분류의 기본 알고리즘

결정트리
(Decisioin Tree)

고려대학교 산업경영공학부 김성범님 인공지능공학연구소 참고

결정 트리의 정의

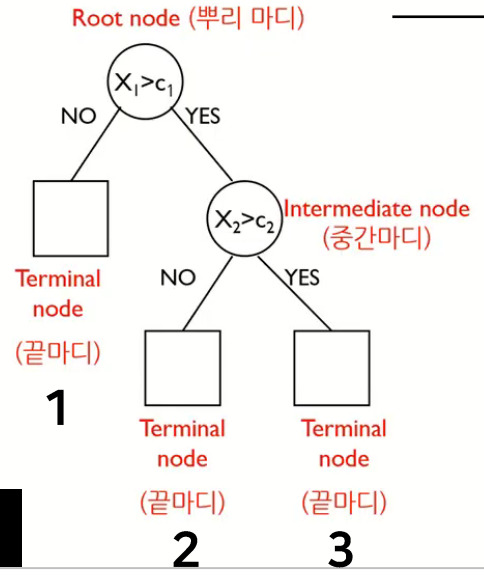
데이터에 내재되어 있는 패턴을 변수의 조합으로 나타내는 예측/분류 모델을 나무의 형태로 만드는 것.

Idea:
질문을 던져서 맞고 틀리는 것에 따라 우리가 생각하고 있는 대상을 좁혀나감
"스무고개"놀이와 비슷한 개념 -> 아키네이터

Data ⇨ Algorithm ⇨ Model (Output)

Input				Output
X_1	X_2	…	X_p	Y

- 데이터를 2개 혹은 그 이상의 부분집합으로 분할 → 데이터가 균일해지도록 분할
- 분류: 비슷한 범주를 갖고 있는 관측치끼리 모음
- 예측: 비슷한 수치를 갖고 있는 관측치끼리 모음



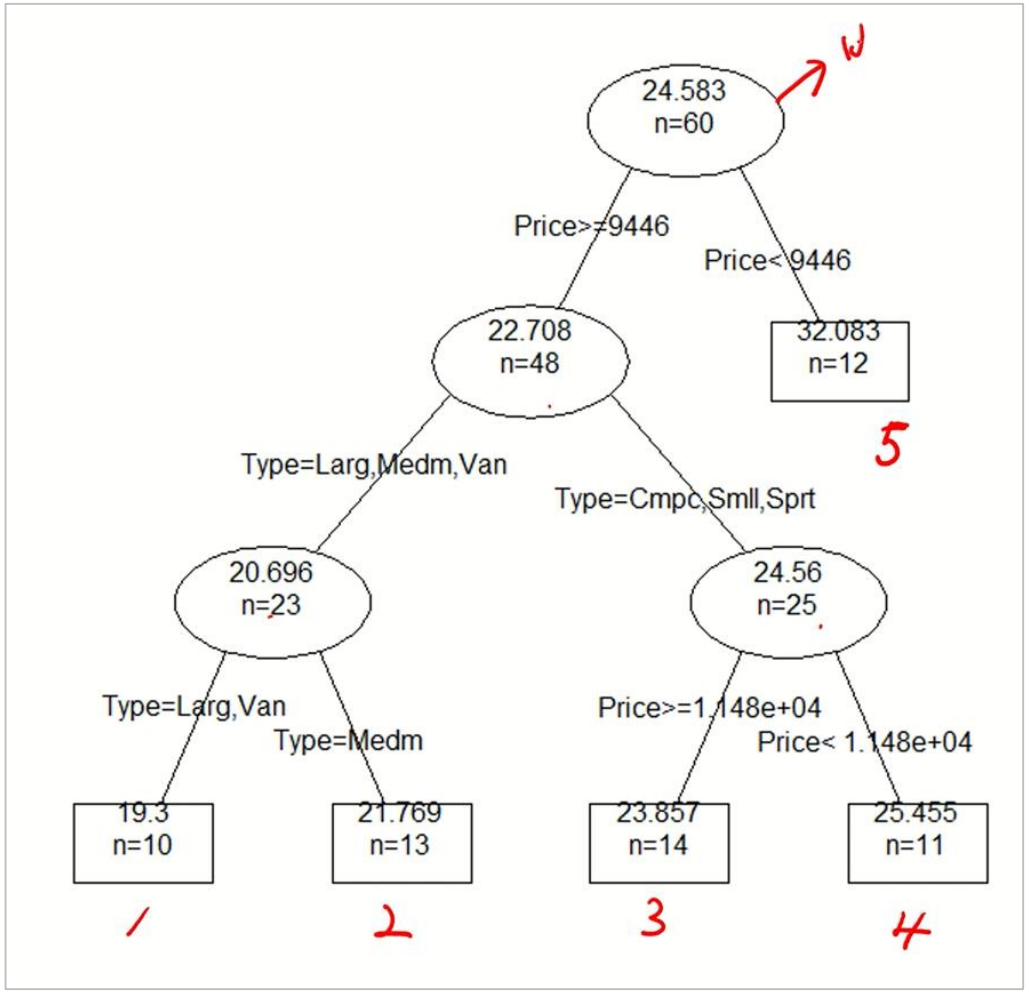
의사결정 나무 모델 개요

→ 끝마디가 3개인 결정트리 모델

데이터가 들어와야하고, 알고리즘을 통해 모델이 구축되는 과정을 거칩니다.

