決定木

Decision Tree

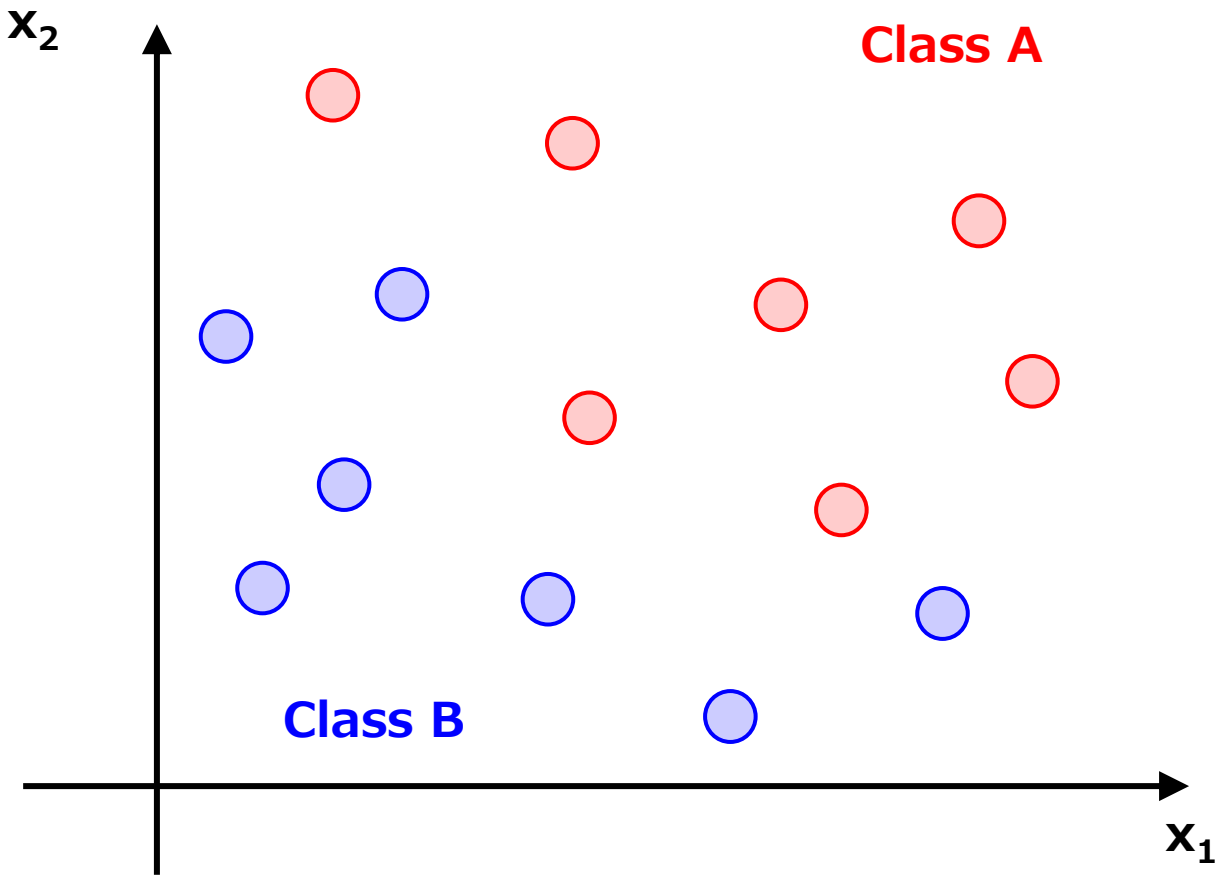
大阪府立大学 工学研究科
清水　悠生

決定木とは

- ✓ 木構造を用いてクラス分類や回帰を行う機械学習手法
- ✓ クラス分類⇒分類木，回帰⇒回帰木
- ✓ Yes/Noで回答可能な質問で構成される
  階層的な木構造を有するため，視覚的にもわかりやすい
- ✓ 本記事では，CART(Classifciation and Regression Trees)
  と呼ばれるアルゴリズムについて解説

