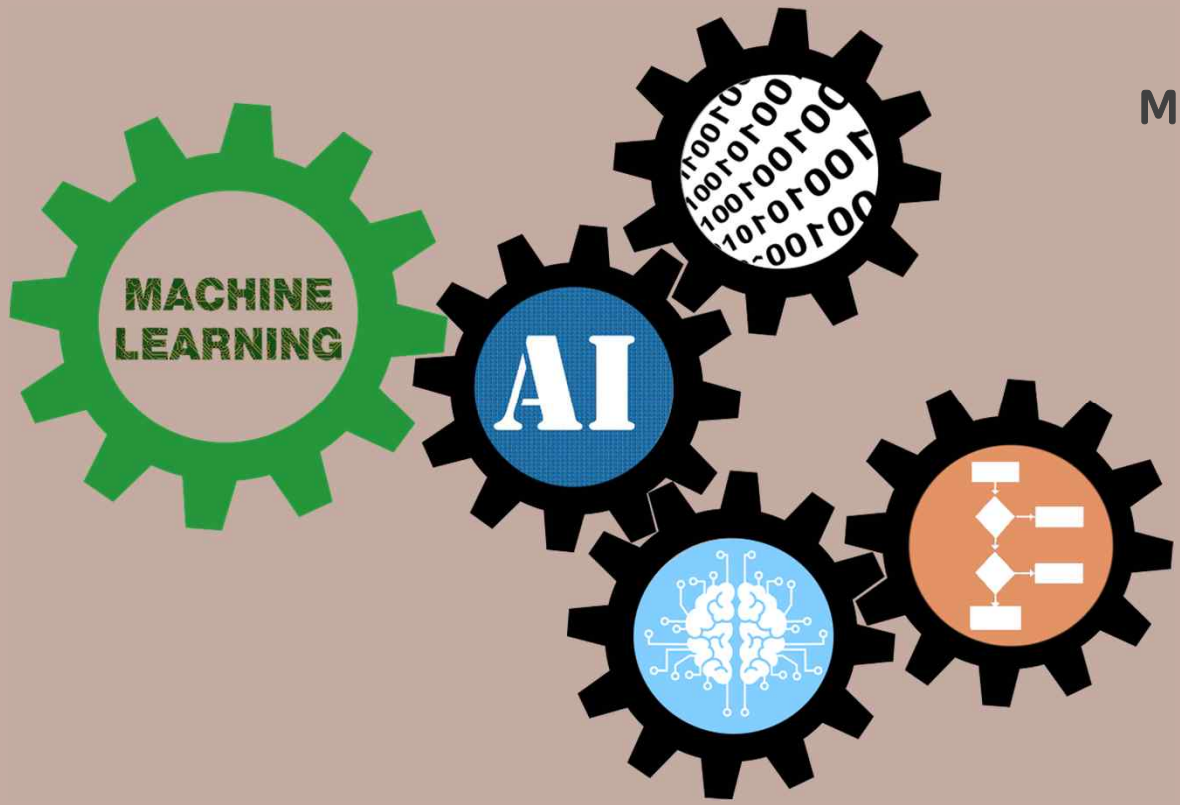



Machine Learning 7강

# 결정 트리와 랜덤 포레스트

컴퓨터과학과 이관용 교수

# 학습목차

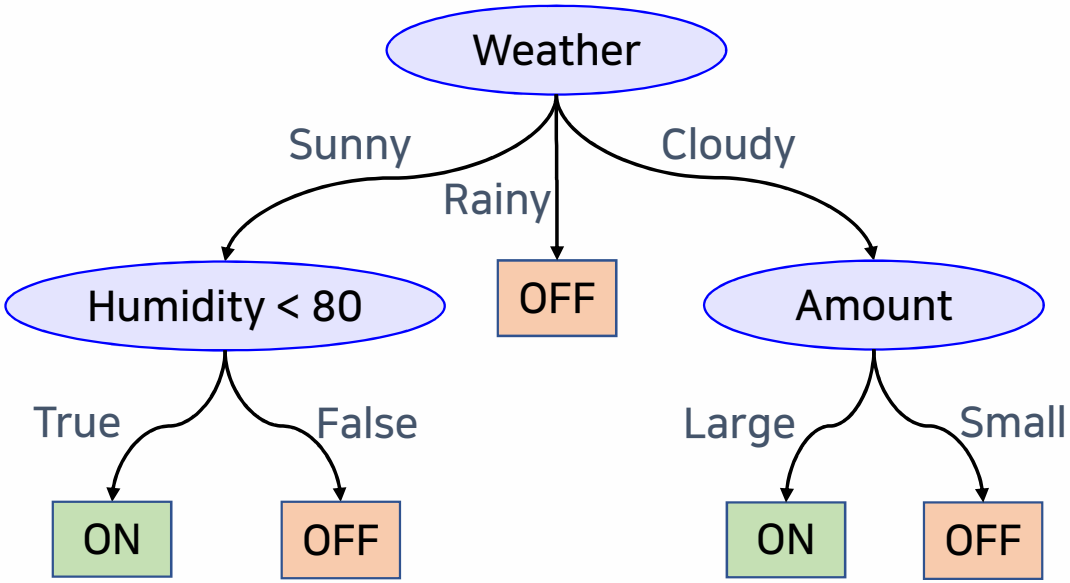
# 1 결정 트리

# 결정 트리 decision tree?

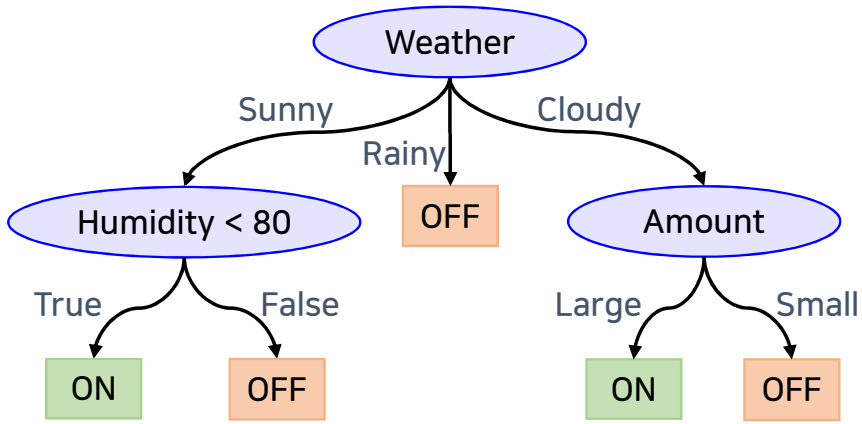
- 주어진 문제에 관해 결정을 내리는 함수를 트리 형태로 구성
  - 분류 문제를 위해 개발
    → 회귀 문제로 확장 → "CART" Classification And Regression Trees
- 뛰어난 설명력 explainability 제공
  - 트리 구조에 각 입력 요소의 역할이 잘 표현되어 학습 결과를 설명
- 과다적합 문제
  - 복잡한 함수의 표현 과정에서 데이터의 노이즈에 민감한 문제
    → 앙상블 학습기법을 결합한 방법("랜덤 포레스트") 등장

# 결정 트리의 예

- 세탁기 동작 여부(ON, OFF)를 결정하는 결정 트리
  - 속성(판단을 내리는 데 사용하는 결정 요인) → 날씨, 습도, 세탁량



# 결정 트리 vs 규칙 기반의 표현




데이터를 이용한 학습을 통해 자동으로 생성

## 규칙의 집합으로의 표현

개발자가 임의로 정의/표현

```
if (Weather == Rainy) then Machine = OFF

if (Weather == Sunny) and (Humidity < 80) then Machine = ON
if (Weather == Sunny) and (Humidity >= 80) then Machine = OFF

if (Weather == Cloudy) and (Amount == Large) then Machine = ON
if (Weather == Cloudy) and (Amount == Small) then Machine = OFF
```

