決定木
Decision Tree
DT

明治大学 理工学部 応用化学科
データ化学工学研究室　金子 弘昌

決定木 (Decision Tree, DT) とは？

✓回帰分析にもクラス分類にも使える

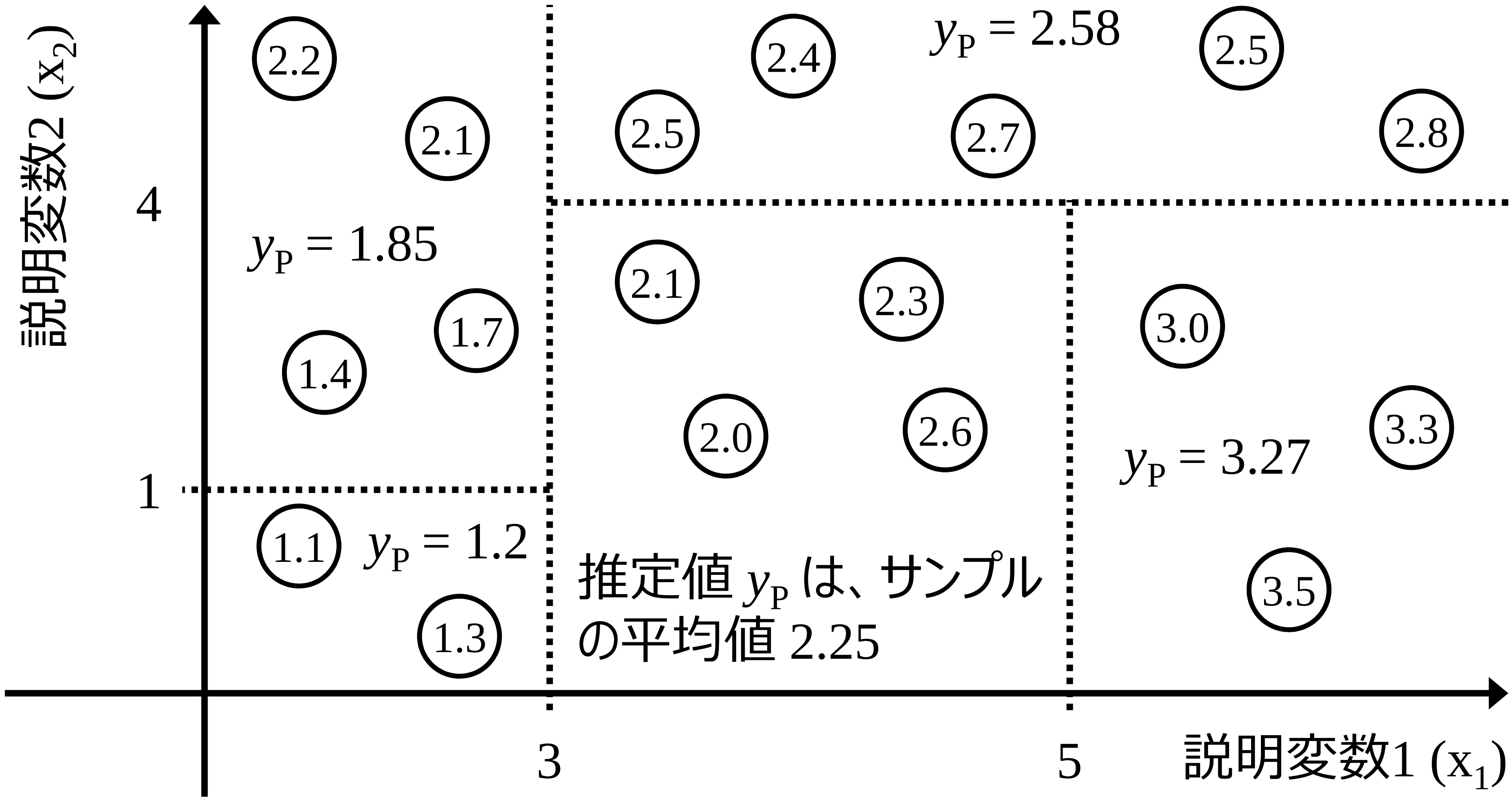
✓回帰モデル・クラス分類モデルが、木のような構造で与えられるため、モデルを直感的に理解しやすい

✓理解しやすい反面、モデルの精度は他の手法と比べて低くなってしまうことが多い

✓今回説明するのは CART (Classification and Regression Tree)