Root node: 뿌리 마디
Intermediate node: 중간 마디
Terminal node: 끝마디

이진분할



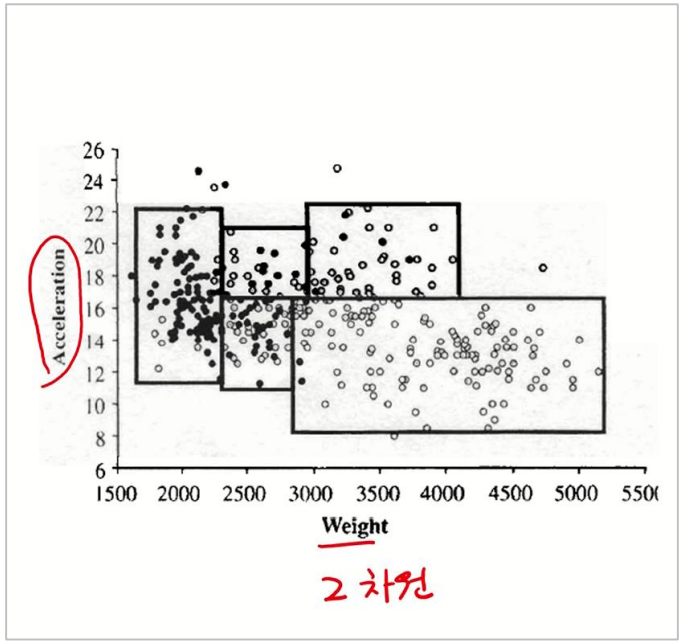
끝마디 5개
중간마디 3개
뿌리노드 1개

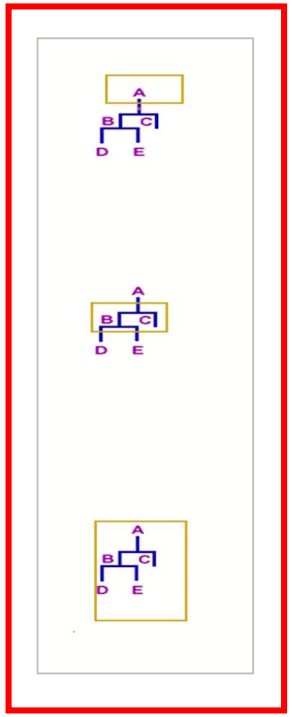
4번의 이진분할에 의해 분할이 됨.

결정 트리 (Decision Tree)

