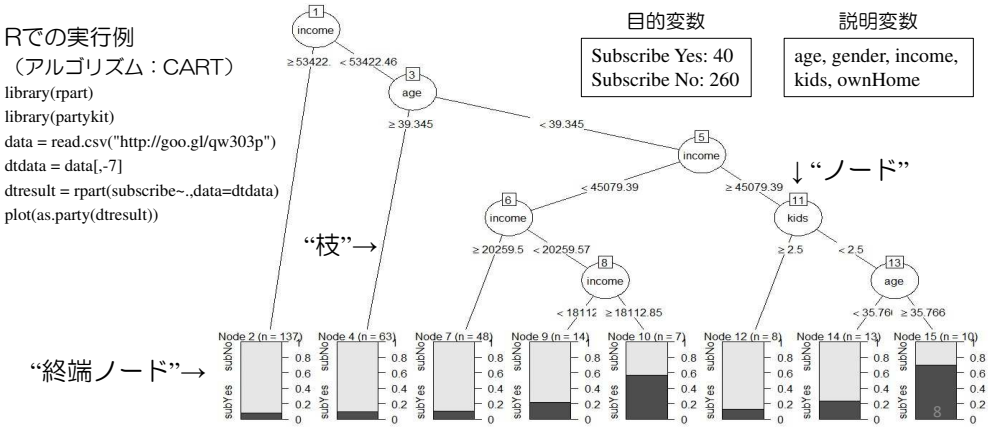
IF-THENルールを用いた非線形回帰・非線形分類の手法

– **回帰木：目的変数が連続変数の場合の決定木**

– **分類木：目的変数が2値変数の決定木**

Rでの実行例
（アルゴリズム：CART）

```
library(rpart)
library(partykit)
data = read.csv("http://goo.gl/qw303p")
dtdata = data[,-7]
dtresult = rpart(subscribe~.,data=dtdata)
plot(as.party(dtresult))
```



決定木生成のアルゴリズム

代表的な決定木生成アルゴリズム

- CART (Classification and Regression Tree, 1984)
  回帰・判別, 2分岐, 分割基準はGini分散指標
- ID3, C4.5, C5.0 (J.R. Quinlan, ID3, 1986)
  判別, 多分岐(ID3は2分岐), 分割基準はエントロピー

決定木生成に必要な基準とプロセス

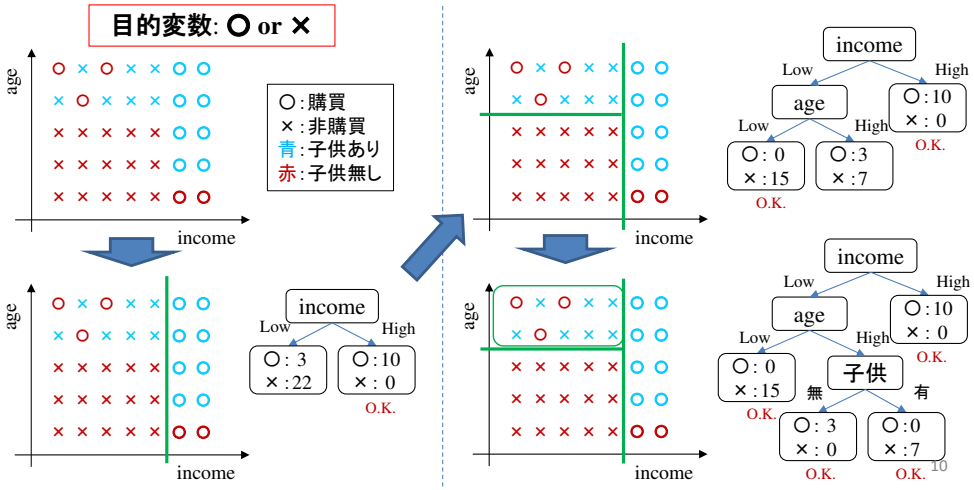
- 1. **木構造の生成：データ分割によりノードと枝を作成**
- 2. **分割基準：良い分岐点を決める**
- 3. **枝刈り：過学習を防ぐ**

