

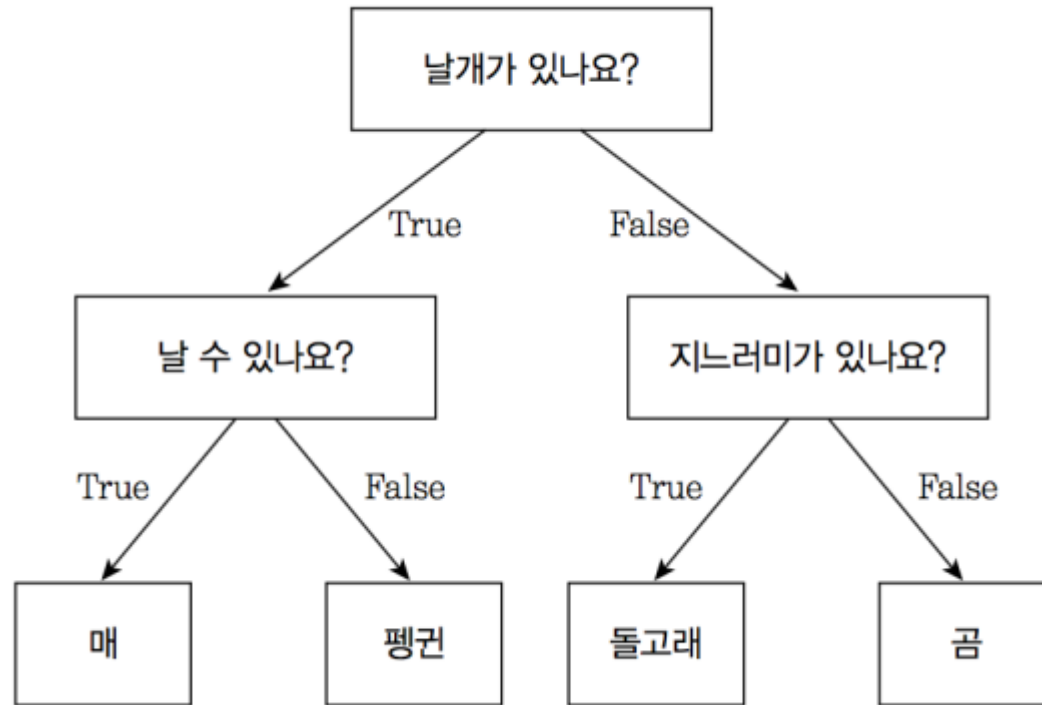
# 의사결정나무

## Decision Tree

# 의사결정나무 (Decision Tree)

1. 개념정리
2. 불순도 (Impurity)
3. 정보획득(Information gain)
4. 학습과정
5. 실습

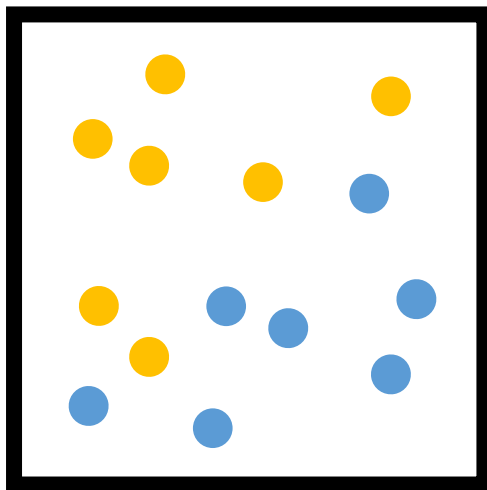
# 개념정리



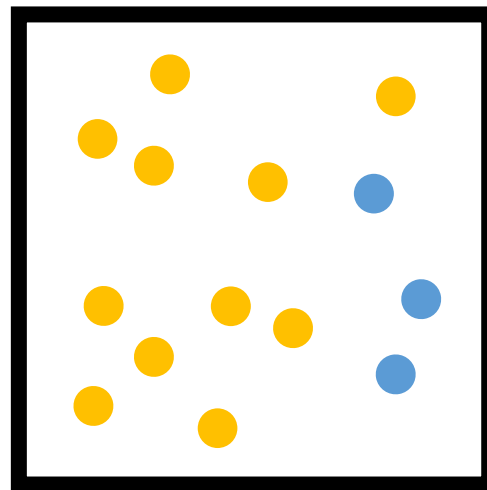
# 불순도 (Impurity)

샘플 집합의 순도를 어떻게 측정할까?