# 결정 트리의 학습

## 학습 데이터의 예

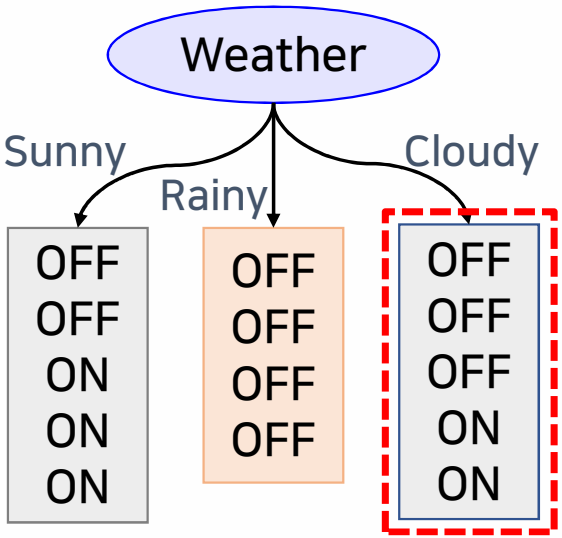
| 날씨 (Weather) | 온도 (Temperature) | 습도 (Humidity) | 세탁량 (Amount) | 세탁기 동작 여부 |
|---|---|---|---|---|
| Cloudy | High | 85 | Small | OFF |
| Cloudy | High | 82 | Large | ON |
| Rainy | High | 90 | Small | OFF |
| Sunny | Medium | 80 | Small | OFF |
| Sunny | Low | 60 | Small | ON |
| Sunny | Low | 52 | Large | ON |
| Rainy | Low | 95 | Large | OFF |
| Cloudy | Medium | 83 | Small | OFF |
| Cloudy | Low | 62 | Small | OFF |
| Sunny | Medium | 46 | Small | ON |
| Cloudy | Medium | 55 | Large | ON |
| Rainy | Medium | 97 | Large | OFF |
| Rainy | High | 93 | Small | OFF |
| Sunny | Medium | 81 | Large | OFF |

# 결정 트리의 학습

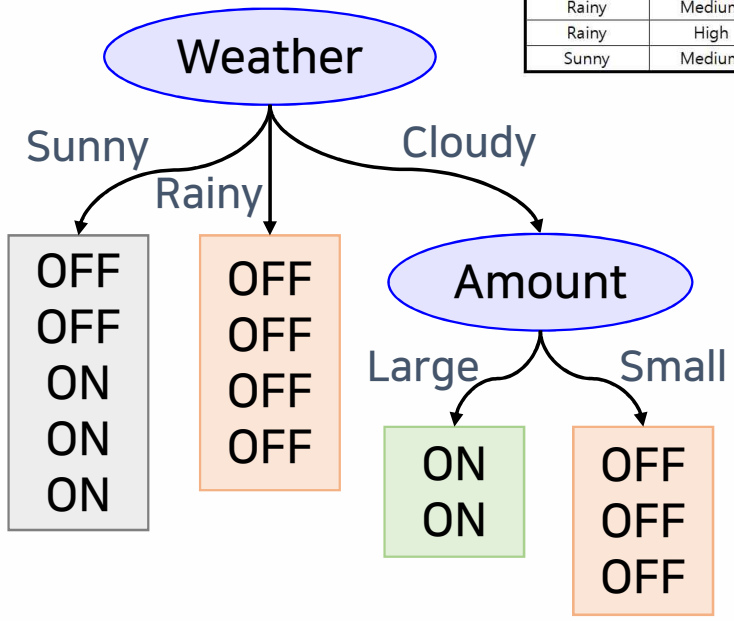
## 학습 과정

| 날씨 (Weather) | 온도 (Temperature) | 습도 (Humidity) | 세탁량 (Amount) | 세탁기 동작 여부 |
|---|---|---|---|---|
| Cloudy | High | 85 | Small | OFF |
| Cloudy | High | 82 | Large | ON |
| Rainy | High | 90 | Small | OFF |
| Sunny | Medium | 80 | Small | OFF |
| Sunny | Low | 60 | Small | ON |
| Sunny | Low | 52 | Large | ON |
| Rainy | Low | 95 | Large | OFF |
| Cloudy | Medium | 83 | Small | OFF |
| Cloudy | Low | 62 | Small | OFF |
| Sunny | Medium | 46 | Small | ON |
| Cloudy | Medium | 55 | Large | ON |
| Rainy | Medium | 97 | Large | OFF |
| Rainy | High | 93 | Small | OFF |
| Sunny | Medium | 81 | Large | OFF |

[step 1]

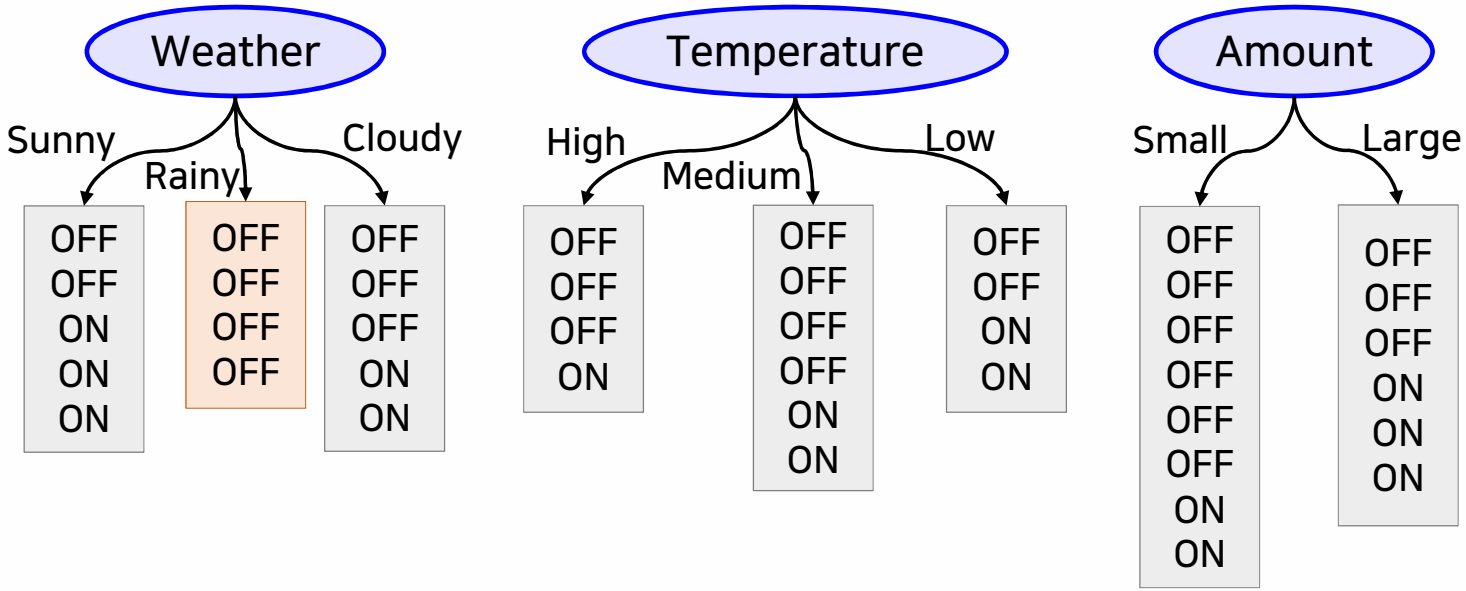



[step 2]



# 결정 트리의 학습

- 각 노드에 어떤 속성(결정 요인)을 배정할 것인가?