예측 모델

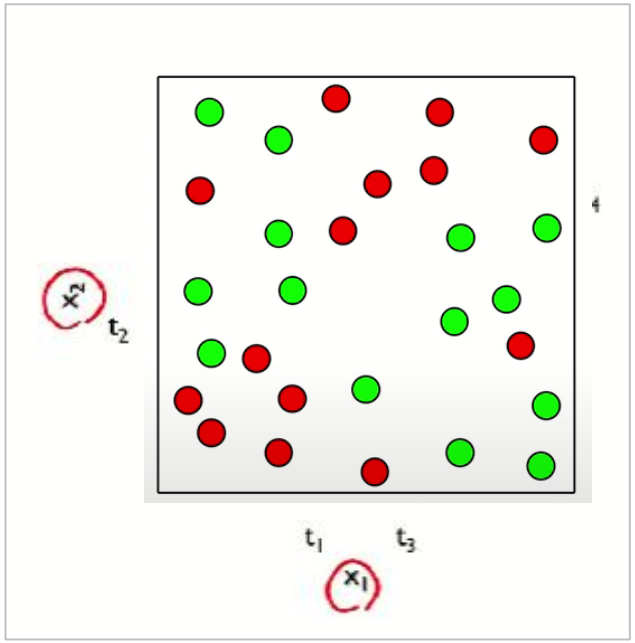
비슷한 수치를 가지고 갖고 있는 관측치들끼리 모음



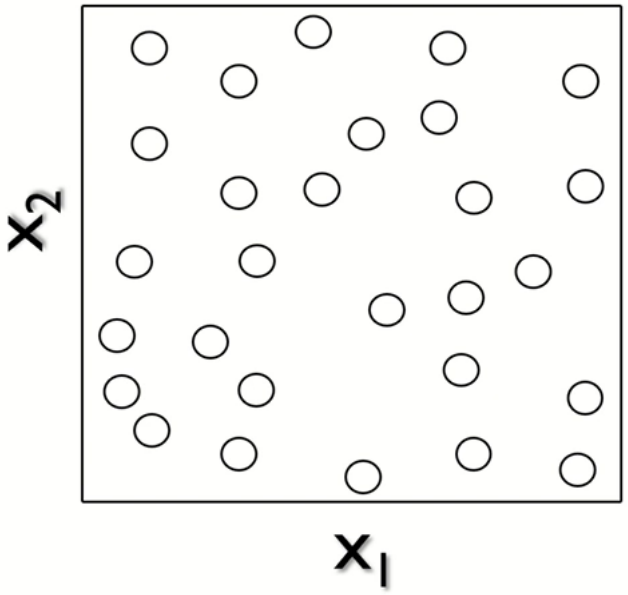


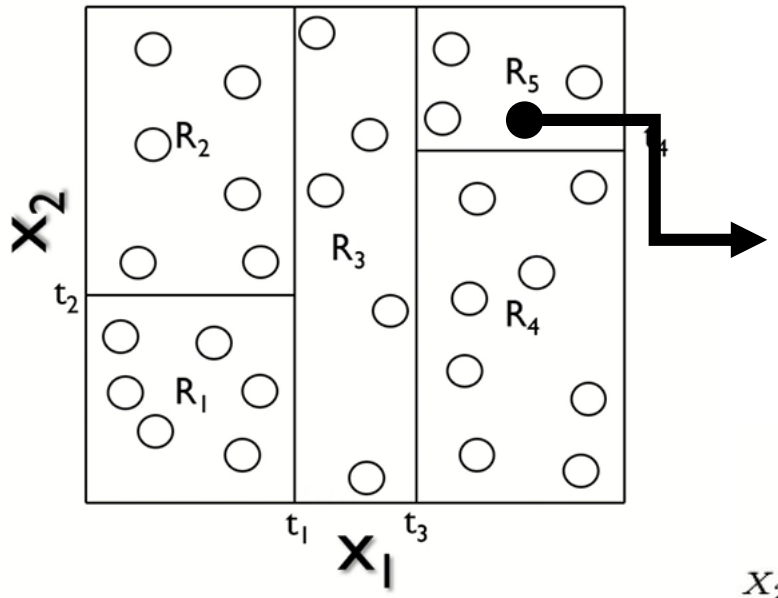
분류 모델

비슷한 범주를 가지고 갖고 있는 관측치들끼리 모음



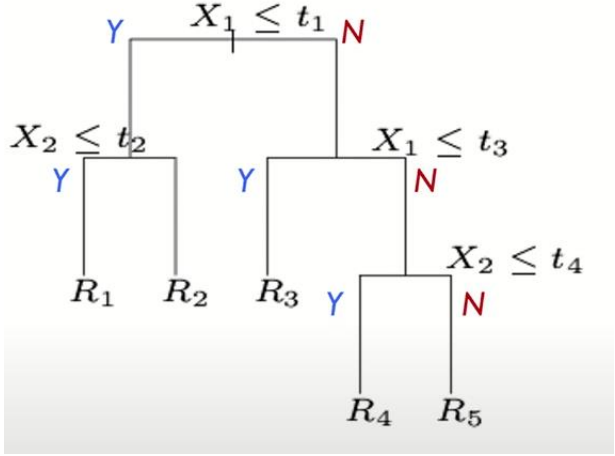
예측나무 모델





들어온 y 예측?

$$\hat{y} = \frac{1+5+3}{3} = 3$$



1. X1이 t1보다 작은지 큰지?
2. X1이 t3보다 작은지 큰지?
3. X2가 t2보다 작은지 큰지?

예측나무 모델

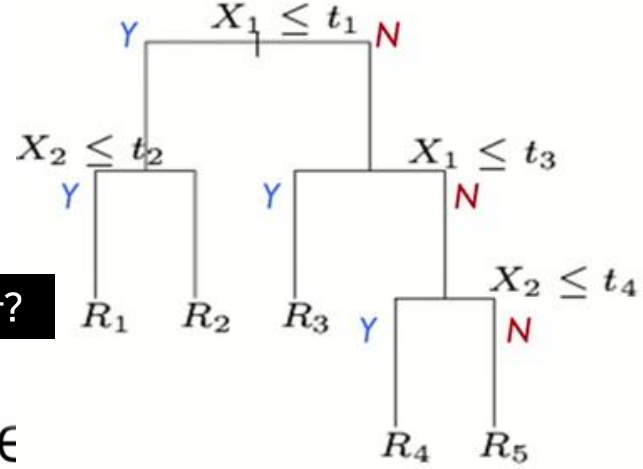
- C_m: 회귀나무모델로 부터 예측한 R_m 부분의 예측값

이진함수
0 -> False
1 -> True

$$\hat{f}(x) = \sum_{m=1}^{5} c_m I\{(x_1, x_2) \in R_m\}$$

$$= c_1 I\{(x_1, x_2) \in R_1\} + c_2 I\{(x_1, x_2) \in R_2\} + c_3 I\{(x_1, x_2) \in$$
$$+ c_4 I\{(x_1, x_2) \in R_4\} + c_5 I\{(x_1, x_2) \in R_5\}$$

