DECISION TREE

DECISION TREE

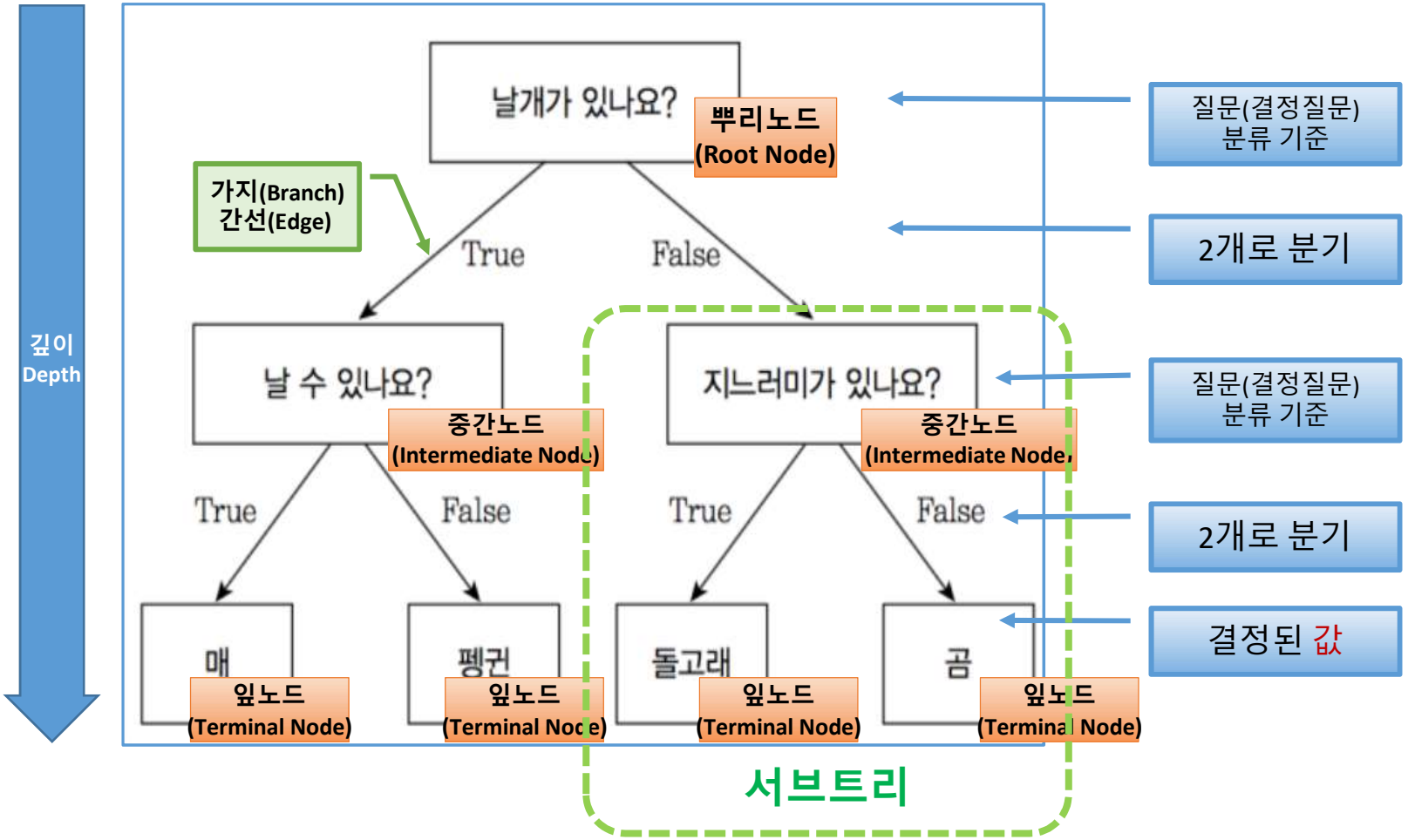
◆ 결정 트리

- **예/아니오 질문(특정 기준)**으로 **학습** 진행 ➔ 스무고개 퀴즈
- **질문(특정 기준)**을 무엇으로 하느냐가 **성능 크게 좌우**
- **질문(특정 기준)**에 따라 **데이터 구분하는 모델** ➔ 결정 트리 모델
- **직관적, 범용성, 해석력이 좋**지만 **데이터에 민감**함
- **데이터 사전 가공**에 대한 **영향이 매우 적음**
- **분류**와 **회귀 모두 가능**한 **지도 학습 모델** 중 하나