# 결정 트리의 학습

## 속성 선택을 위한 평가 기준

### 지니 불순도 $I(N)$ Gini Impurity

✓ 각 노드에 할당된 클래스 레이블이 얼마나 다른지 그 혼합 정도를 측정

$$I(N) = \sum_{i=1}^{K} p_i(1-p_i) = 1 - \sum_{i=1}^{K} (p_i)^2$$

$K \rightarrow$ 클래스의 개수

$p_i \rightarrow$ 노드 $N$에 할당된 데이터 그룹에 속한 $i$번째 클래스의 비율

### 지니 평가지수 $G(R_a)$ Gini criterion

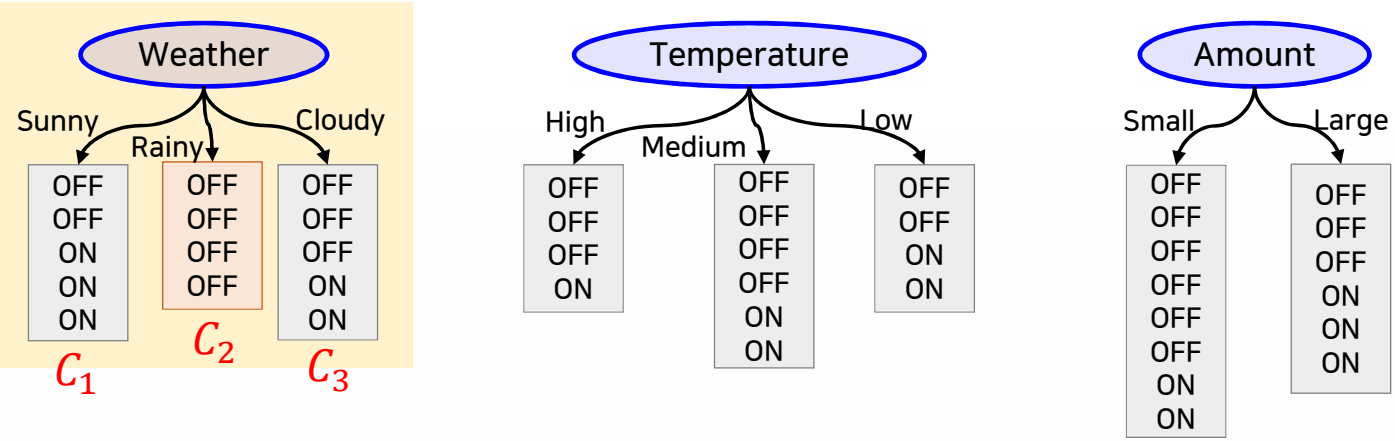
✓ 속성 $a$를 갖는 부모 노드 $R_a$에서 자식 노드들의 지니 불순도의 가중합

$$G(R_a) = \sum_{i=1}^{M} \frac{|C_i|}{|R_a|} I(C_i)$$

$M$개의 자식 노드 $\rightarrow C_1, C_2, \cdots, C_M$

$|R_a|, |C_1|, \cdots, |C_M| \rightarrow$ 각 노드에 속하는 데이터 개수

# 결정 트리의 학습




**지니 불순도** $I(N) = 1 - \sum_{i=1}^{K}(p_i)^2$

$$I(C_1) = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 0.48$$

$$I(C_2) = 1 - \left(\frac{0}{4}\right)^2 - \left(\frac{4}{4}\right)^2 = 0$$

$$I(C_3) = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = 0.48$$

**지니 평가지수** $G(R_a) = \sum_{i=1}^{M} \frac{|C_i|}{|R_a|} I(C_i)$

$$G(R_{weather}) = \frac{5}{14} \times 0.48 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.48 = \mathbf{0.3429}$$

$$G(R_{Temperature}) = 0.4405$$

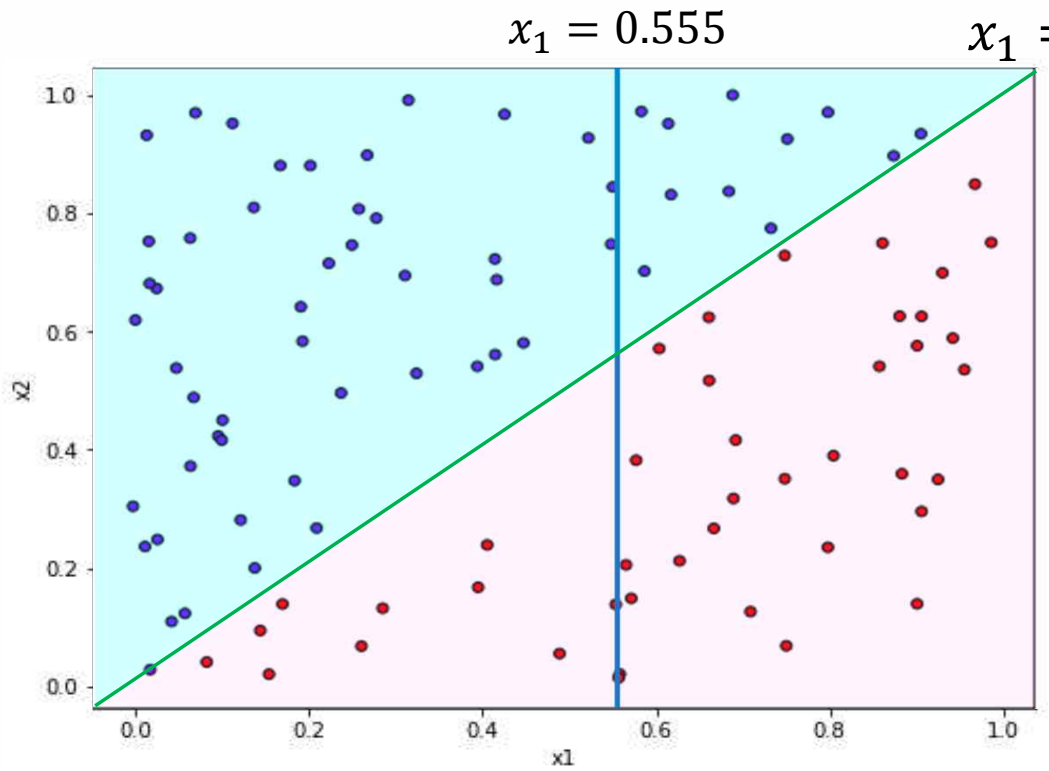
$$G(R_{Amount}) = 0.4048$$

# 속성 노드 선택을 위한 다양한 평가지수

- **정보 이득** information gain
  - 데이터 집합의 분할 전후의 엔트로피의 차이
    - ✓ 엔트로피 → 데이터의 혼잡도. 낮을수록 데이터의 순도가 높음
- **분산 감소량** reduction in variance
  - 모든 노드에 대한 분산의 가중 평균
    - ✓ 분산 → 데이터의 동질성을 표시. 데이터가 완전히 같으면 분산은 0
  - 회귀 문제에서 주로 사용
- **Chi-square**
  - 부모 노드와 하위 노드 간 차이의 통계적 유의성을 활용