우리가 고려하고 있는 x1, x2가 Rm 지역에 있는가?

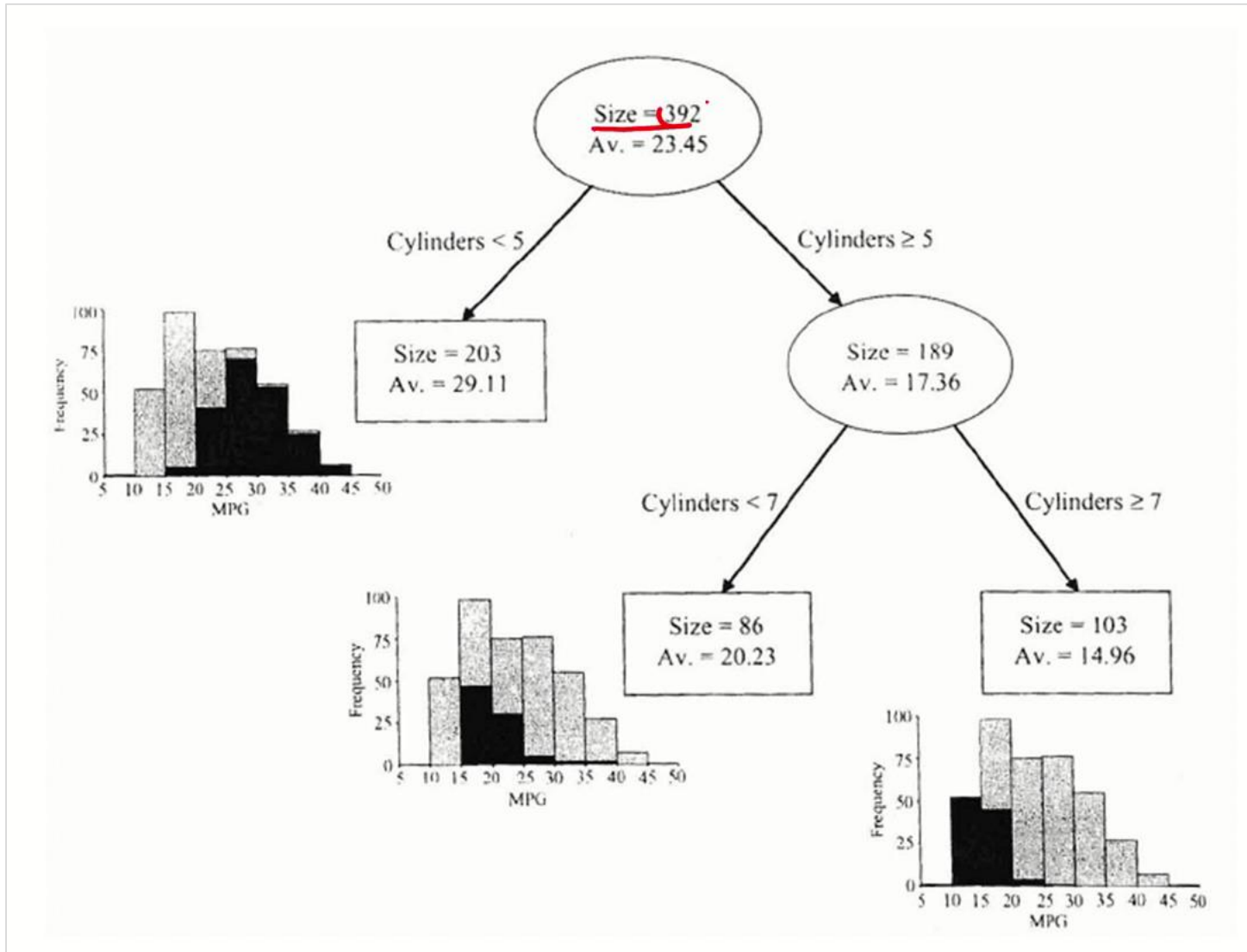


데이터가 들어와서 R3지역에 있다고 했을때, 어떻게 표현할 수 있을까?