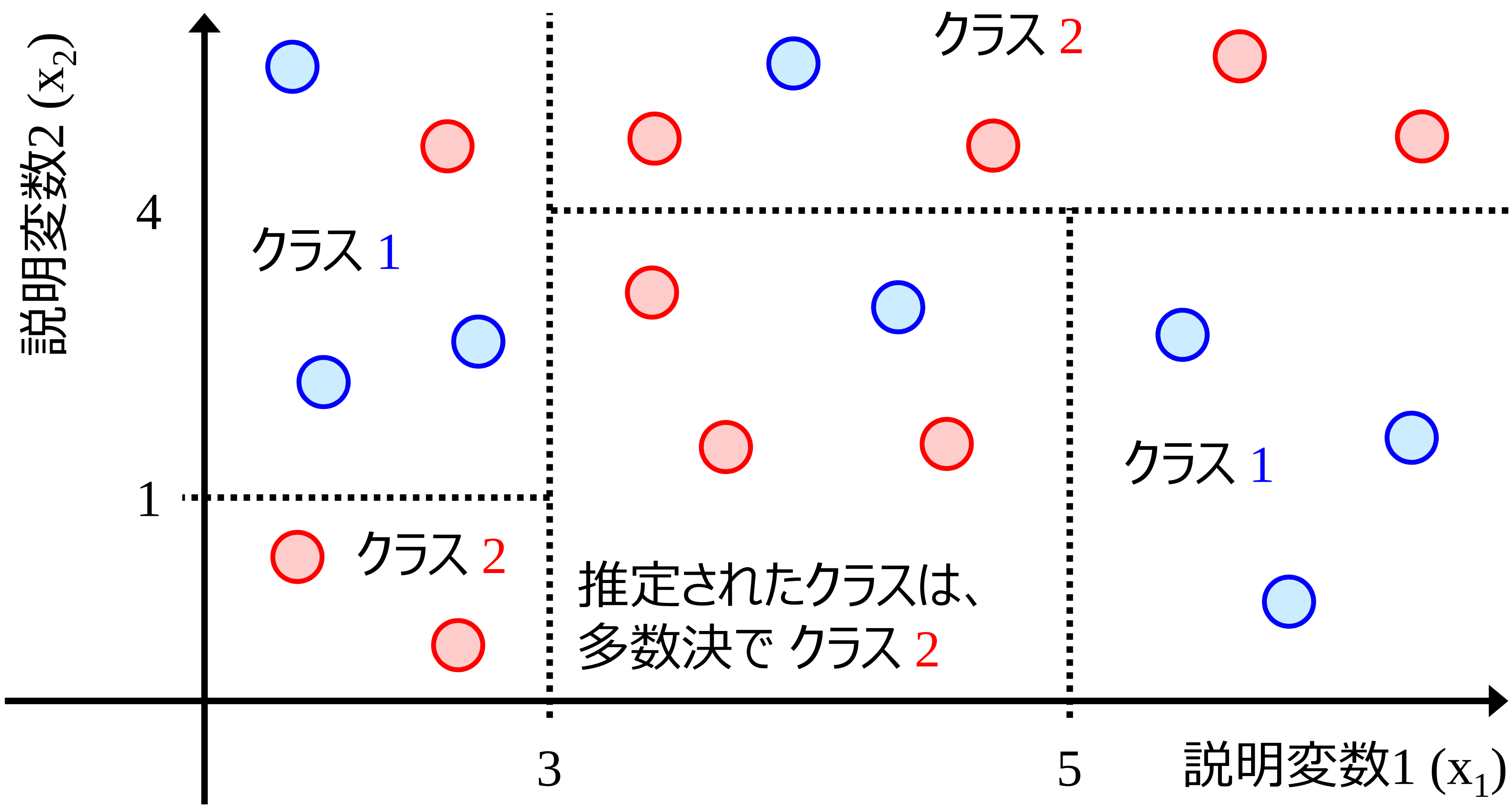
決定木でできることのイメージ (回帰分析)