# 결정 트리를 이용한 분류

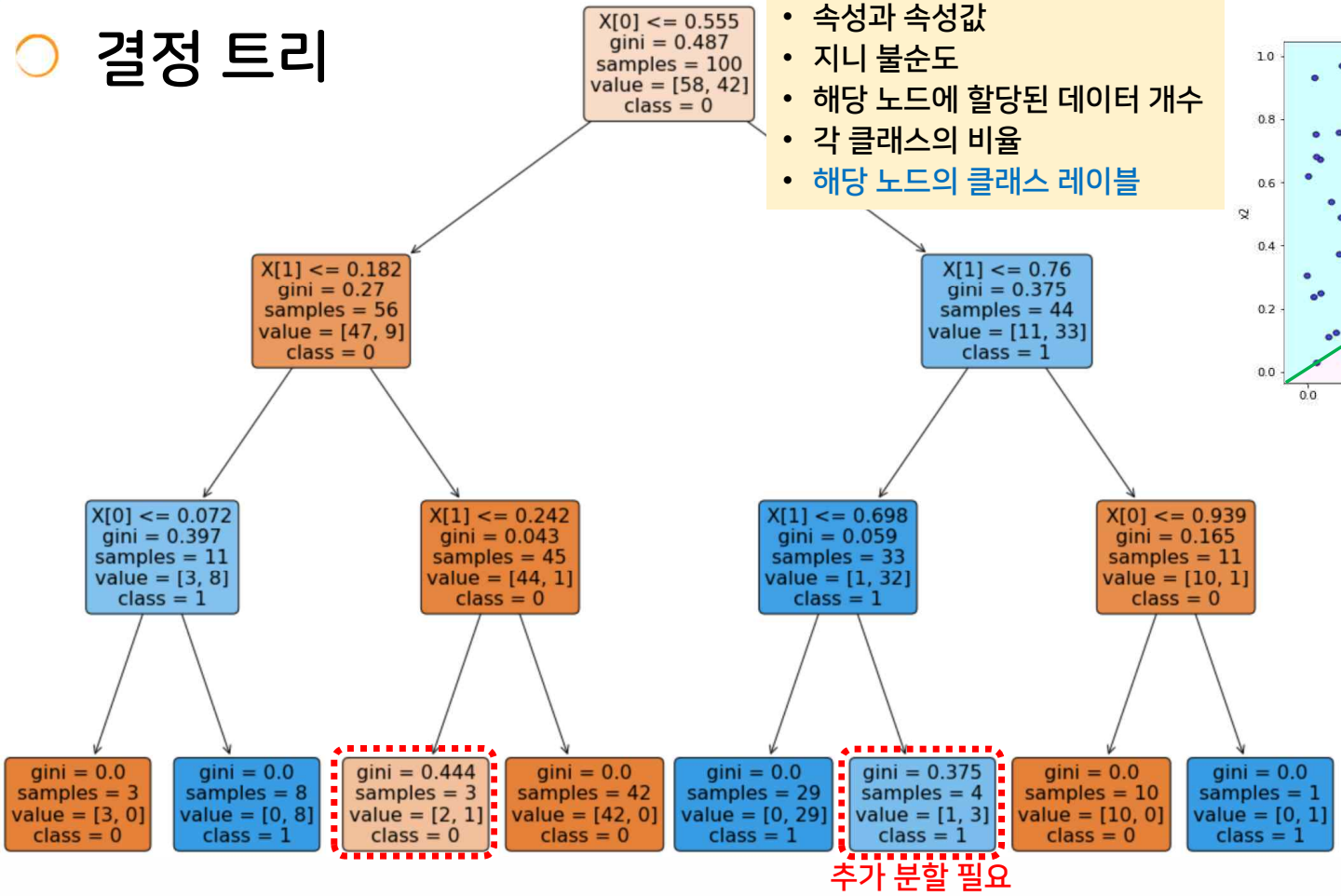
- 2차원 데이터의 예: $x_1, x_2 \in (0,1)$

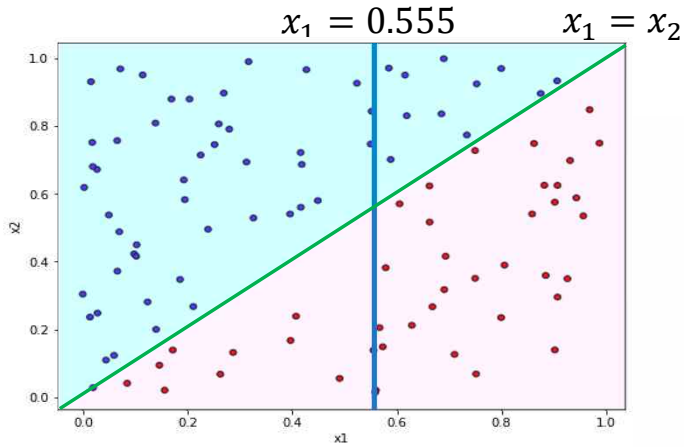


# 결정 트리를 이용한 분류

## 결정 트리

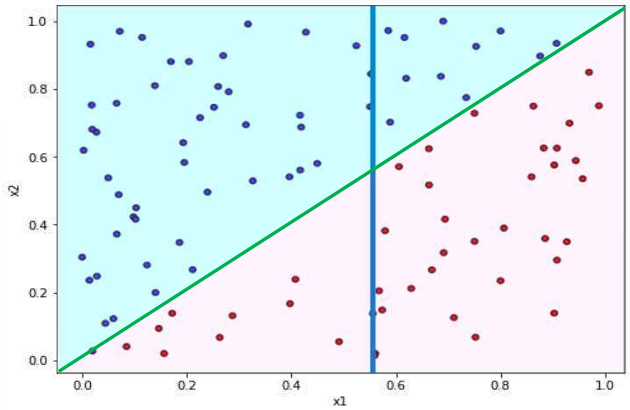
- 속성과 속성값
- 지니 불순도
- 해당 노드에 할당된 데이터 개수
- 각 클래스의 비율
- 해당 노드의 클래스 레이블