$= c_1 \cdot 0 + c_2 \cdot 0 + c_3 \cdot 1 + c_4 \cdot 0 + c_5 \cdot 0$

$= c_3$

예측나무 모델


Size = 392
Av. = 23.45
Cylinders < 5
Cylinders ≥ 5
Size = 203
Av. = 29.11
Size = 189
Av. = 17.36
Cylinders < 7
Cylinders ≥ 7
Size = 86
Av. = 20.23
Size = 103
Av. = 14.96
Frequency
MPG
0
25
50
75
100
5 10 15 20 25 30 35 40 45 50

예측나무 모델링 프로세스

- 데이터를 M 개로 분할: $R_1, R_2, \ldots, R_M$ 끝마디가 M개

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m)$$

- 최상의 분할은 다음 비용함수(cost function)를 최소로 할 때 얻어짐

$$\min_{C_m} \sum_{i=1}^{N} (y_i - f(x_i))^2$$

실제숫자 y와 모델에서 나온 y값의 차이의 제곱의 합을 최소화 시키려면 Cm이 뭐가 되어야 할까?

$$= \min_{C_m} \sum_{i=1}^{N} (y_i - \sum_{m=1}^{M} c_m I(x \in R_m))^2$$

$$\hat{c}_m = ave(y_i | x_i \in R_m)$$

- 각 분할에 속해 있는 y값들의 평균으로 예측했을 때 오류가 최소

분할변수(j)와 분할점(s)은 어떻게 결정할까?

$$R_1(j,s) = \{x | x_j \leq s\}$$
$$R_2(j,s) = \{x | x_j > s\}$$

비용함수가 최소화 되려면 C의 값은 해당 부분집합의 최소값

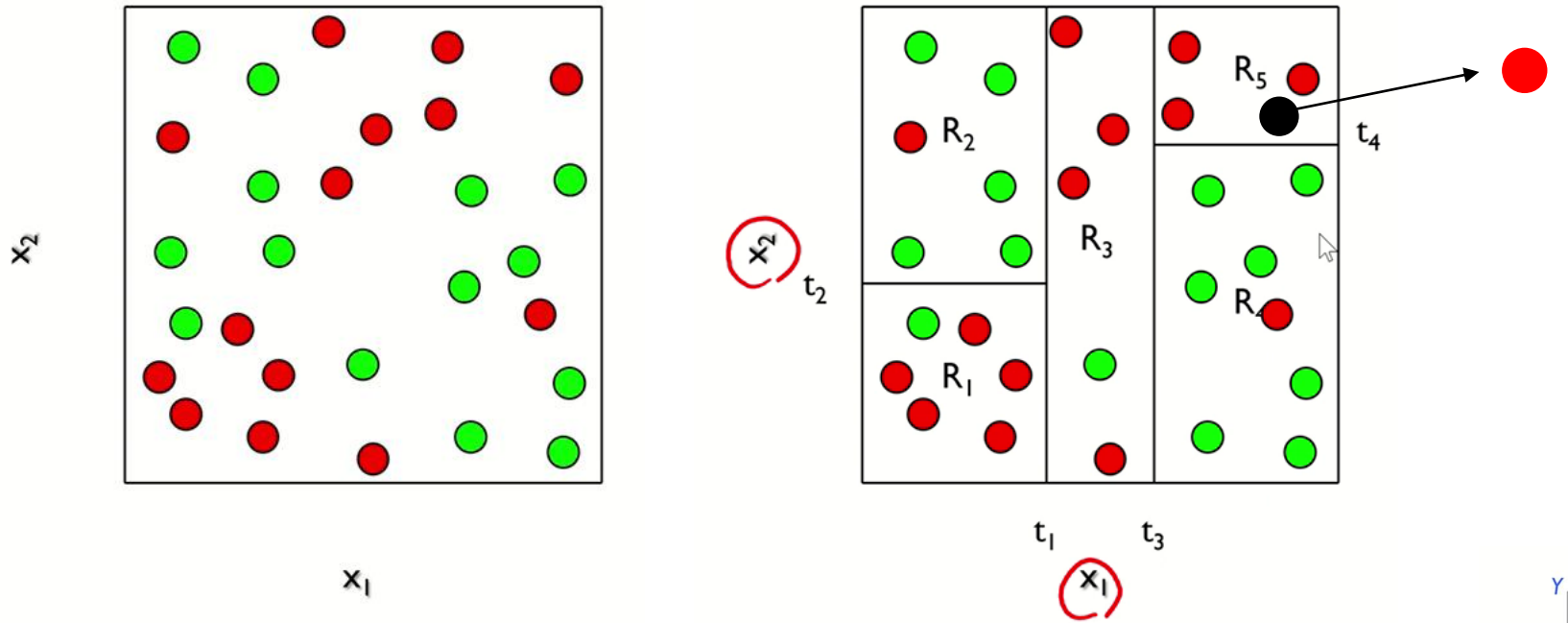
$$\underset{j,s}{\operatorname{argmin}} \left[\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2 \right]$$