(a)



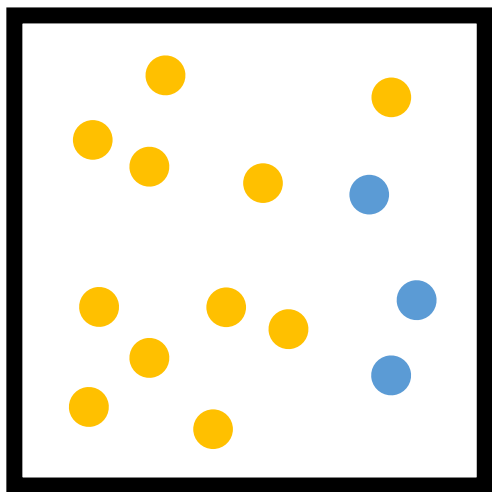
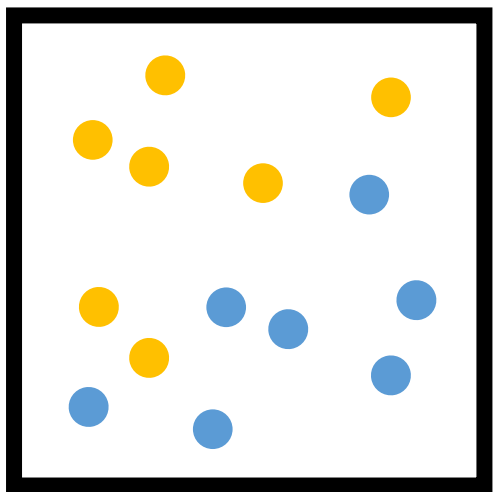
(b)



둘중 불순도가 더 높은 것은?

# 불순도 지표: Gini

$$Gini(p_A) = 1 - \sum_{i=1}^m (p_{Ai})^2$$

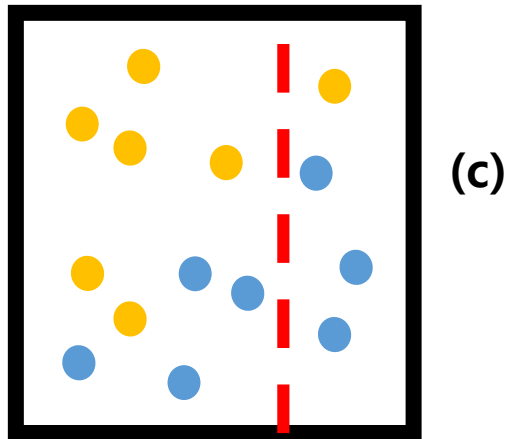
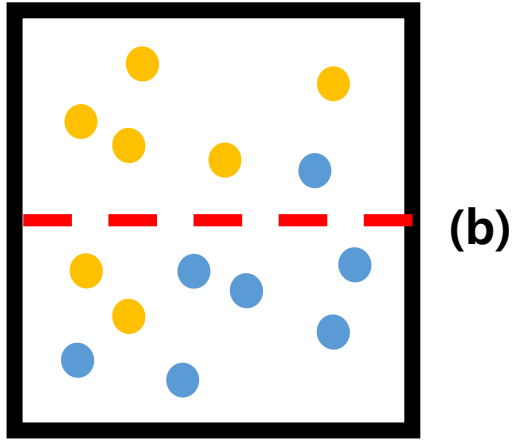
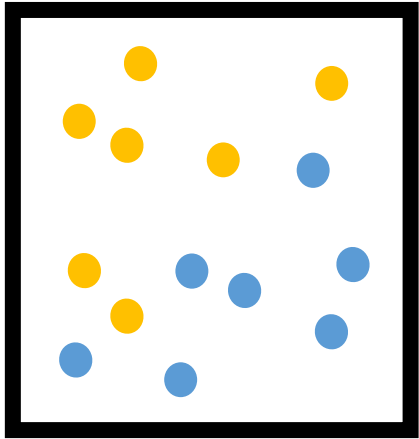


