

Machine Learning

# 6장 Classification and Regression Trees (CART)

---

고려대학교 통계학과  
박유성



# Contents

**01** Introduction

**02** Regression Tree

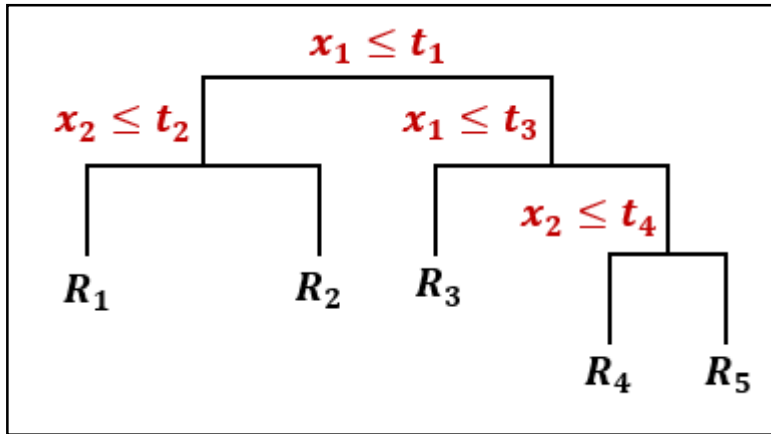
**03** Classification Tree

# 01 Introduction

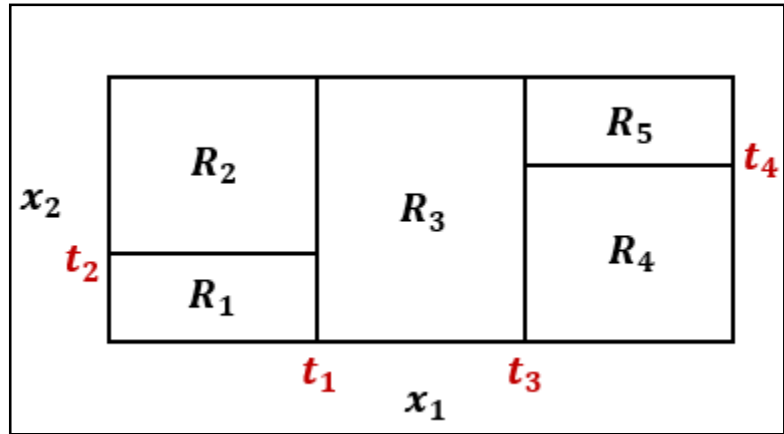
- CART의 다른 이름: “Decision Tree Learning”
- 목적: Target value  $y$ 를 예측 (분류).  $y$ 는 연속변수 or 이산변수 (class)
- 절차
  - [Step 1] 특성변수  $x = (x_1, x_2, \dots, x_d)^T$ 의 샘플을  $d$ 차원 상의  $M$ 개 직사각형 cell들로 partition
    - ▶ Cell (Region) 구성: Training data를 이용
    - ▶ Generalization error 측정: Test data를 이용
  - [Step 2] 각 cell (region)을 대표하는  $y$ 값 할당 (이 값을 각 cell에서의 추정치로 사용)
    - ▶ If  $y$ 가 연속변수, 각 cell에서의 평균값을 할당 → “Regression Tree”
    - ▶ If  $y$ 가 이산변수, 각 cell에서 비중이 가장 큰 class를 할당 → “Classification Tree”

# Graphical Example

- (Example) 특성변수 두 개( $x_1$  and  $x_2$ ,  $d=2$ ), cell 개수 5개인 경우



<Decision Tree by  $x_1$  and  $x_2$ >



<Decision Tree에 의한 직사각형 cells>

- (기호)  $t_1, t_2, t_3, t_4$  : 이항분할 (binary partition) 값.  $\rightarrow$  “node”
- 핵심 결정사항
  - 분할 시 특성변수의 순서를 어떻게 정할 것인지?
  - 각 이항분할 값 (node)은 어떻게 결정할 것인지?