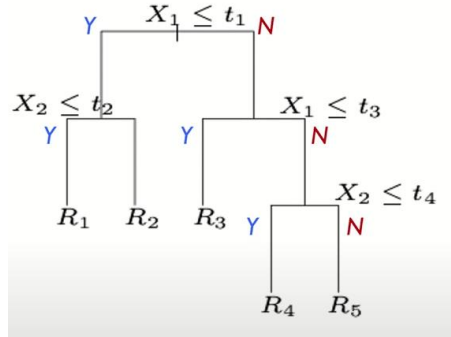
$$= \underset{j,s}{\operatorname{argmin}} \left[\sum_{x_i \in R_1(j,s)} (y_i - \hat{c}_1)^2 + \sum_{x_i \in R_2(j,s)} (y_i - \hat{c}_2)^2 \right]$$

$$\hat{c}_1 = ave(y_i | x_i \in R_1(j,s) \text{ and } \hat{c}_2 = ave(y_i | x_i \in R_2(j,s)$$

J, s 를 바꿔가면서 최소가 되는 값을 찾아나감

분류나무 모델


x2
x1
R1
R2
R3
R4
R5
t1
t2
t3
t4


$X_1 \leq t_1$
Y
N
$X_2 \leq t_2$
$X_1 \leq t_3$
$X_2 \leq t_4$
R_1
R_2
R_3
R_4
R_5

분류나무 모델

범주

- 각 관측치 마다 반응변수 값 $y_i = 1,2,\ldots K,$. 즉, K개의 클래스 존재

끝노드에 속해있는 관측치의 개수

- R_m: 끝노드 m에 해당하며 N_m 관측치 개수를 가지고 있음
- $\hat{p}_{mk}$: 끝노드 m에서 k 클래스에 속해 있는 관측치의 비율

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k)$$

1 1 1
0 0

- 끝노드 m으로 분류된 관측치는 $k(m)$ 클래스로 분류

$$k(m) = \underset{k}{\operatorname{argmax}}\, \hat{p}_{mk}$$

= argmax(0.6, 0.3, 0.1)
= 1

3개 Class 끝노드 1
P11 = 0.6
P12 = 0.3
P13 = 0.1

분류나무 모델

X1, X2 -> R3

$$\hat{f}(x) = \sum_{m=1}^{5} k(m) I\{(x_1, x_2) \in R_m\}$$

$$= k(1) I\{(x_1, x_2) \in R_1\} + k(2) I\{(x_1, x_2) \in R_2\} + k(3) I\{(x_1, x_2) \in R_3\}$$

$$+ k(4) I\{(x_1, x_2) \in R_4\} + k(5) I\{(x_1, x_2) \in R_5\}$$ = K(3)

$$k(m) = \underset{k}{\operatorname{argmax}}\, \hat{p}_{mk}$$

해당 끝노드에 들어간 여러가지 범주에 해당하는 데이터들의 비율 중 가장 큰 확률을 가지고 있는 것

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k)$$

분류나무 모델

- 분류 모델에서의 비용함수 (불순도 측정)

Misclassification rate: $\frac{1}{N_m}\sum_{i\in R_m} I\left(y_i \neq k(m)\right) = 1 - \hat{p}_{(mk)m}$