#メモ：数多くの決定木作成アルゴリズムが提案されている。
各アルゴリズムの違いの詳細には触れないが、上記の1~3は共通して必要

木構造生成のイメージ

データの分割によりノードと枝を順々に生成

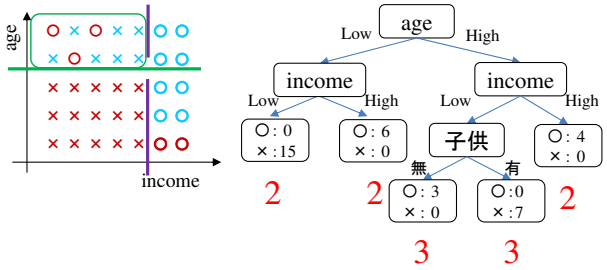
- **空間の分割を繰り返すことで, 非線形な領域に分類**



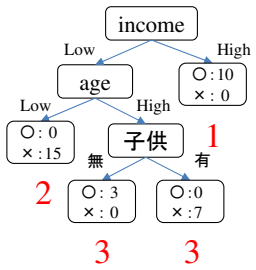
良い決定木とは？

単純な構造(コストが小さい)が望ましい

- **コスト：全終端ノードに至るまでのノードの数の合計**
- **データ分割のパターンは膨大だが, なるべくコストの小さい決定木を生成したい**



コスト=2+2+3+3+2=12

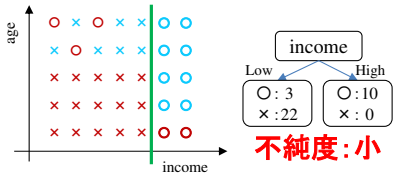


コスト=2+3+3+1=9

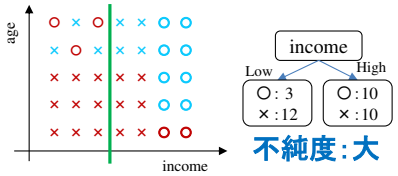
決定木の分割基準

目的変数の違いが明確になるよう分割

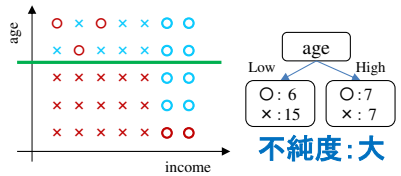
- **目的変数の一様性を不純度として定量的に定義**
- **不純度が小さくなるように分割**



～説明変数内での比較～



～説明変数間での比較～



不純度の例：Gini分散指標

#メモ：Gini分散指標はGini係数とも呼ばれている。
経済統計でのGini係数の定義とは値の大小関係が逆なので注意

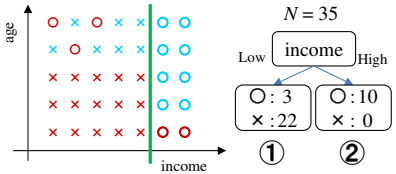
Gini 分散指標(GI)

※Gini係数：経済統計における各集団間の格差の指標

– $GI = 1 - \sum_c^M \left(\frac{1}{N}\sum_i^N x_i^{(c)}\right)^2$

c : クラスの変数 (下例では、$c = \{1,2\}, M = 2$)

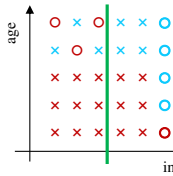
$x_i^{(c)}$: データ i がクラス c に属していたら 1, otherwise 0.

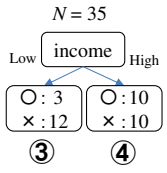


①のGI = $1 - \left(\frac{3}{25}\right)^2 - \left(\frac{22}{25}\right)^2 = 0.21$

②のGI = $1 - \left(\frac{10}{10}\right)^2 - \left(\frac{0}{10}\right)^2 = 0$

分割の不純度 $=0.21 \times \left(\frac{25}{35}\right) + 0 \times \left(\frac{10}{35}\right) = 0.15$





③のGI = $1 - \left(\frac{3}{15}\right)^2 - \left(\frac{12}{15}\right)^2 = 0.32$

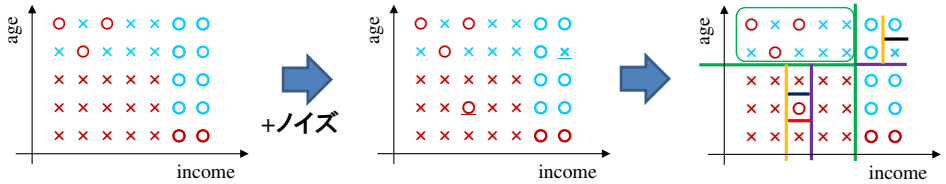
④のGI = $1 - \left(\frac{10}{20}\right)^2 - \left(\frac{10}{20}\right)^2 = 0.5$

分割の不純度 $=0.32 \times \left(\frac{15}{35}\right) + 0.5 \times \left(\frac{20}{35}\right) = 0.42$