## 02 Regression Tree

### ■ Algorithm

- [Step 1] 각 특성변수  $x_j$  ( $j=1, \dots, d$ )의 관측치를 오름차순 정렬: (예)  $\{-5, 1, -3, 3\} \rightarrow \{-5, -3, 1, 3\}$
- [Step 2] 각 특성변수  $x_j$  ( $j=1, \dots, d$ )에 대하여, 정렬된 관측치를 partition (즉, node  $t_k$ 를 구함)
  - ▶  $SSE(x_j) = \sum_{i \in l} (y_i - \bar{y}_l)^2 + \sum_{i \in r} (y_i - \bar{y}_r)^2$ 를 최소화하는  $x_j$ 의 관측치를  $x_j$ 의 노드로 선택 ( $j=1, \dots, d$ )  
(※  $l$ 은 partition의 왼쪽 cell,  $r$ 은 partition의 오른쪽 cell을 의미)
    - ➔ 즉, Regression Tree의 손실함수는 SSE (Sum of Squared Error)
  - ▶ (예)  $x_1$ 의 정렬된 관측치가  $\{-1, 0, 2, 3\}$ 이면 4개의  $SSE(x_1)$ 이 계산되고, 만일 이 중  $x_1=0$ 이 최소  $SSE(x_1)$ 를 가지고 있다면  $x_1$ 의 node는 0이 됨.
- [Step 3] [Step 2]에서 구한 최소  $SSE(x_j)$ 의 값이 가장 작은 특성변수 ( $x^*$ ) 기준으로 Split 수행
  - ▶ 2개의 region이 생성 (left, right)
- [Step 4] [Step 3]에서 구한 2개의 region 각각에 대하여, [Step 1] – [Step 3]을 반복 수행
  - ▶ 이 결과 얻게 되는 region들에 대해서도 [Step 1] – [Step 3]을 반복

# 예측치 계산 및 과적합 방지

## ■ 예측치 계산

- (기호)  $R_1, \dots, R_M$ : [Step 1] - [Step 4] 결과 최종적으로 만들어진 regions (cells)

- 새로운 관측치  $x_0$ 에 대한 regression tree의 예측치:  $\hat{y} = \sum_{m=1}^M \bar{y}_m I(x_0 \in R_m)$

( $I(x_0 \in R_m) = 1$  if  $x_0 \in R_m$ , 0 otherwise.  $\bar{y}_m = \frac{1}{N_m} \sum_{x_i \in R_m} y_i$ , where  $N_m = \text{size of } \{x_i \in R_m\}$ )

## ■ 과대적합 방지: 가지치기 (Pruning)

- 각 cell에 관측치가 오직 한 개 남을 때 까지 [Step 4] 반복 가능 → 과적합의 문제 발생

- 해결: Split의 깊이를 제한. → “가지치기 (Pruning)”

- (기호)  $T_M$ : 충분히 큰 Tree. (예) 각 region에 5개 이하의 관측치가 남을 때 까지만 키운 Tree

- 아래  $C_\alpha(T)$ 를 최소로 하는 tree  $T$ 를 구함 ( $|T|$ :  $T_M$  보다 작은 Tree  $T$ 에서의 총 cell의 개수)

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha |T|, \text{ where } Q_m(T) = \frac{1}{N_m} \sum_{x_i \in R_m} (y_i - \bar{y}_m)^2.$$

-  $\alpha$ 의 결정: by Cross-validation (제9장)

# 03 Classification Tree

## 불순도 (Impurity) 측도

- (기호)  $\hat{P}_{mk}$ : region  $R_m$ 에서 class  $k$ 가 차지하는 비율 ( $\hat{P}_{mk(m)}$ : region  $R_m$ 에서  $\hat{P}_{mk}$ 의 최대값)

Gini Index	$I_G(R_m) = \sum_{m=1}^K \hat{P}_{mk}(1 - \hat{P}_{mk}) = 1 - \sum_{m=1}^K \hat{P}_{mk}^2$	(K: class의 총 개수)
Cross Entropy	$I_C(R_m) = - \sum_{m=1}^K \hat{P}_{mk} \log_2(\hat{P}_{mk})$	
Misclassification Error	$I_E(R_m) = 1 - P_{mk(m)}$	