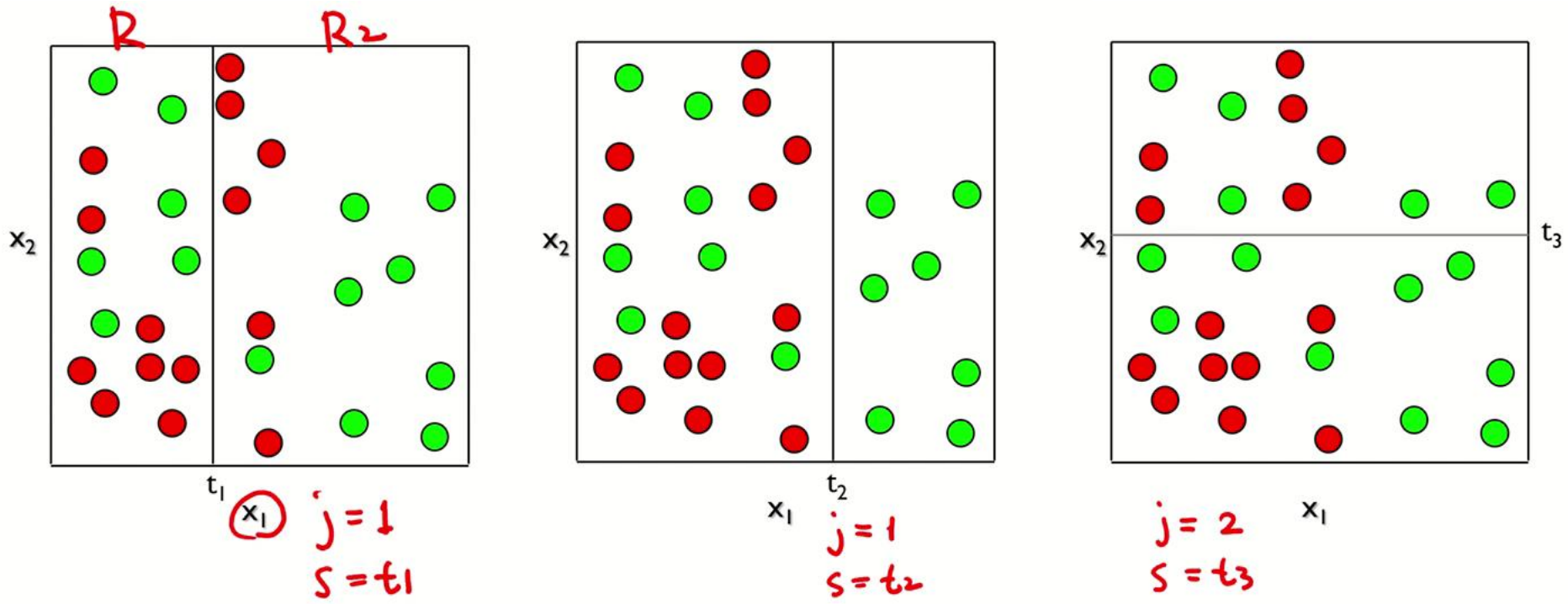
실제 범주와 모델에서 나온 범주가 얼마나 잘 매칭 되었는지

Gini Index: $\sum_{k\neq k'}^{K} \hat{p}_{mk}\hat{p}_{mk'} = \sum_{k=1}^{K} \hat{p}_{mk}\left(1-\hat{p}_{mk}\right)$

Cross-entropy : $-\sum_{k=1}^{K} \hat{p}_{mk}\, log\hat{p}_{mk}$

분할변수(j)와 분할점(s)은 어떻게 결정할까?

$$R_1(j,s) = \{x | x_j \leq s\}$$
$$R_2(j,s) = \{x | x_j > s\}$$



모든 경우의 수 고려

분할변수(j)와 분할점(s)은 어떻게 결정할까?

- Misclassification rate을 비용함수를 사용했을 때,

$$\underset{j,s}{arg\min}\left[\frac{1}{N_{R_1(j,s)}}\sum_{x_i\in R_1(j,s)} I\big(y_i \neq k(m)\big) + \frac{1}{N_{R_2(j,s)}}\sum_{x_i\in R_2(j,s)} I\big(y_i \neq k(m)\big)\right]$$

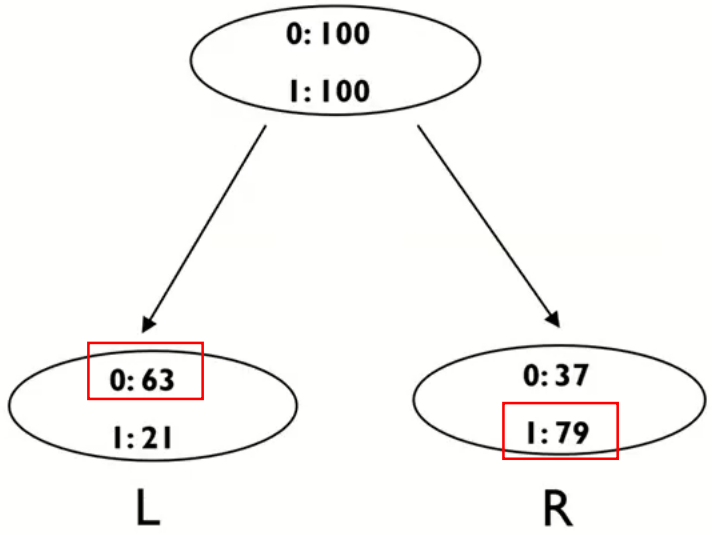
$$k(m) = \underset{k}{\operatorname{argmax}}\,\hat{p}_{mk} \qquad \hat{p}_{mk} = \frac{1}{N_m}\sum_{x_i\in R_m} I(y_i = k)$$

분류나무 모델링 프로세스

- 분할법칙
  - 분할변수와 분할기준은 목표변수의 분포를 가장 잘 구별해주는 쪽으로 정함
  - 목표변수의 분포를 잘 구별해주는 측도로 순수도(purity) 또는 불순도(impurity)를 정의
  - 예를 들어 클래스 0과 클래스 1의 비율이 45%와 55%인 노드는 각 클래스의 비율이 90%와 10%인 마디에 비하여 순수도가 낮다 (또는 불순도가 높다)라고 해석
  - 각 노드에서 분할변수와 분할점의 설정은 불순도의 감소가 최대가 되도록 선택