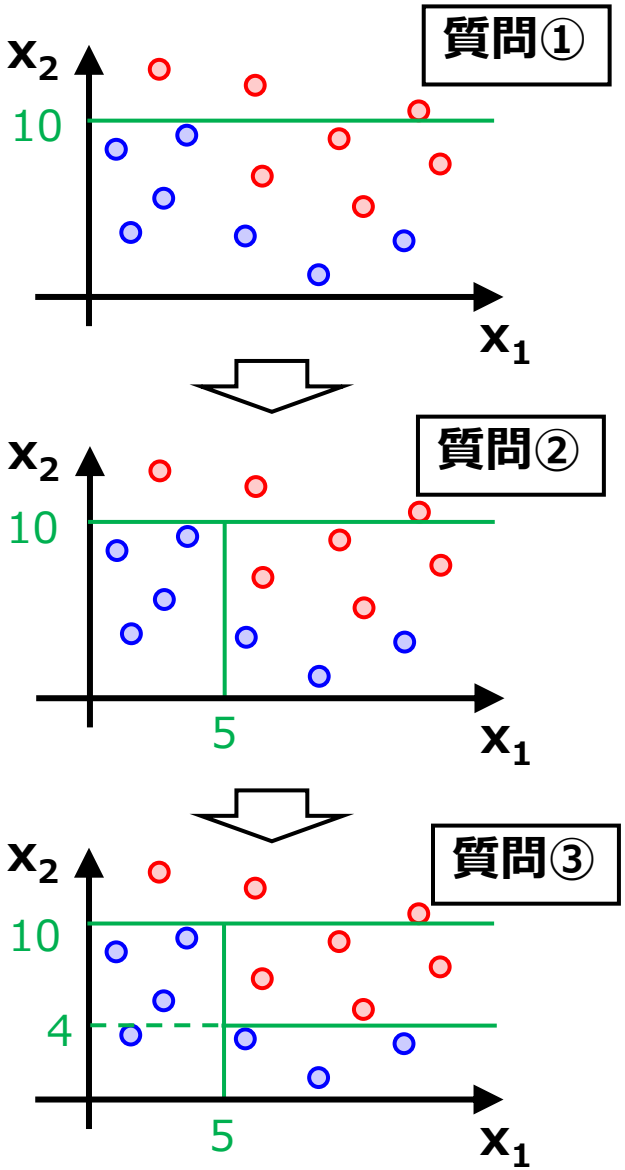
分類木の問題

✓ 2つのクラスの分類境界を求める問題を考える

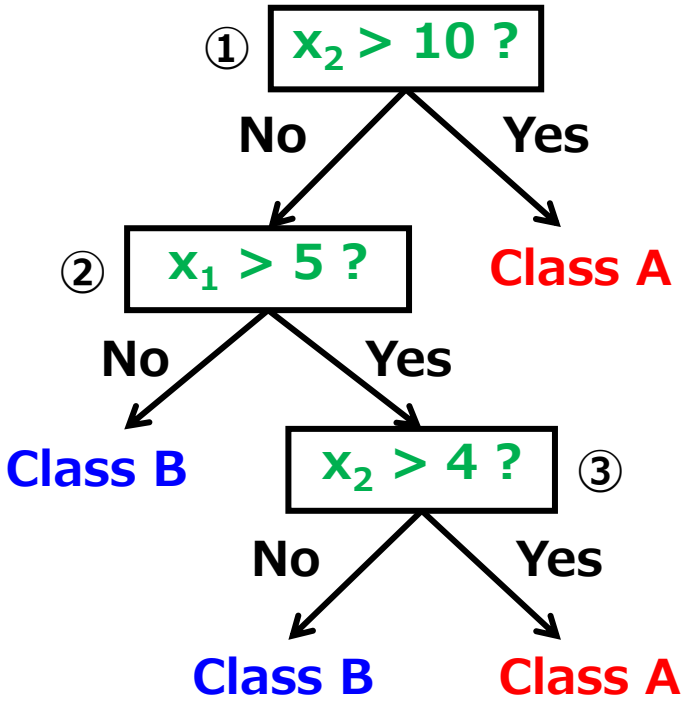


分類木の構築イメージ

✓ 2つのクラスを分類可能なYes/No形式の質問を構築する

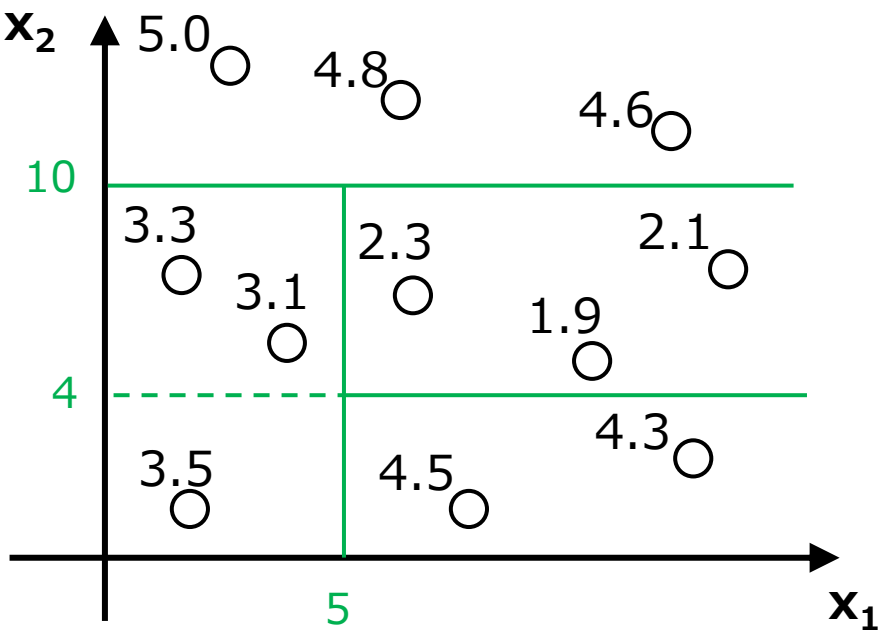


構築する分類木



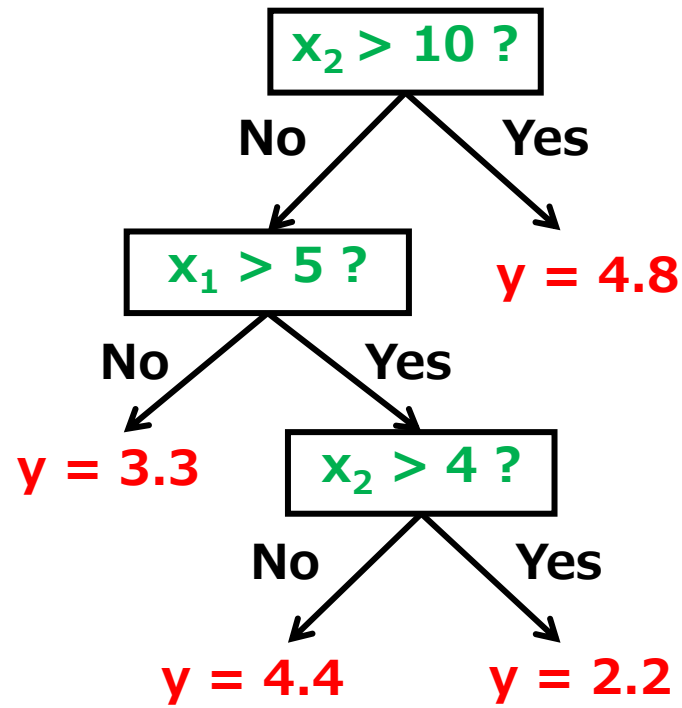
回帰木のイメージ

- ✓ 回帰木も同様にYes/No形式の質問を構築する
- ✓ 回帰木の出力は，例えば領域内の平均値



$y^{(i)}$ ○ **$y^{(i)}$: 各データの出力値**

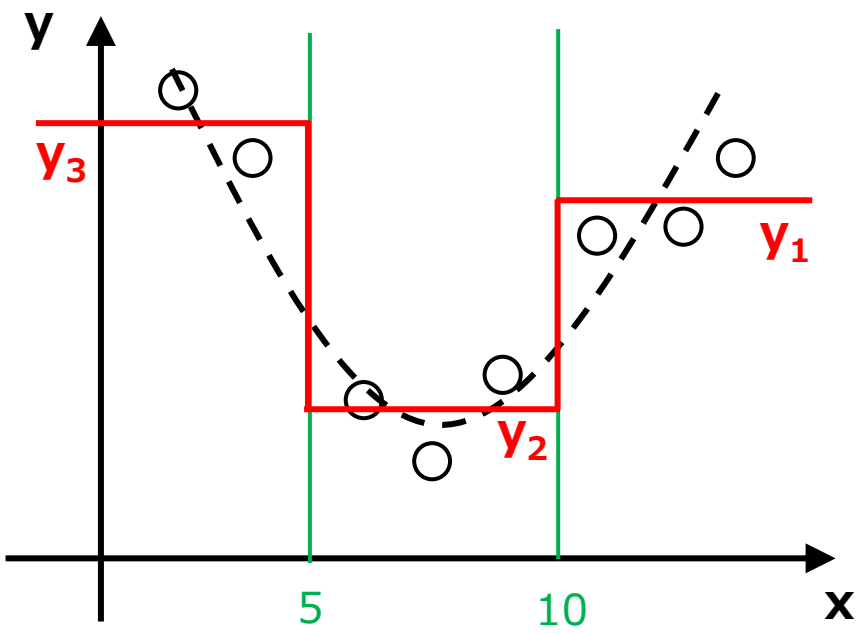