## Information Gain

- Classification Tree에서 분할 (split) 변수 ( $x_j$ ) 및 node 선택의 기준

- 상위 cell  $R$ 에서 두 개의 영역  $R_l$  (왼쪽)과  $R_r$  (오른쪽)로 나누어질 때,  $N, N_{R_l}, N_{R_r}$ 을 각각  $R, R_l, R_r$ 에 포함된 관측치의 개수라고 하면 ( $N = N_{R_l} + N_{R_r}$ ),

$$IG(R, x_j) = I(R) - \frac{N_{R_l}}{N} I(R_l) - \frac{N_{R_r}}{N} I(R_r).$$

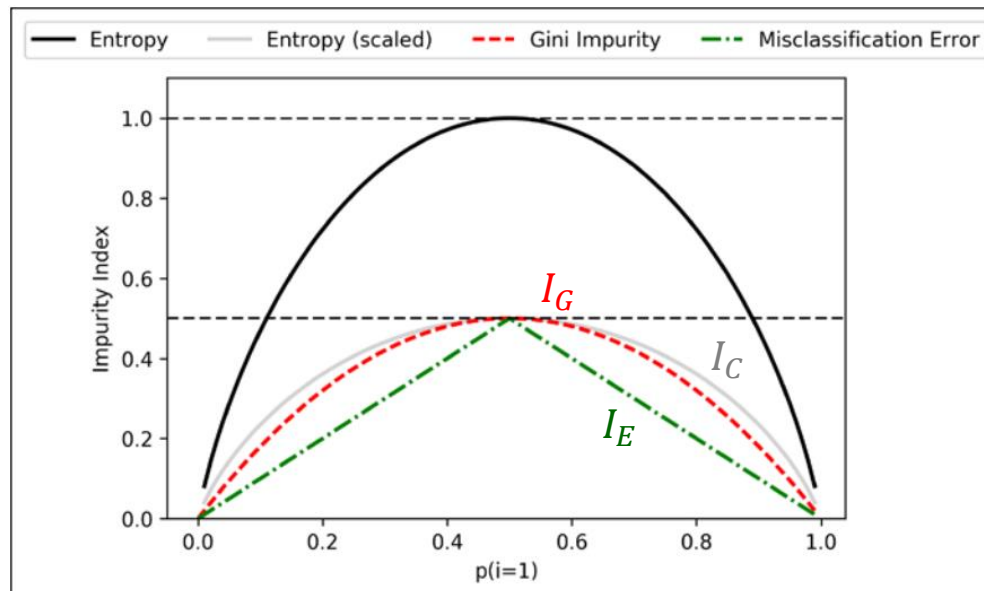
-  $IG(R, x_j)$ 가 최대가 되도록 분할변수 ( $x_j$ )와 node를 선택

# 불순도 (Impurity)의 성질

- (예)  $y$ 가 두 개의 class  $\{1, 2\}$ 를 갖는 경우

Gini Index	$I_G(R_m) = 1 - (\hat{P}_{m1}^2 + \hat{P}_{m2}^2)$	$(\hat{P}_{m1} + \hat{P}_{m2} = 1)$
Cross Entropy	$I_C(R_m) = -(\hat{P}_{m1} \log_2 \hat{P}_{m1} + \hat{P}_{m2} \log_2 \hat{P}_{m2})$	
Misclassification Error	$I_E(R_m) = 1 - \max(\hat{P}_{m1}, \hat{P}_{m2})$	