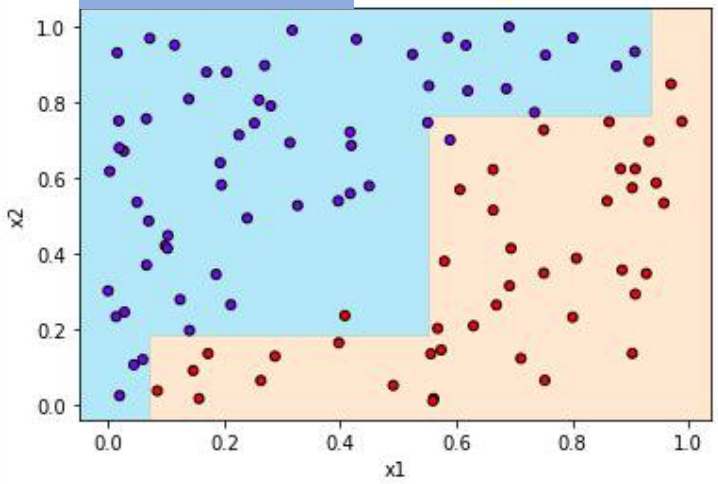



# 결정 트리를 이용한 분류

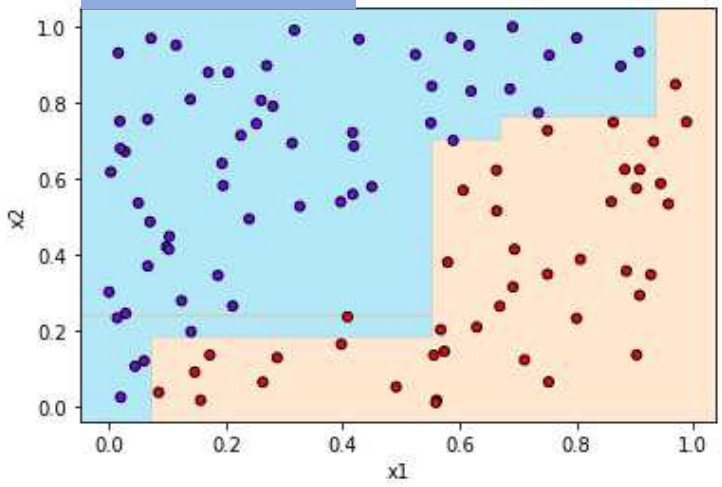
- 결정 경계




깊이 4인 경우

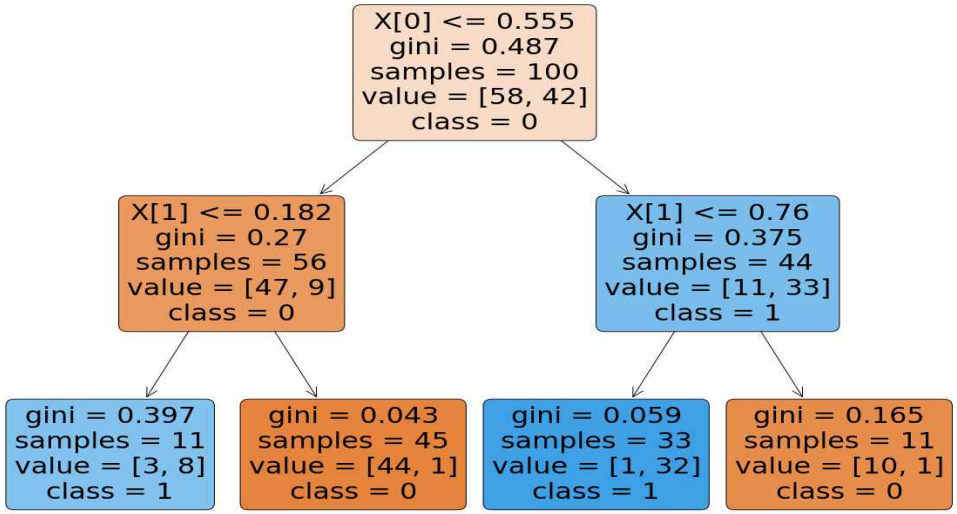



깊이 5인 경우 → 모든 학습 데이터의 처리가 완전히 끝난 상태

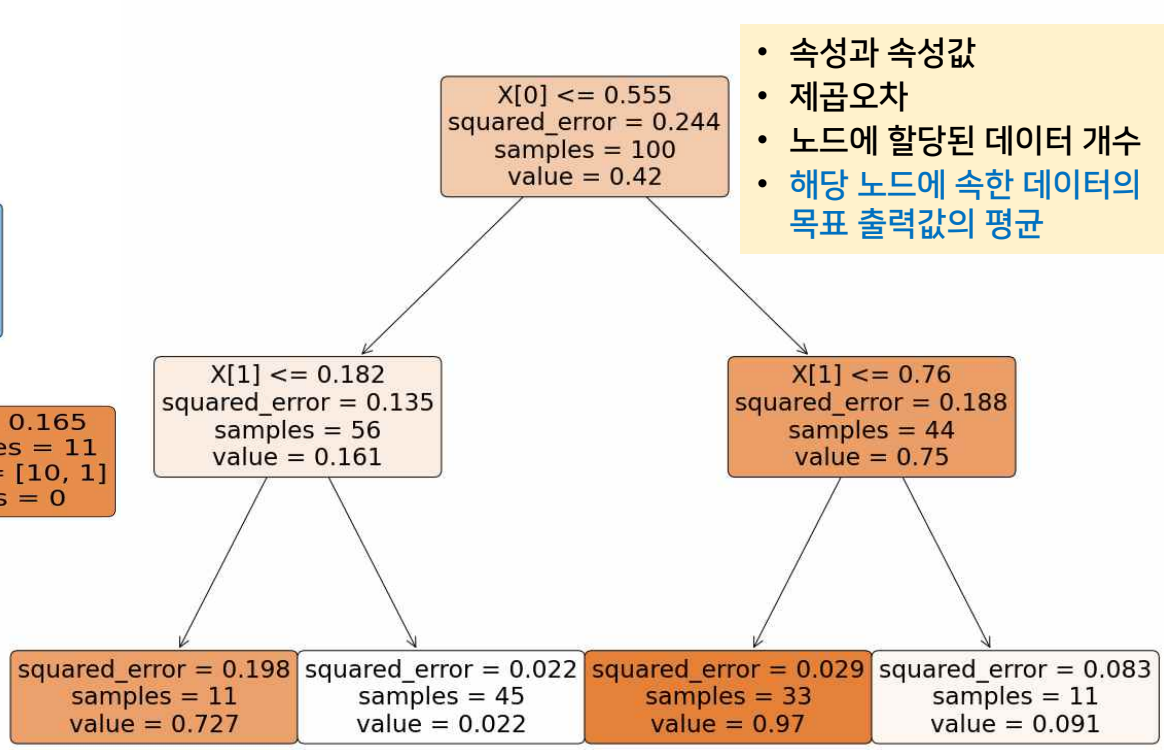


# 회귀를 위한 결정 트리

- 회귀 문제 → $y = f(x_1, x_2)$ $y \in \{0,1\}$




분류 문제로 접근한 경우의 결정 트리

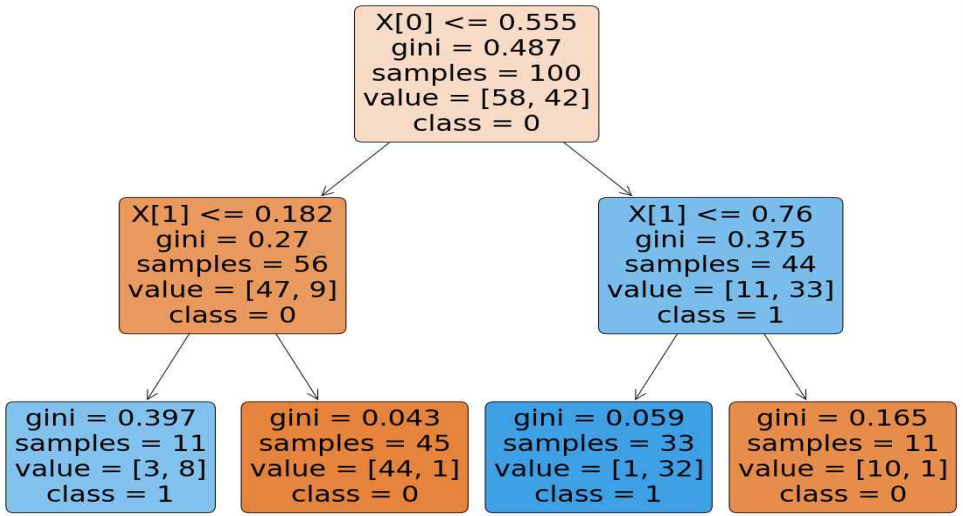



- 속성과 속성값
- 제곱오차
- 노드에 할당된 데이터 개수
- 해당 노드에 속한 데이터의 목표 출력값의 평균

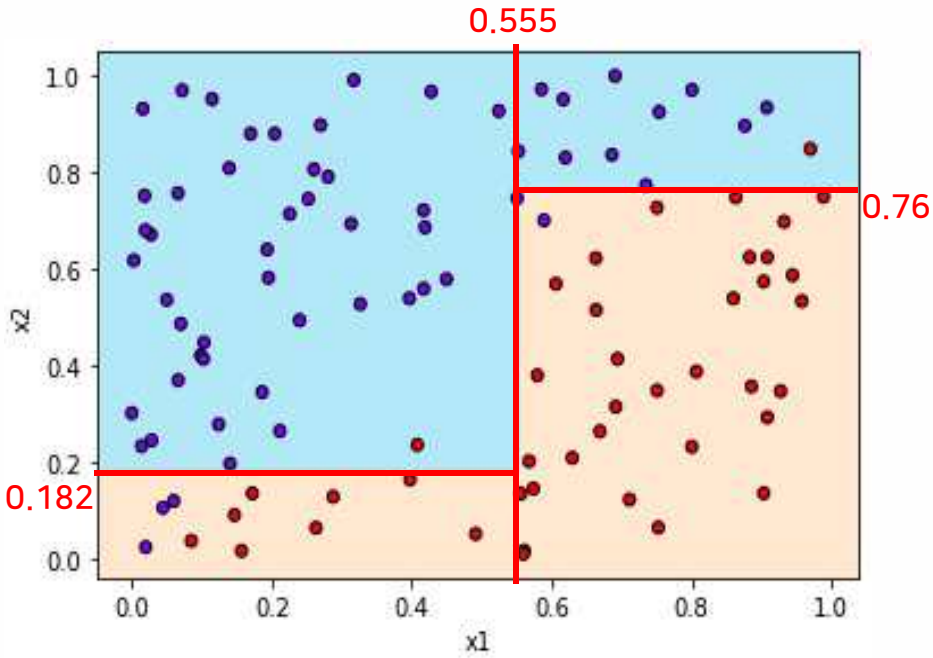
회귀 문제로 접근한 경우의 결정 트리

# 회귀를 위한 결정 트리

- 문제 유형에 따른 결정 트리의 출력: **분류 문제의 경우**



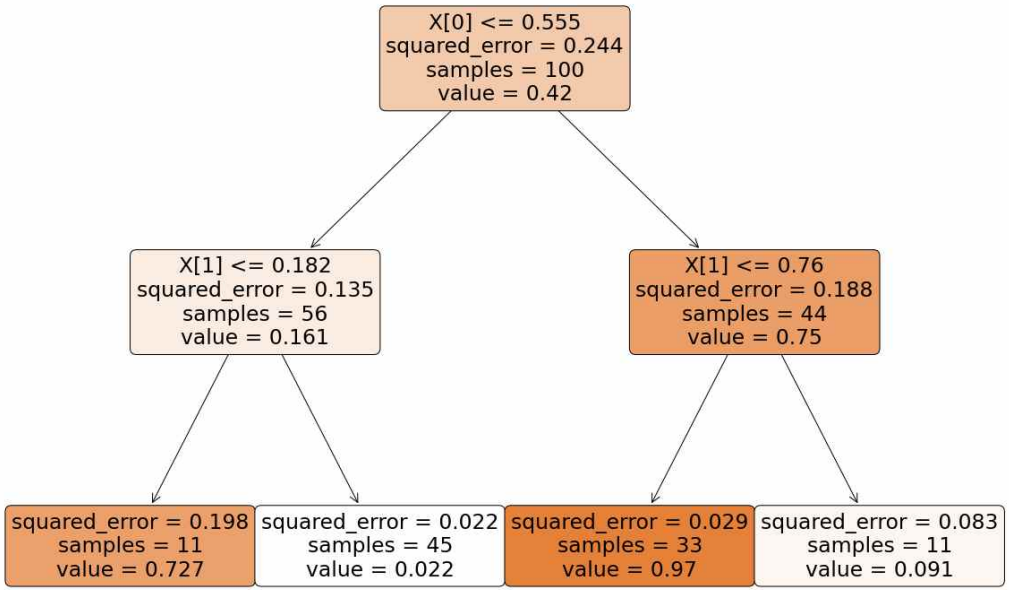

결정 트리