構築する回帰木



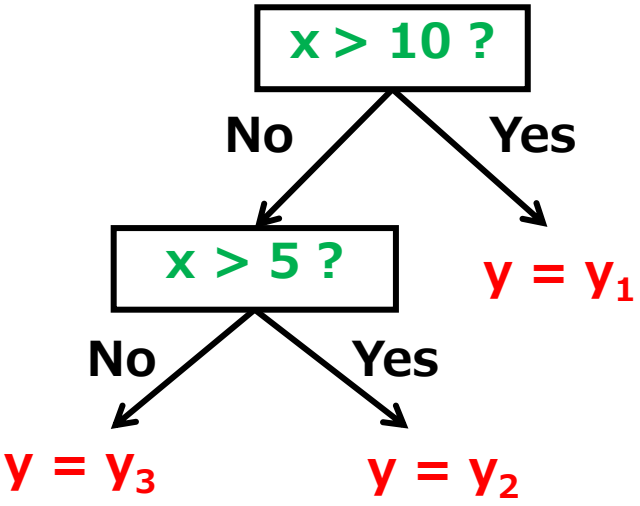
※各領域の平均値を出力とする

回帰木のイメージ（1次元）

- ✓ 1次元のほうが直感的に理解しやすい
- ✓ 回帰木によって回帰曲線（曲面）を構築する
- ✓ 回帰曲線（曲面）はステップ状になる



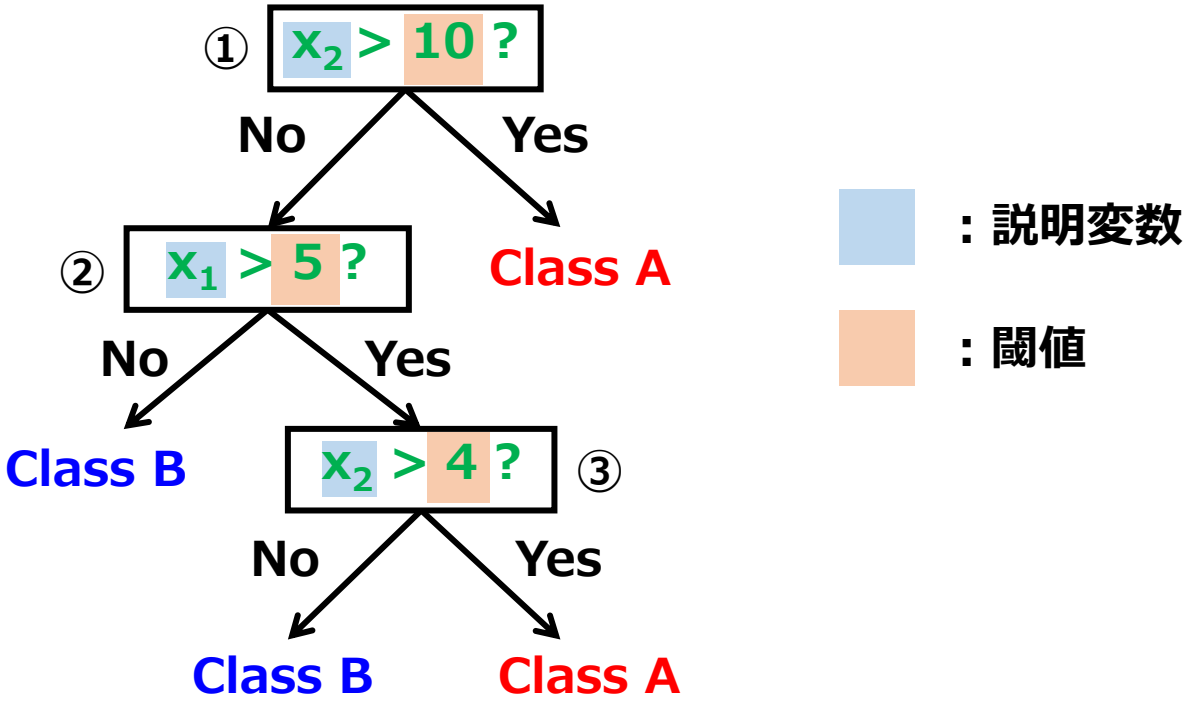
構築する回帰木



※各領域の平均値を出力とする

説明変数と閾値をどうやって選択するか？

✓ 説明変数と閾値の全ての組み合わせにおいて
損失関数を計算し，損失が最小となる組み合わせを選択



分類木で扱う損失関数

✓ 分類木で扱う損失関数には以下のようなものが存在

ジニ係数

$$H(Q_m) = \sum_k p_{mk}(1 - p_{mk})$$