(a)

(a)

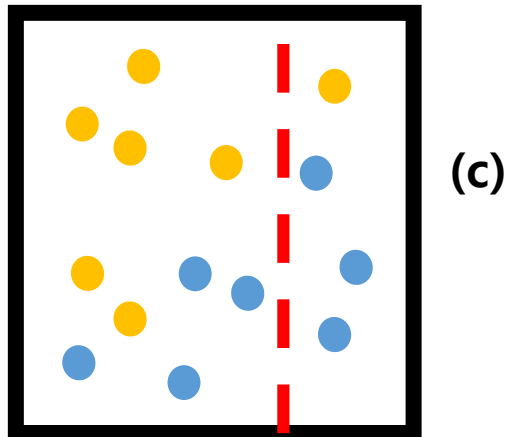
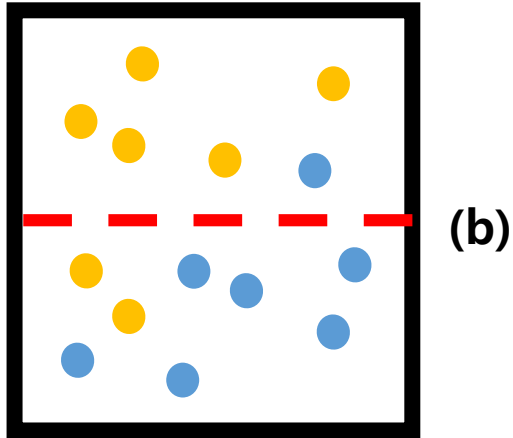
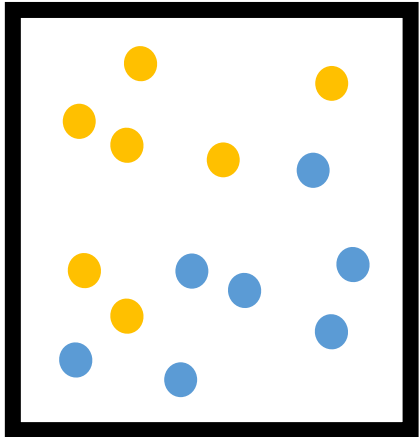
따라서 (a) 영역의 불순도가 더 높다

# 정보획득 (Information gain)



# 정보획득 (Information gain)

$$InformationGain = E_{parent} - w1 * E_{children1} - w2 * E_{children2} - \dots$$