  ➔ CART(Classification And Regression Tree)라고도 함

DECISION TREE

◆ 결정 트리

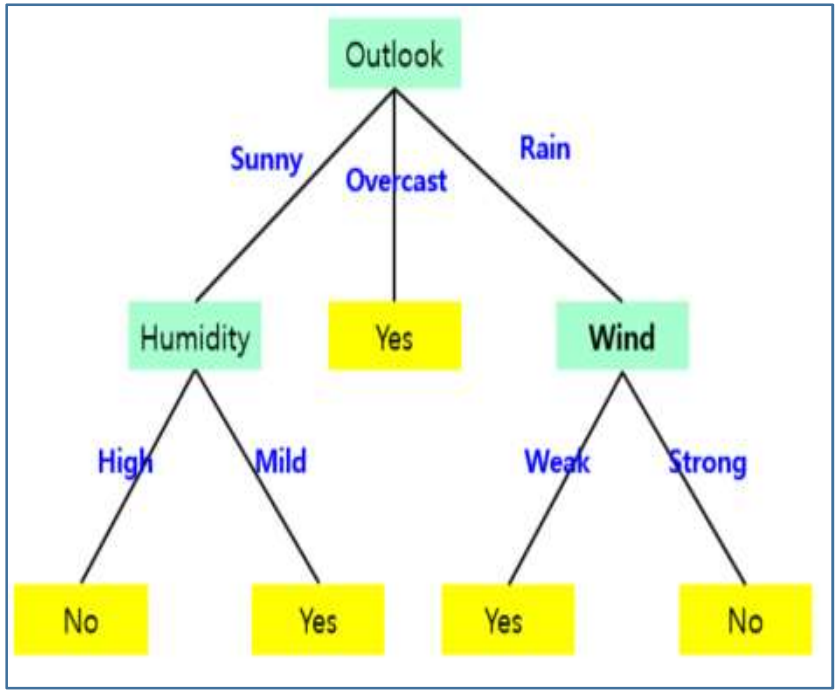


DECISION TREE

◆ 결정 트리

➢ 범주형 타입 입력 & 출력

속성/특성/피쳐					타겟
Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

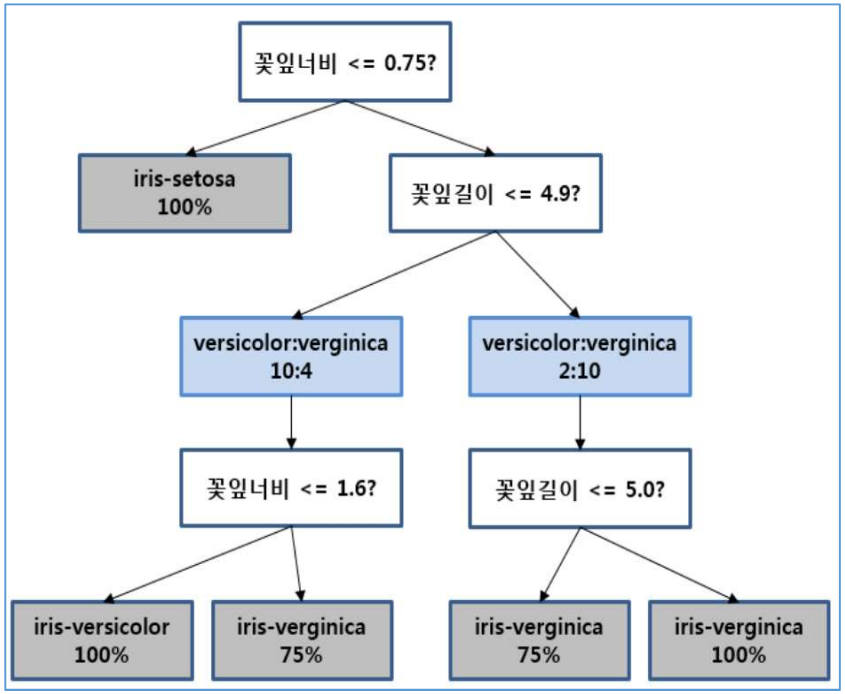


DECISION TREE

◆ 결정 트리

➢ 수치형 타입 입력&출력

	속성/특성/피쳐			타겟
SepalLength	SepalWidth	PetalLength	PetalWidth	Name
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3.0	1.4	0.2	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa
4.6	3.1	1.5	0.2	Iris-setosa
5.0	3.6	1.4	0.2	Iris-setosa
5.4	3.9	1.7	0.4	Iris-setosa
4.6	3.4	1.4	0.3	Iris-setosa
5.0	3.4	1.5	0.2	Iris-setosa
4.4	2.9	1.4	0.2	Iris-setosa
4.9	3.1	1.5	0.1	Iris-setosa
5.4	3.7	1.5	0.2	Iris-setosa
4.8	3.4	1.6	0.2	Iris-setosa
4.8	3.0	1.4	0.1	Iris-setosa