n ・・・ 目的変数 y の値が n のサンプル

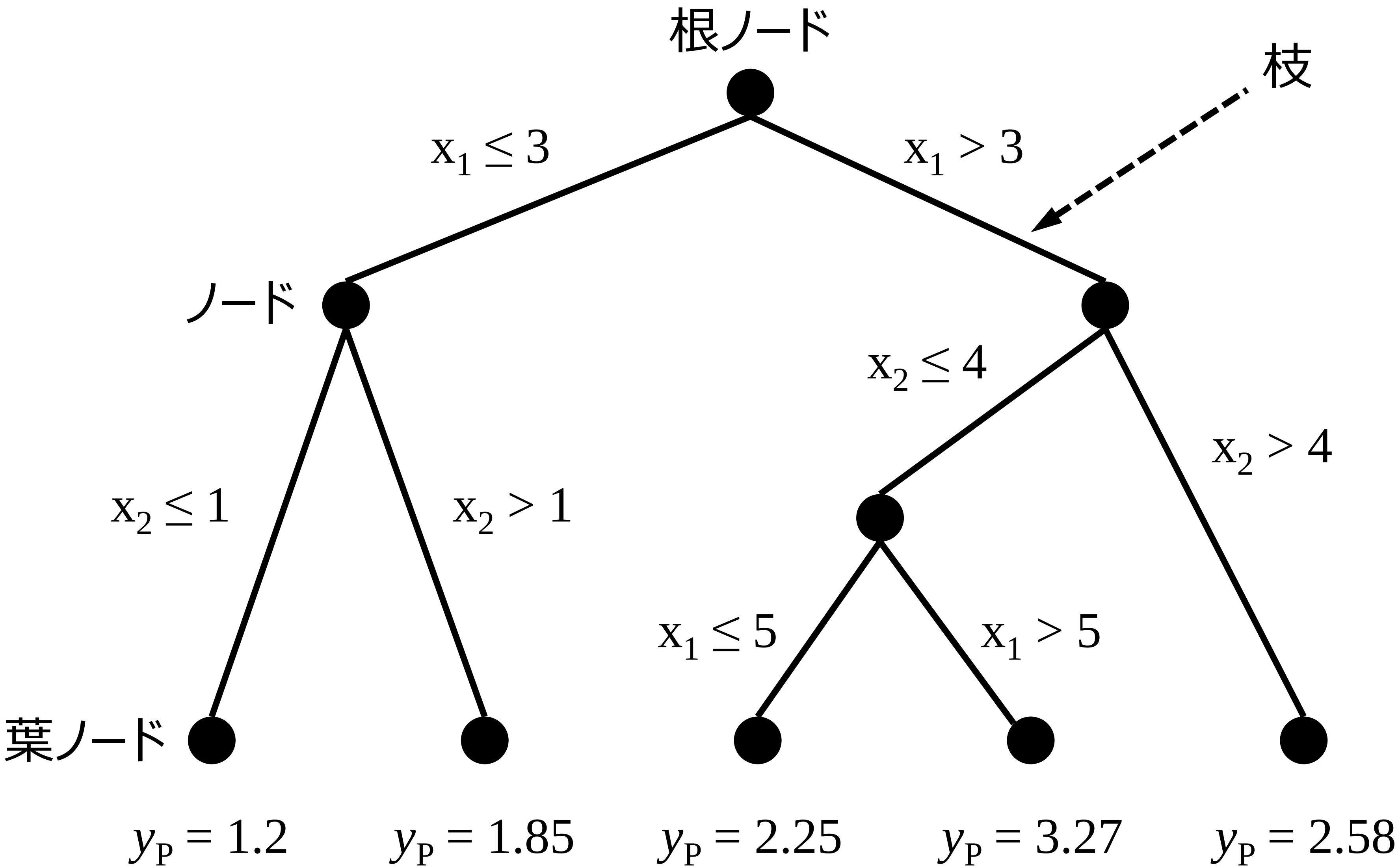


決定木のでできることのイメージ (クラス分類)