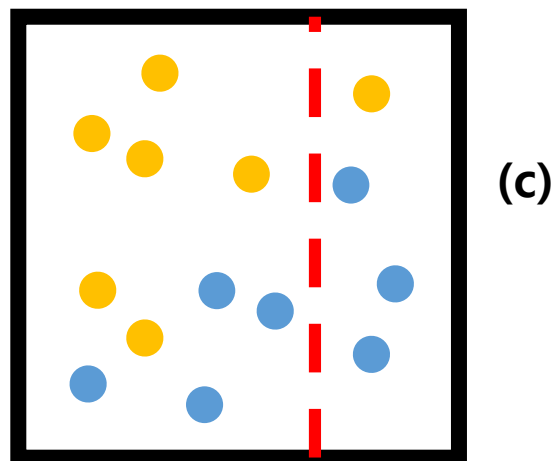
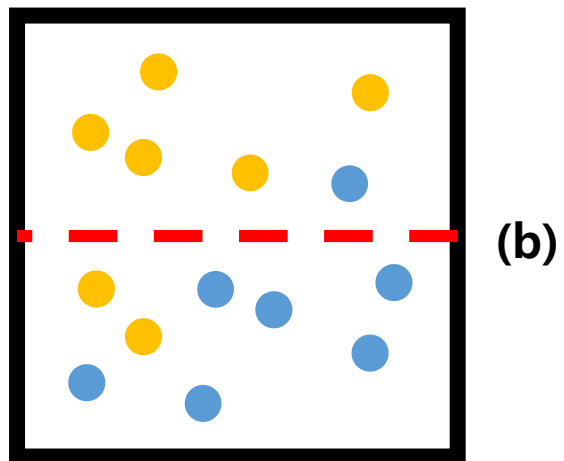
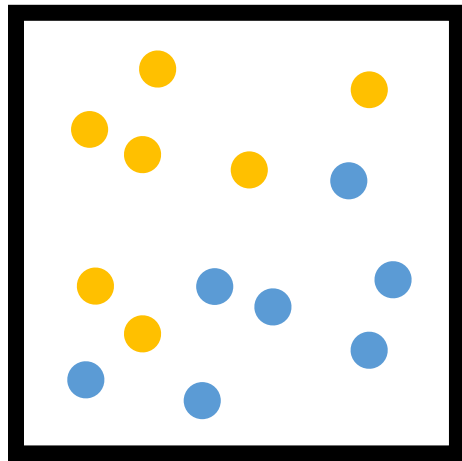
\*  $E_{parent}$ : 부모의 엔트로피

\*  $E_{children}$ : 자식의 엔트로피

\*  $w$ : weighted average, 가중평균

# 정보획득 (Information gain)

샘플 집합의 순도를 어떻게 측정할까?





# 학습과정

## 1. ROOT 노드의 불순도 계산

날짜	바람	습도	온도	날씨
1	약	높음	높음	비
2	강	정상	낮음	맑음
3	약	정상	낮음	맑음
4	강	높음	낮음	비
5	약	정상	높음	맑음
6	강	높음	높음	비
7	강	높음	낮음	맑음
8	약	정상	낮음	맑음
9	강	높음	높음	비
10	강	높음	높음	비