交差エントロピー

$$H(Q_m) = -\sum_k p_{mk}\log(p_{mk})$$

Misclassification

$$H(Q_m) = 1 - \max_k(p_{mk})$$

$H(Q_m)$: m番目のノードの損失関数

k: 分類するクラス数
（2クラス分類 ⇒ k=1,2）

p_{mk}: m番目のノードにおける
クラス k のサンプルの割合

回帰木で扱う損失関数

✓ 回帰木で扱う損失関数には以下のようなものが存在

平均二乗誤差

$$H(Q_m) = \frac{1}{N_m}\sum_j \left(y^{(j)} - \bar{y}_m\right)$$

N_m: m番目のノードのサンプル数

$y^{(j)}$: m番目のノードの
j番目の目的変数の値

$\bar{y}_m$: m番目のノードの
全ての目的変数の平均値

Half Poisson Deviance

$$H(Q_m) = \frac{1}{N_m}\sum_j \left(y^{(j)} \log\frac{y^{(j)}}{\bar{y}_m} - y^{(j)} + \bar{y}_m\right)$$

平均絶対誤差

$$H(Q_m) = \frac{1}{N_m}\sum_j \left|y^{(j)} - \mathrm{median}\left(y^{(j)}\right)\right|$$

決定木のメリット・デメリット

✓ メリット

- 容易に可視化可能で，解釈がしやすい
- スケールに依存せず，標準化や正規化の必要がない

✓ デメリット

- 過学習しやすく，汎化性能が低い傾向にある
- 回帰木において，外挿が不可能