DECISION TREE

◆ 결정 트리

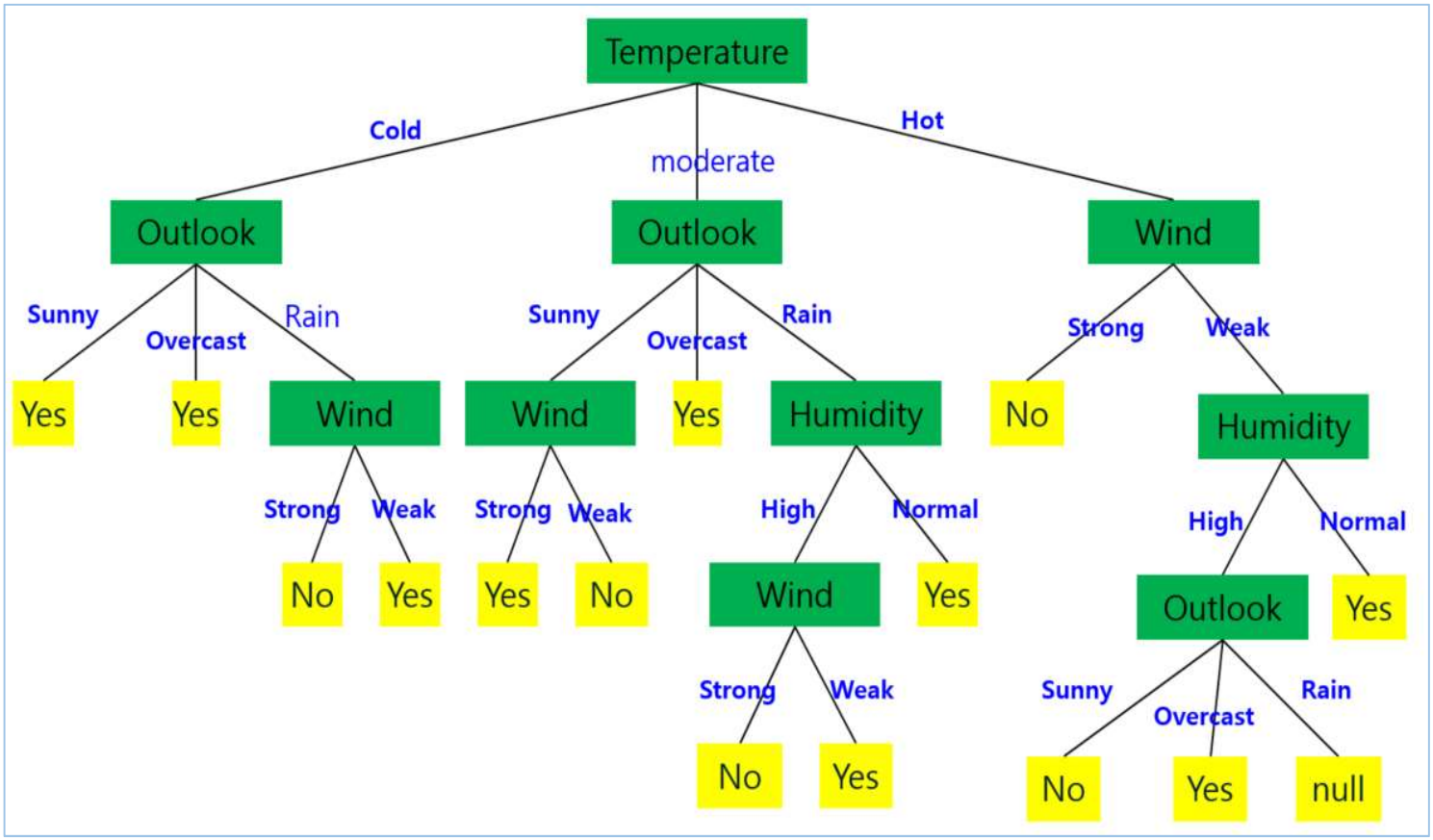
➢ 동작 알고리즘

- 모든 데이터 포함한 **하나의 노드(Root Node)로 구성된 트리에서 시작**
- **반복적 노드 분할 과정**
  - **분할 속성(splitting attribute)**을 선택
  - 속성값에 따라 **서브 트리(subtree)를 생성**
  - 데이터를 속성값에 따라 **분배**