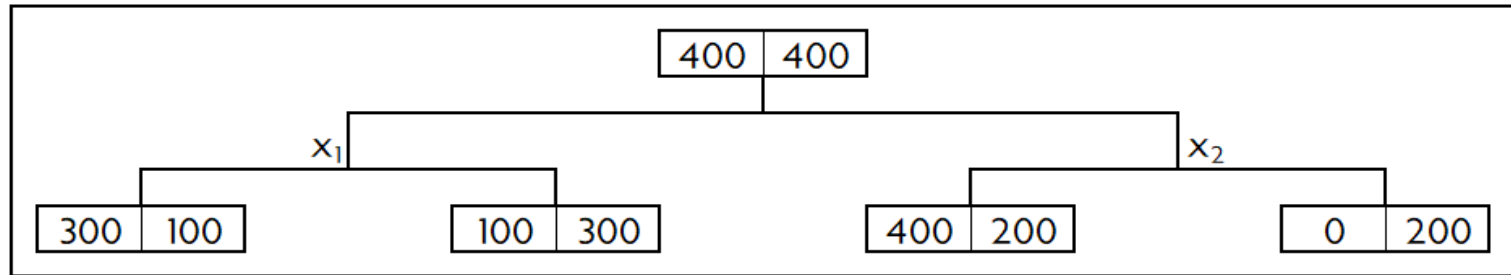
- 세 불순도 모두  $\hat{P}_{m1}=0.5$  ( $\Leftrightarrow$  분류 의미 X)일 때 최대이고,  $\hat{P}_{m1}$ 이 0 or 1로 접근할 때 점점 낮아짐.





# Example

- 부모 cell로 부터  $x_1$ 과  $x_2$ 에 의해 분류된 예



- (Gini Index 사용)  $I_G(R) = 1 - \left[ \left( \frac{4}{8} \right)^2 + \left( \frac{4}{8} \right)^2 \right] = 0.5$

▶  $x_1$  기준:  $I_G(R_l) = 1 - \left[ \left( \frac{3}{4} \right)^2 + \left( \frac{1}{4} \right)^2 \right] = 0.375$ ,  $I_G(R_r) = 1 - \left[ \left( \frac{1}{4} \right)^2 + \left( \frac{3}{4} \right)^2 \right] = 0.375$ .

$$\therefore IG_G(R, x_1) = 0.5 - \frac{4}{8} 0.375 - \frac{4}{8} 0.375 = 0.125.$$

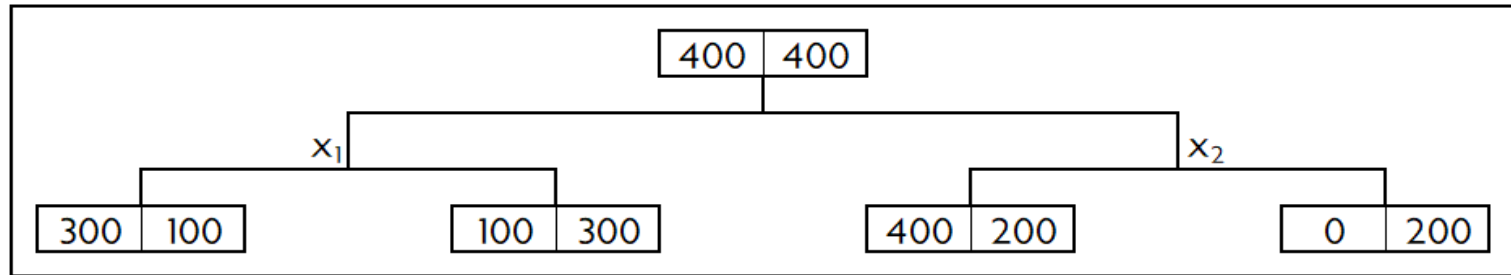
▶  $x_2$  기준:  $I_G(R_l) = 1 - \left[ \left( \frac{4}{6} \right)^2 + \left( \frac{2}{6} \right)^2 \right] = 0.444$ ,  $I_G(R_r) = 1 - \left[ \left( \frac{0}{2} \right)^2 + \left( \frac{2}{2} \right)^2 \right] = 0$ .

$$\therefore IG_G(R, x_2) = 0.5 - \frac{6}{8} 0.444 - \frac{2}{8} 0 = 0.167.$$

→  $IG_G(R, x_1) < IG_G(R, x_2)$  이므로, Gini Index에 의한 하위 cell 분류 기준은  $x_2$ .