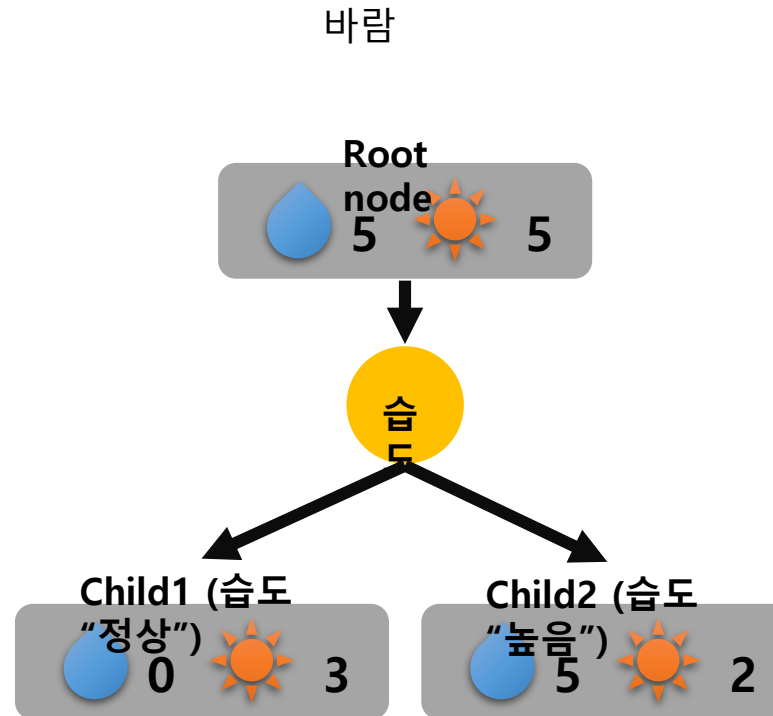
Root node



# 학습과정

## 2. 각 송근성별 자식노드의 불순도 계산

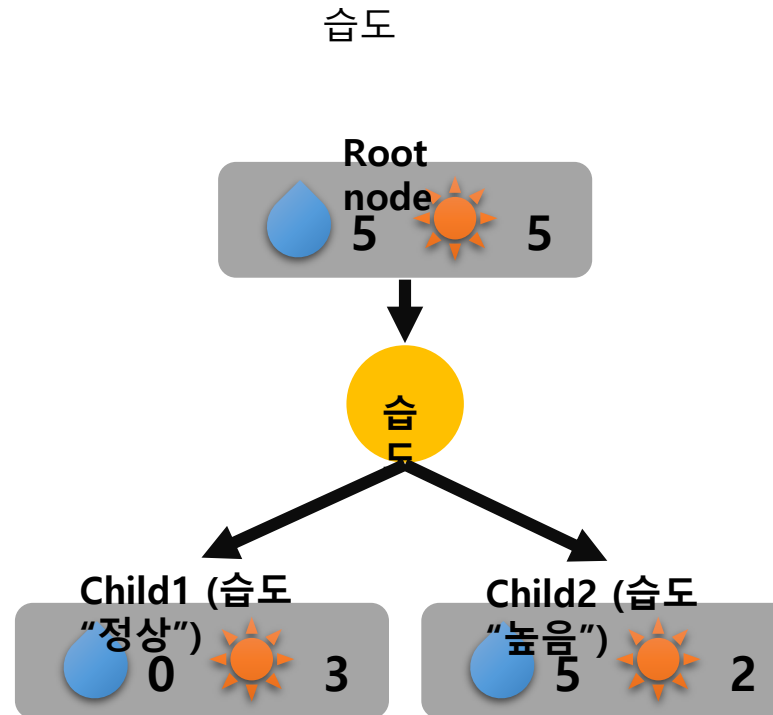
날짜	바람	습도	온도	날씨
1	약	높음	높음	비
2	강	정상	낮음	맑음
3	약	정상	낮음	맑음
4	강	높음	낮음	비
5	약	정상	높음	맑음
6	강	높음	높음	비
7	강	높음	낮음	맑음
8	약	정상	낮음	맑음
9	강	높음	높음	비
10	강	높음	높음	비



# 학습과정

## 2. 각 송근성별 자식노드의 불순도 계산

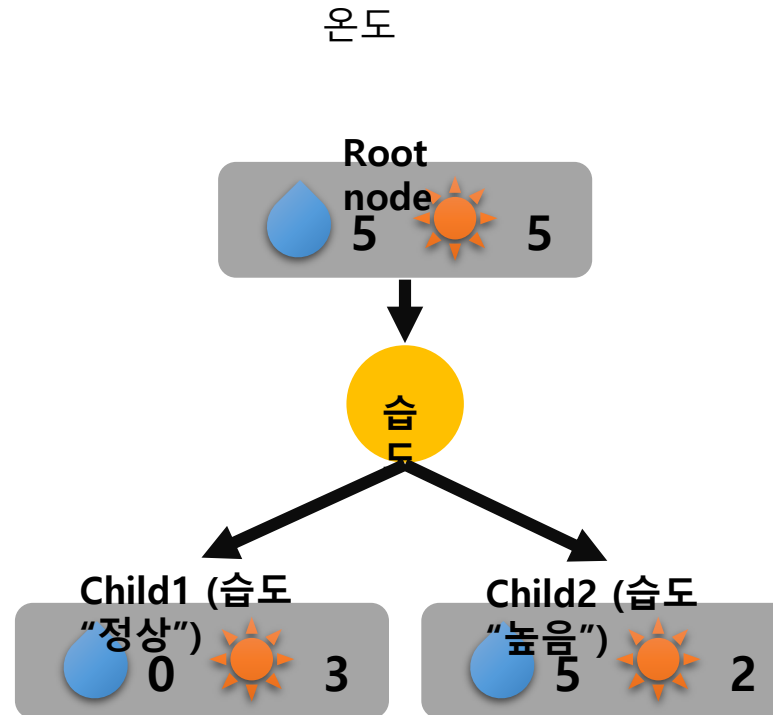
날짜	바람	습도	온도	날씨
1	약	높음	높음	비
2	강	정상	낮음	맑음
3	약	정상	낮음	맑음
4	강	높음	낮음	비
5	약	정상	높음	맑음
6	강	높음	높음	비
7	강	높음	낮음	맑음
8	약	정상	낮음	맑음
9	강	높음	높음	비
10	강	높음	높음	비