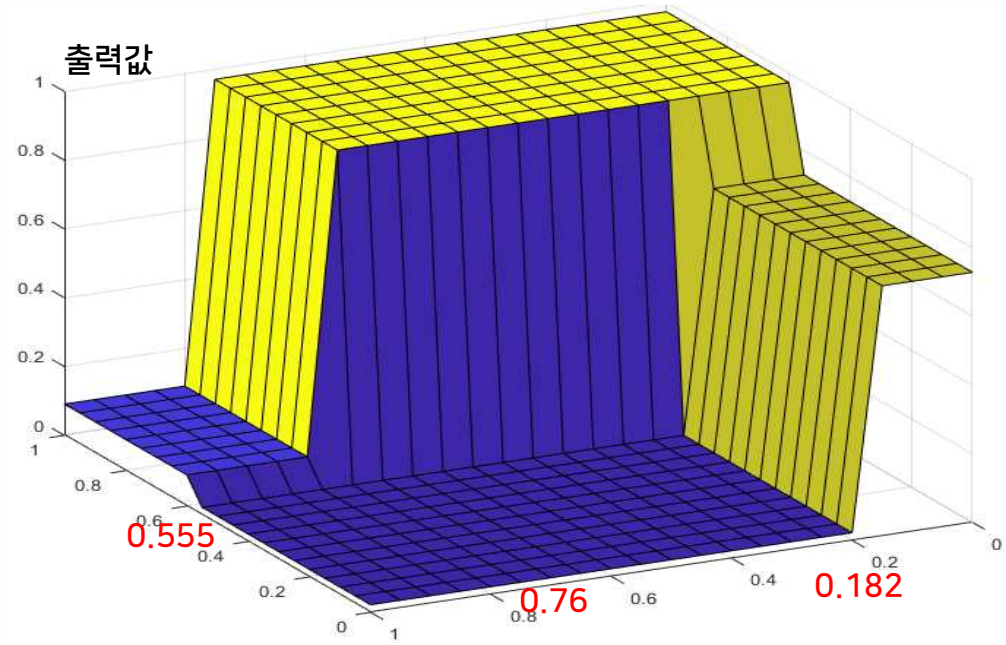

결정 영역

# 회귀를 위한 결정 트리

- 문제 유형에 따른 결정 트리의 출력: **회귀 문제의 경우**

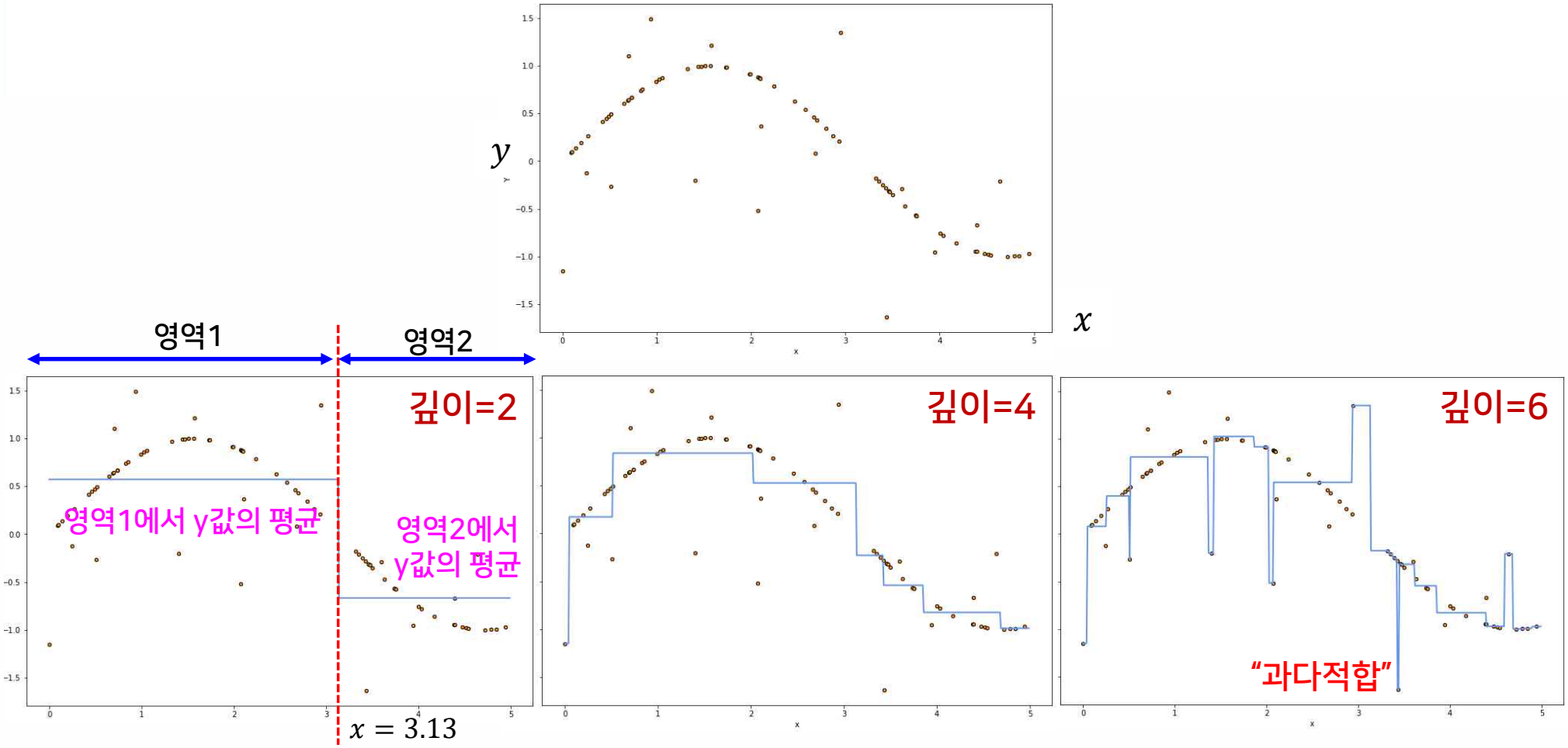



결정 트리




회귀함수

# 일반적인 회귀 문제의 결정 트리



# 결정 트리의 문제

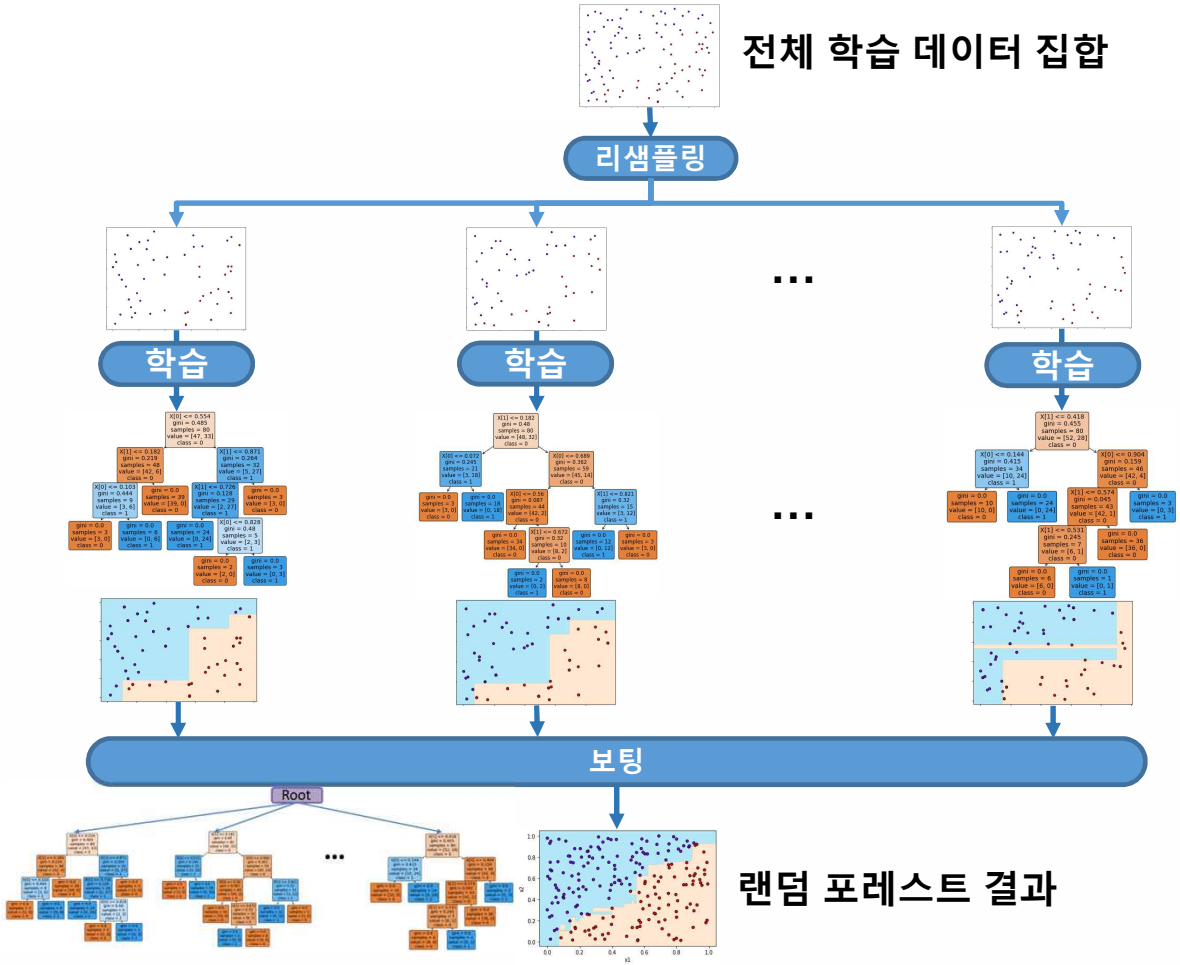
- **과다적합**
  - 모든 학습 데이터에 대해 완벽한 학습
- **간단한 해결책**
  - 조기종료 early stopping
    - ✓ 데이터를 더 분할해도 성능이 향상되지 않을 때 노드의 분할을 종료
  - 가지치기 pruning
    - ✓ 전체 트리를 만든 후 불필요한 노드들을 제거
- **발전된 해결책** → "랜덤 포레스트" random forest

# 2 랜덤 포레스트

# 랜덤 포레스트?

- 결정 트리와 앙상블 학습기법을 결합한 방법
  - 배깅 방법으로 데이터를 리샘플링하여
    $M$개의 결정 트리를 학습하고 결합하는 방법
    - ✓ 결합 방법 → 주로 보팅법(분류 문제), 출력값의 평균(회귀 문제)