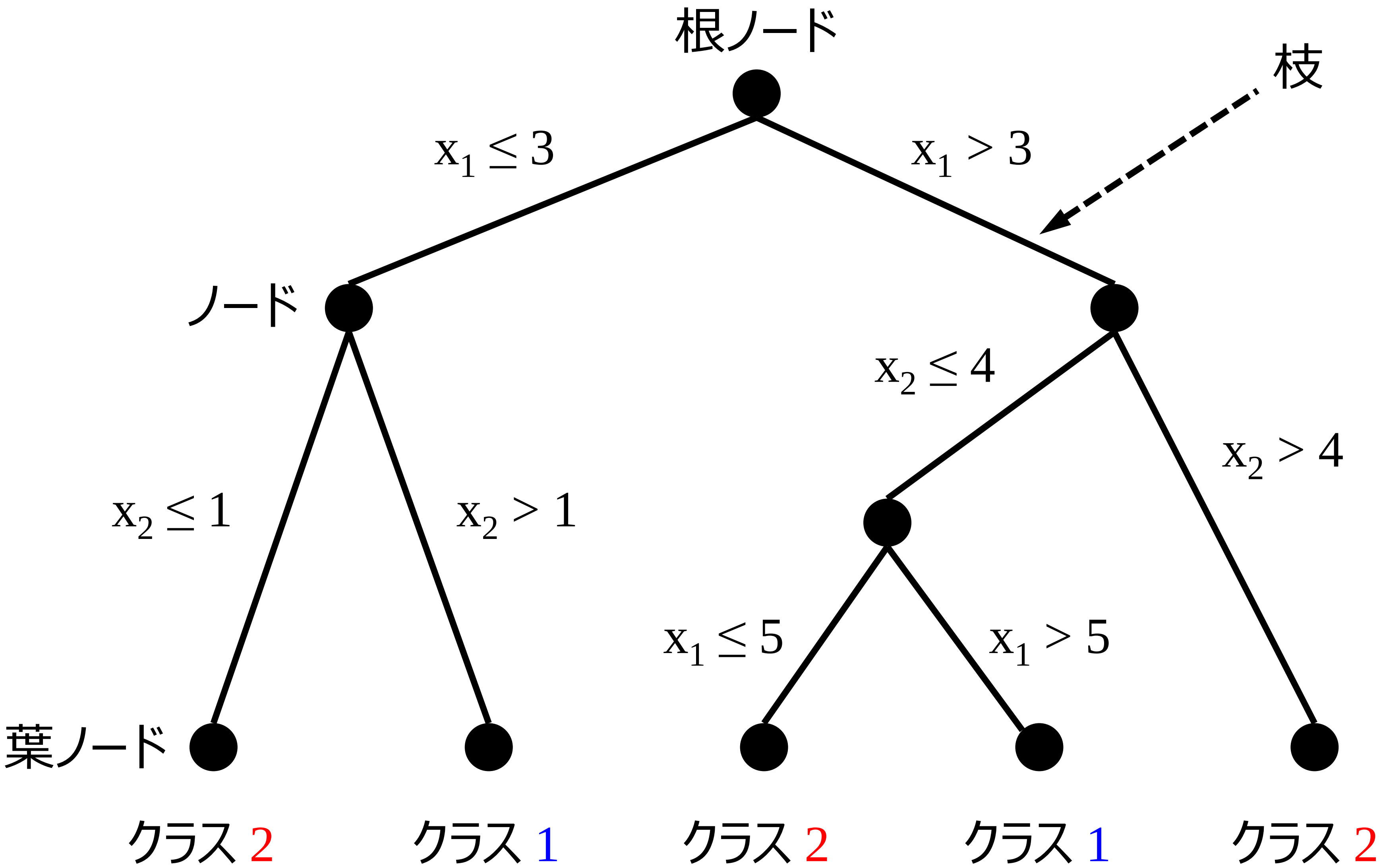
◯ ・・・ クラスが 1 のサンプル　　◯ ・・・ クラスが 2 のサンプル



決定木モデルの木構造 (回帰分析)



決定木モデルの木構造 (クラス分類)



決定木のアルゴリズム

✓どのように木を作るか？

- 根ノードから、２つずつ葉ノードを追加していき、木を成長させる

✓どのように２つの葉ノードを追加するか？

✓つまり、どのように説明変数を選んで、どのようにしきい値を選ぶか？

- 説明変数としきい値とのすべての組み合わせにおいて、
  評価関数 E の値を計算し、それが最も小さい組み合わせにする

回帰分析における評価関数 E

✓目的変数の誤差の二乗和

- それぞれの葉ノードにおける目的変数の推定値は、同じ葉ノードにあるサンプルの平均値で与えられる

$$E = \sum_{i=1}^{n} E_i$$

$$E_i = \sum_{j=1}^{m_i} \left(y_i^{(j)} - y_{\mathrm{P}i} \right)^2$$

$$y_{\mathrm{P}i} = \frac{1}{m_i} \sum_{j=1}^{m_i} y_i^{(j)}$$

n : 葉ノードの数
E_i : 葉ノード i の評価関数
m_i : 葉ノード i におけるサンプル数
$y_j^{(i)}$: 葉ノード i における、j 番目のサンプルの目的変数の値
$y_{\mathrm{P}i}$: 葉ノード i における目的変数の推定値

クラス分類における評価関数 E

✓交差エントロピー誤差関数

$$E_i = -\sum_{k=1}^{K} p_{ik} \ln p_{ik}$$

K : クラスの数
p_{ik} : 葉ノード i における、クラス k のサンプルの割合

✓ジニ係数

$$E_i = \sum_{k=1}^{K} p_{ik} \left(1 - p_{ik}\right)$$

いずれも、 $E = \sum_{i=1}^{n} E_i$

(ジニ係数のほうがよく使われるかな・・・)

いつ木の成長を止めるか？

✓クロスバリデーションの誤差が最小になるように深さを決める

✓１つの葉ノードにおける最小サンプル数を決め (３とか)、
とりあえずすべて木を生成させる

✓葉ノードを２つずつ枝刈りしていく

- 下の基準 C が大きくなったら枝刈りストップ

$$C = E + \lambda n$$

E : 評価関数
n : 葉ノードの数
λ : 木の精度と複雑度との間のトレードオフを決める重み

- λ はクロスバリデーションで決める