DECISION TREE

◆ 결정 트리

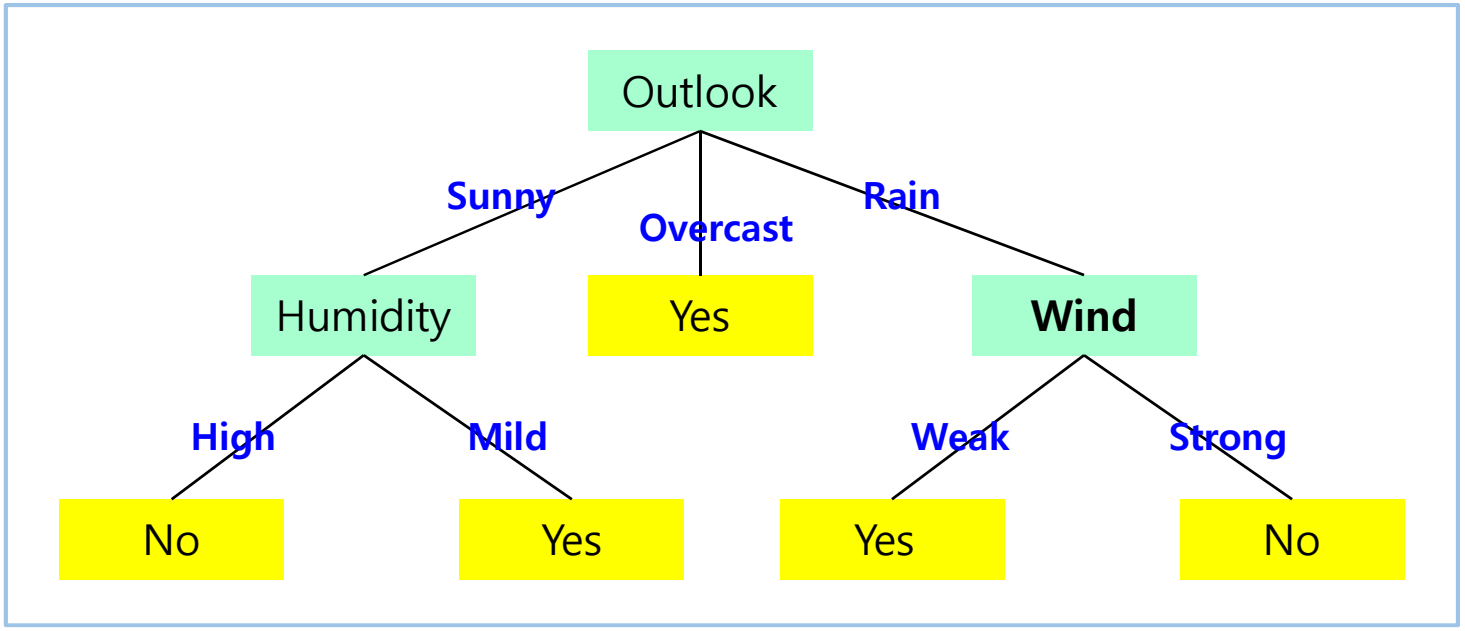
➢ 동일 문제 - 분할 속성에 따른 복잡한 트리



DECISION TREE

◆ 결정 트리

➢ 동일 문제 - 분할 속성에 따른 간단한 트리



DECISION TREE

◆ 결정 트리

➢ 분할 속성(Spliting Attribute) 결정

- 결정 트리 구성 및 성능에 가장 큰 영향
- 속성 결정 기준

  ➔ 분할 후 **가능한 많은 동일 분류의 데이터가 모이는 속성**