# Example

- 부모 cell로 부터  $x_1$ 과  $x_2$ 에 의해 분류된 예 (계속)



- (Cross Entropy 사용)  $I_C(R) = -\frac{4}{8}\log_2\left(\frac{4}{8}\right) - \frac{4}{8}\log_2\left(\frac{4}{8}\right) = 1$

▶  $x_1$  기준:  $I_C(R_l) = -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) = 0.811$ ,  $I_C(R_r) = -\frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{3}{4}\log_2\left(\frac{3}{4}\right) = 0.811$ .

$$\therefore IG_C(R, x_1) = 1 - \frac{4}{8}0.811 - \frac{4}{8}0.811 = 0.189.$$

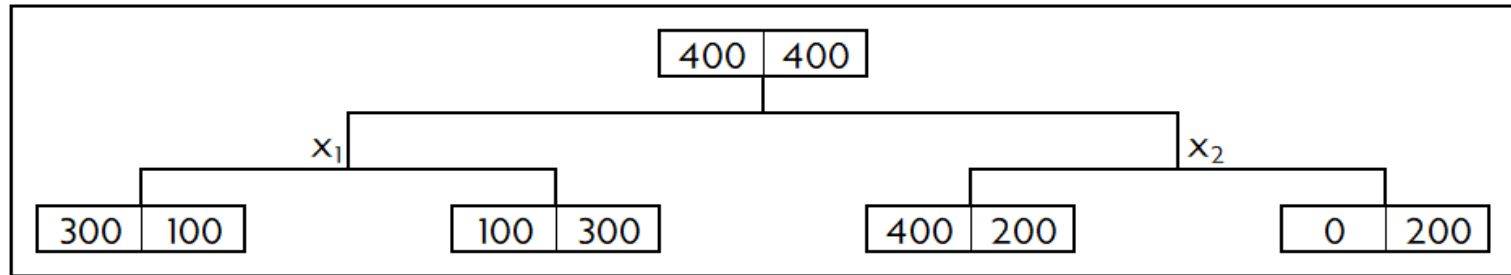
▶  $x_2$  기준:  $I_C(R_l) = -\frac{4}{6}\log_2\left(\frac{4}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right) = 0.918$ ,  $I_C(R_r) = -\frac{0}{2}\log_2\left(\frac{0}{2}\right) - \frac{2}{2}\log_2\left(\frac{2}{2}\right) = 0$ .

$$\therefore IG_C(R, x_2) = 0.5 - \frac{6}{8}0.918 - \frac{2}{8}0 = 0.311.$$

→  $IG_C(R, x_1) < IG_C(R, x_2)$  이므로, Cross Entropy에 의한 하위 cell 분류 기준은  $x_2$ .

# Example

- 부모 cell로 부터  $x_1$ 과  $x_2$ 에 의해 분류된 예 (계속)



- (Misclassification Error 사용)  $I_E(R) = 1 - \max\left[\frac{4}{8}, \frac{4}{8}\right] = 0.5$

▶  $x_1$  기준:  $I_E(R_l) = 1 - \max\left[\frac{3}{4}, \frac{1}{4}\right] = 0.25$ ,  $I_E(R_r) = 1 - \max\left[\frac{1}{4}, \frac{3}{4}\right] = 0.25$ .

$$\therefore IG_E(R, x_1) = 0.5 - \frac{4}{8}0.25 - \frac{4}{8}0.25 = 0.25.$$

▶  $x_2$  기준:  $I_E(R_l) = 1 - \max\left[\frac{4}{6}, \frac{2}{6}\right] = 0.333$ ,  $I_E(R_r) = 1 - \max\left[\frac{0}{2}, \frac{2}{2}\right] = 0$ .

$$\therefore IG_E(R, x_2) = 0.5 - \frac{6}{8}0.333 - \frac{2}{8}0 = 0.25.$$

→  $IG_E(R, x_1) = IG_E(R, x_2)$  이므로, Misclassification Error에 의한 하위 cell 분류 기준 선택 불가.

- Pruning에서는 불순도 측도로서 misclassification error를 사용함.

Q & A