  - "포레스트" → $M$개의 서로 다른 결정 트리의 결합
    "랜덤" → 결정 트리 간의 차이가 랜덤으로 추출된 데이터 샘플에 기인
  - 장점
    - ✓ 간단한 학습기의 결합으로 복잡한 함수 표현 및 일반화 성능 향상
    - ✓ 높은 설명 능력, 빠른 학습

# 랜덤 포레스트의 생성 과정



# 랜덤 포레스트의 학습

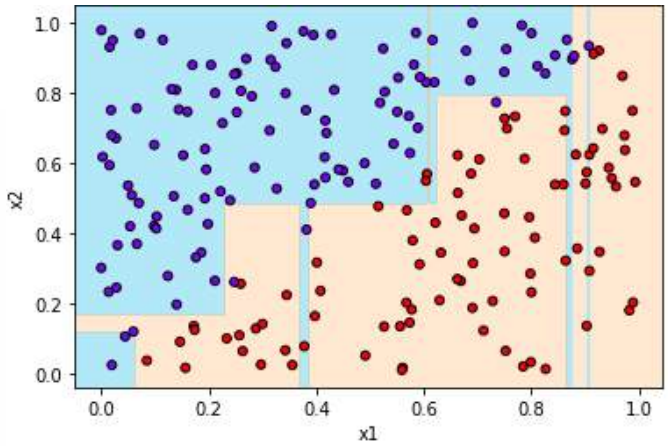
① $N$개의 데이터로 이루어진 학습 데이터 집합 $X$를 준비하고, 각 결정 트리의 학습에 사용될 데이터 집합의 크기 $\widetilde{N}$을 정한다.$(\widetilde{N} \leq N)$

② $i$번째 결정 트리를 학습하기 위해 트리의 깊이를 결정하고, 학습 데이터 집합 $X$로부터 $\widetilde{N}$개의 데이터를 랜덤하게 선출하여 데이터 집합 $X_i$를 만든다. 이때 같은 데이터가 중복해서 선출되는 것도 허락한다(복원추출).

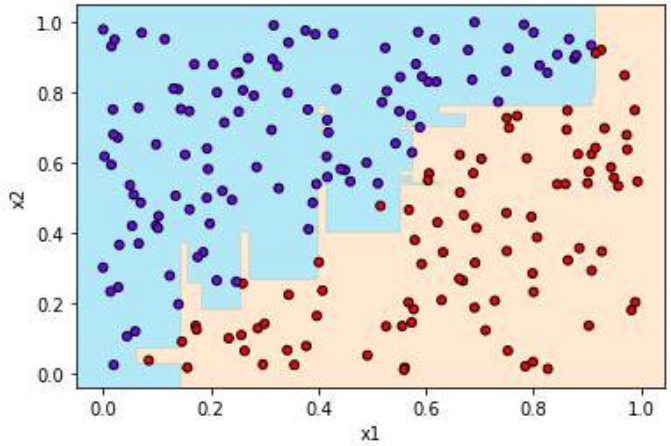
③ 데이터 집합 $X_i$를 이용하여 결정 트리를 학습하여 $i$번째 판별함수(또는 회귀함수) $h_i(\boldsymbol{x})$를 얻는다.