# 학습과정

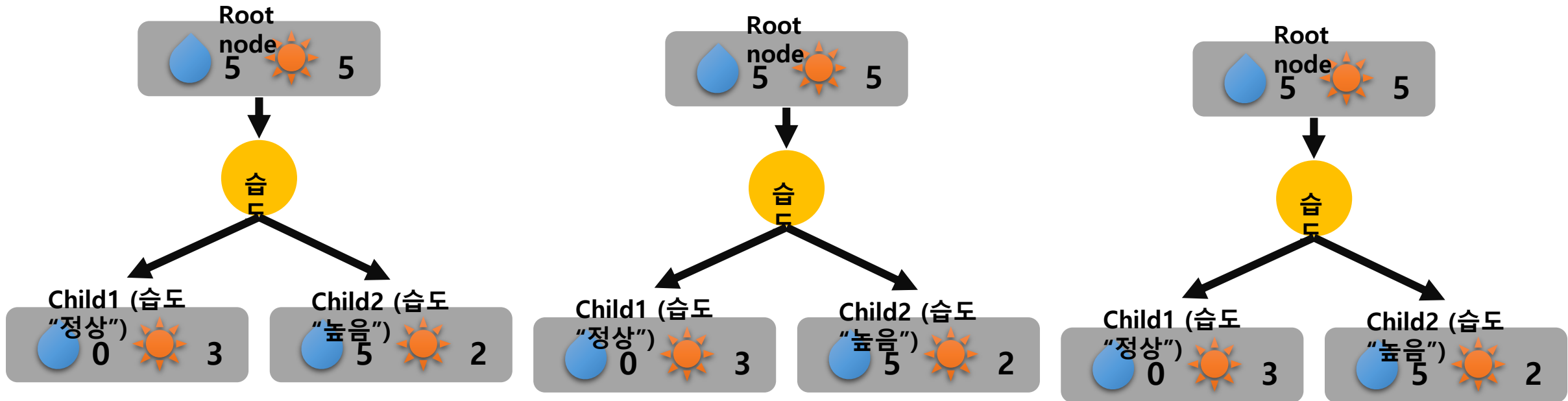
## 2. 각 송근성별 자식노드의 불순도 계산

날짜	바람	습도	온도	날씨
1	약	높음	높음	비
2	강	정상	낮음	맑음
3	약	정상	낮음	맑음
4	강	높음	낮음	비
5	약	정상	높음	맑음
6	강	높음	높음	비
7	강	높음	낮음	맑음
8	약	정상	낮음	맑음
9	강	높음	높음	비
10	강	높음	높음	비