분류나무 모델링 프로세스

- 오분류율(misclassification rate)



- L 노드의 오분류율 = 21 / (63+21) = 21 / 84 = 0.25
- R 노드의 오분류율 = 37 / (37+79) = 37 / 116 = 0.32
- 총 오분류율 = (37/116) · (116/200) + (21/84) · (84/200) = 0.29

R L

분류나무 모델링 프로세스

- 예제: 어떤 노드의 구성이 다음과 같을 때, **Gini**와 **entropy** 지수를 계산하여라

Gini 지수 계산

$$\varphi(g) = \sum_j P_j(g)(1 - P_j(g))$$

$$= \left(\frac{6}{7} \times \frac{1}{7}\right) + \left(\frac{1}{7} \times \frac{6}{7}\right)$$

$$= 0.2449$$

Entropy 지수 계산

$$\varphi(g) = -\sum_j P_j(g) log P_j(g)$$

$$= -\frac{6}{7} log \frac{6}{7} - \frac{1}{7} log \frac{1}{7}$$

$$= 0.1781$$

Information Gain

- Entropy (S) $= \sum_{i=1}^{c} -p_i \log_2 p_i$

c is the number of class, p_i is the proportion of S belong to class i.

- For example, suppose S is a collection of 14 examples [9+, 5-].

$$Entropy[9+, 5-] = \sum_{i=1}^{2} -p_i \log_2 p_i = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0.94$$ 꽤나 혼재되어 있다 !

- **Information gain (IG):** the expected reduction in entropy caused by partitioning the data according to this variable. (특정 변수를 사용했을 때 entropy 감소량)

- IG(S,A): Information gain of a variable A. (변수A를 사용했을 때 entropy 감소량)

$$IG(S, A) = Entropy(S) - \sum_{v \in value(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

총 데이터

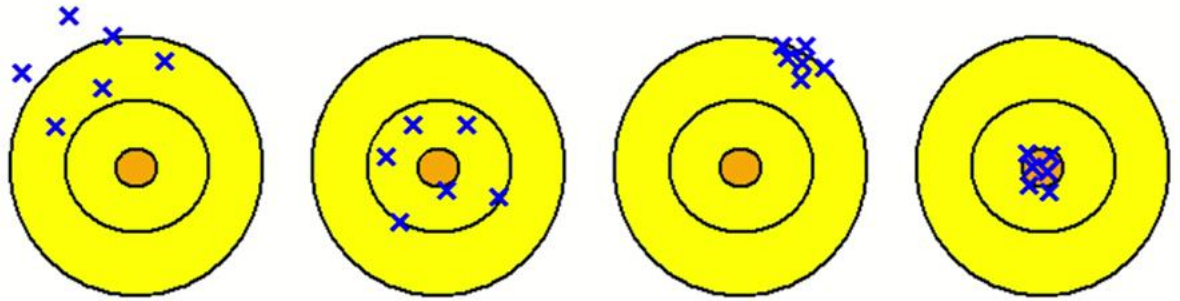
특정 변수를 사용했을 때 얼마나 엔트로피를 감소 시켰는지

- value (A): the set of all possible values for a variable A.

- S_v: the subset of S for which variable A has value v.

트리 모델의 단점

- 계층적 구조로 인해 중간에 에러가 발생하면 다음 단계로 에러가 계속 전파
- 학습 데이터의 미세한 변동에도 최종 결과 크게 영향
- 적은 개수의 노이즈에도 크게 영향
- 나무의 최종노드 개수를 늘리면 과적합 위험 (Low Bias, Large Variance)



Bias	High	Low	High	Low
Variance	High	High	Low	Low

- 해결방안 → 랜덤 포레스트 (Random forest)

DecisionTreeClassifier 파라미터

min_samples_split	노드를 분할하기 위한 최소한의 샘플 데이터 수	디폴트: 2 작게 설정될수록 과적합 가능성이 증가
min_samples_leaf	말단 노드(leaf)가 되기 위한 최소한의 샘플 데이터 수	특정 클래스의 데이터가 극도로 작을 수 있으므로 이 경우는 작게 설정할 필요가 있다.
max_features	최적의 분할을 위해 고려할 최대 피처 개수	최적의 분할을 위해 고려해야할 최대 피처 개수
Max_depth	트리의 최대 깊이를 규정	깊이가 깊어지면 min_samples_split 설정대로 최대 분할하여 과적합할 수 있으므로 적절한 값으로 제어 필요
Max_leaf_nodes	말단 노드의 최대 개수	