決定木～枝刈りとは？

木構造の過学習

– データ分割を進めると既知データは100%正しく分類できるが，将来得られるデータの予測・判別精度は低下



100%正しく分類

枝刈りによる過学習の防止

– 事前枝刈り

ある基準で木構造の生成を途中で止める。高速に木構造を生成

– 事後枝刈り

大きな木構造を作成した後で，不要な枝を刈る。精度が良い

事後枝刈りの例

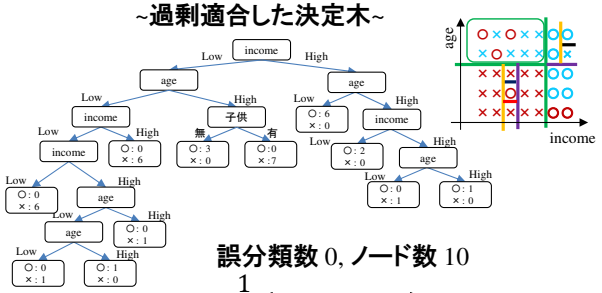
悲観的枝刈り

– 誤分類数とノード数による評価指標を小さくするよう枝刈り

メリット：簡単・単純，デメリット：理論的裏付けがない

– 評価指標：$EI = \frac{1}{N}$(誤分類数 + 0.5 × ノード数)

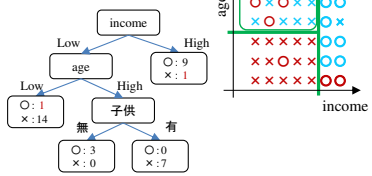
~過剰適合した決定木~



誤分類数 0, ノード数 10

$EI = \frac{1}{35}(0 + 0.5 \times 10) = 0.143$

~枝刈りした決定木~



誤分類数 2, ノード数 3

$EI = \frac{1}{35}(2 + 0.5 \times 3) = 0.1$

決定木使用の注意事項

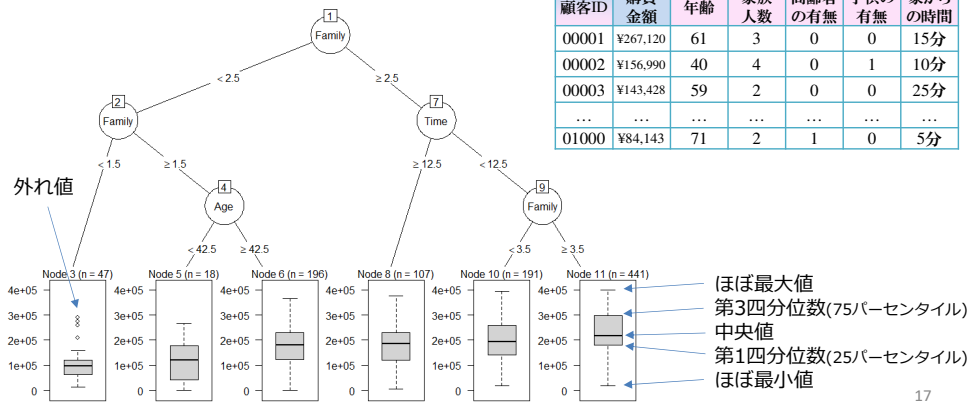
- データの変動に対して，構築される木構造やルールが敏感に変動する
- 分割基準，枝刈り法によって異なる結果が出力される
- ランダムフォレスト法やXGBoost法は決定木の予測精度を向上させる主要な手法。ただし，回帰・判別結果の説明は決定木の方が分かりやすい
- コストが最小の木構造を見つけるには全探索が必要
  – コスト最小化の計算量は組み合わせ爆発(NP困難)。前述の分割法は近似解法であり，良さそうな決定木の作成方法

#メモ：もしも、最適な木構造生成の効率的(多項式時間で解ける)なアルゴリズムを見つけたらアメリカのクレイ数学研究所から100万ドル貰えます。数学の最難関未解決問題(N≠NP予想)で扱われるクラスの一つです。

演習問題

線形回帰モデルの回で扱ったスーパーマーケットデータに対して，1年間の購買金額を目的変数として決定木分析（回帰木）を行った。この結果を解釈しなさい

dtresult = rpart(Sales~.,data=data,cp=0.008)



外れ値

顧客ID	購買金額	年齢	家族人数	高齢者の有無	子供の有無	家からの時間
00001	¥267,120	61	3	0	0	15分
00002	¥156,990	40	4	0	1	10分
00003	¥143,428	59	2	0	0	25分
…	…	…	…	…	…	…
01000	¥84,143	71	2	1	0	5分

ほぼ最大値
第3四分位数(75パーセンタイル)
中央値
第1四分位数(25パーセンタイル)
ほぼ最小値