④ ②~③ 과정을 $M$번 반복하여 서로 다른 $M$개의 결정 트리를 생성하고, 이들을 결합하여 최종 판별함수(또는 회귀함수) $f(h_1, h_2, \cdots, h_M)$을 찾는다.

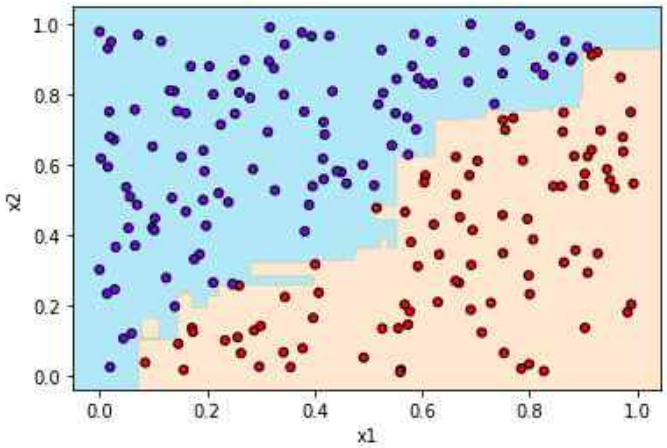
# 랜덤 포레스트를 이용한 분류



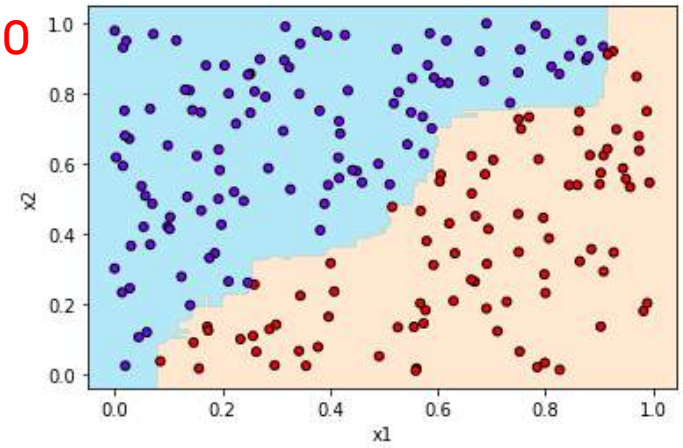

M=1




M=3

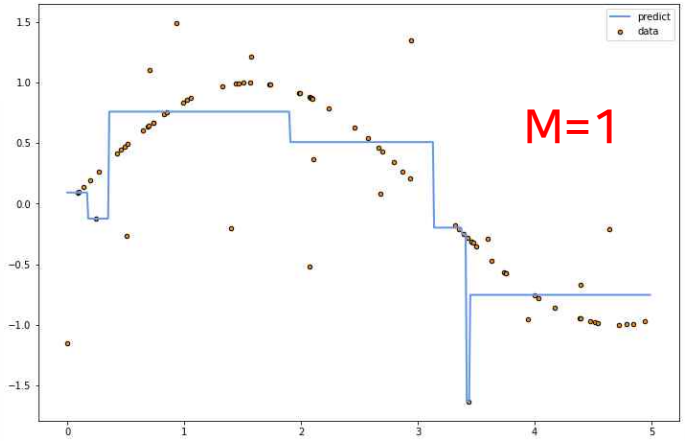


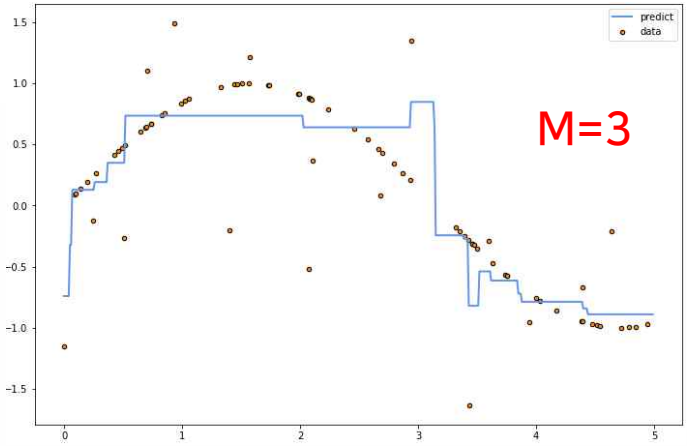

M=10

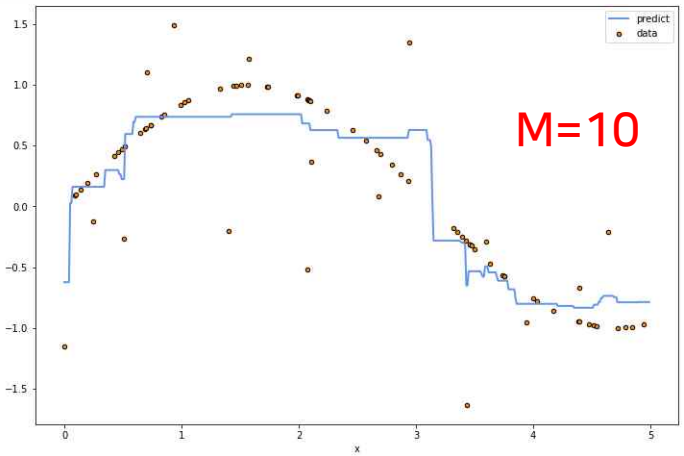


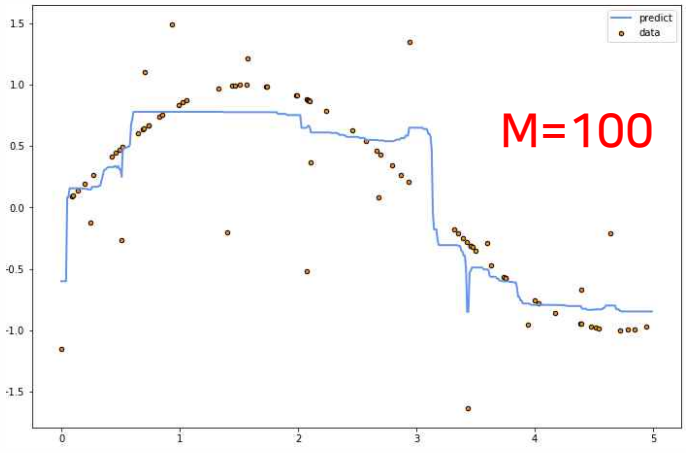

M=100

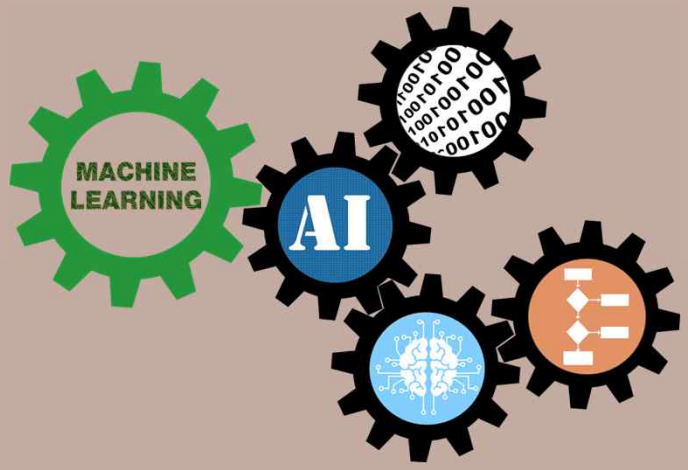
# 랜덤 포레스트를 이용한 회귀






다음시간안내

제8강

# SVM과 커널법