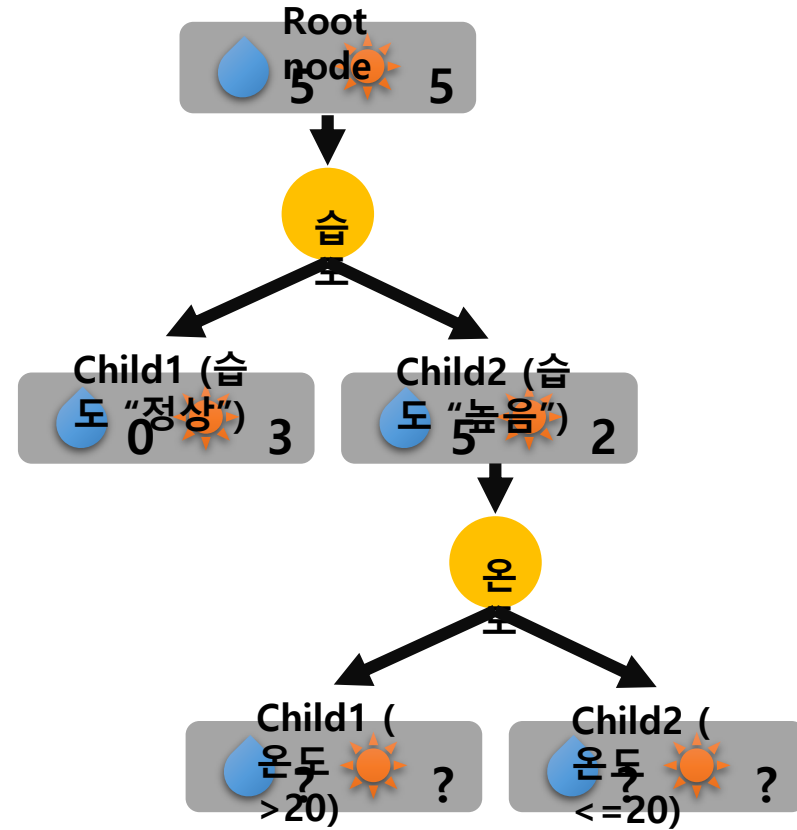
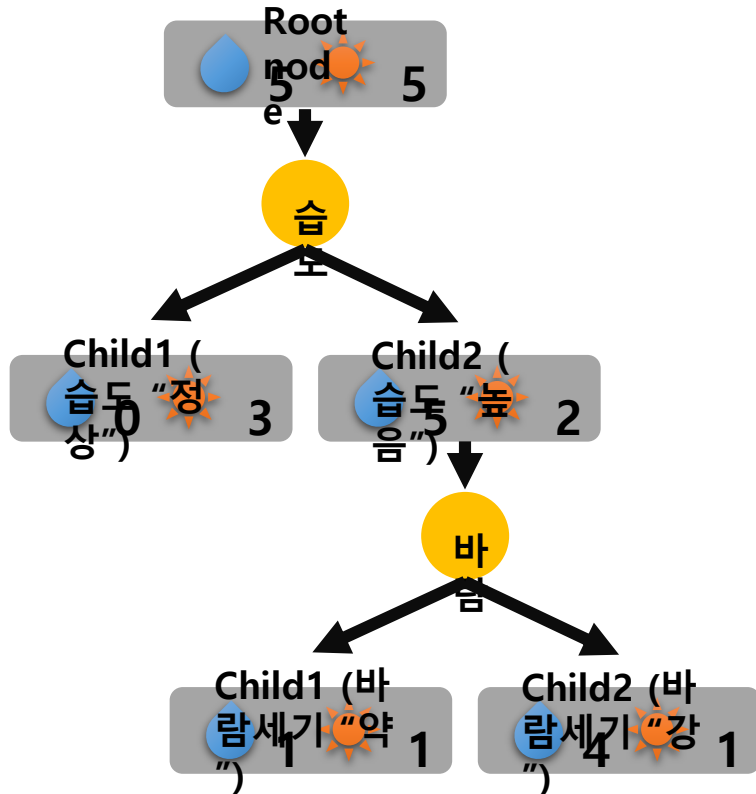
# 학습과정

3. 가장 높은 Information gain 을 가진 조건으로 분류



# 학습과정

4. 모든 leaf 노드의 불순도가 0이 될때까지 2-3 반복



# 데이터 구하기

날짜	바람	습도	온도	날씨
1	약	높음	높음	비
2	강	정상	낮음	맑음
3	약	정상	낮음	맑음
4	강	높음	낮음	비
5	약	정상	높음	맑음
6	강	높음	높음	비
7	강	높음	낮음	맑음
8	약	정상	낮음	맑음
9	강	높음	높음	비
10	강	높음	높음	비