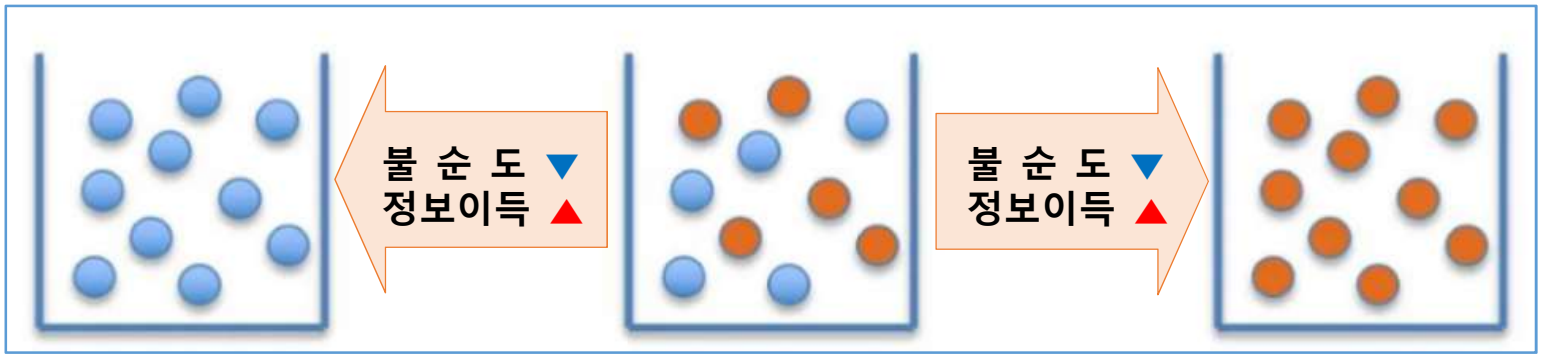
  ➔ 분할 후 동일 분류 데이터 모이는 정도 측정 필요

DECISION TREE

◆ 결정 트리

➢ 분할 속성(Spliting Attribute) 결정

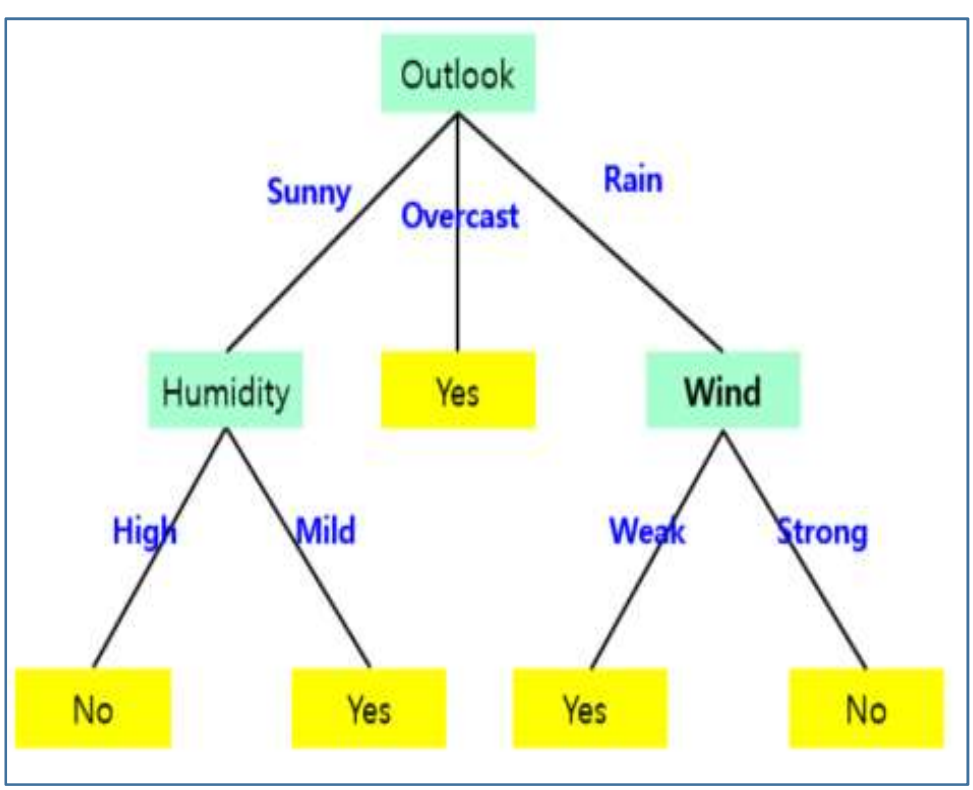
- **고려 사항**
  - **불순도(Impurity)** : 서로 다른 데이터가 얼마나 섞여 있는지 의미
  - **정보 이득(IG)** : 분할 후 불순도의 차이

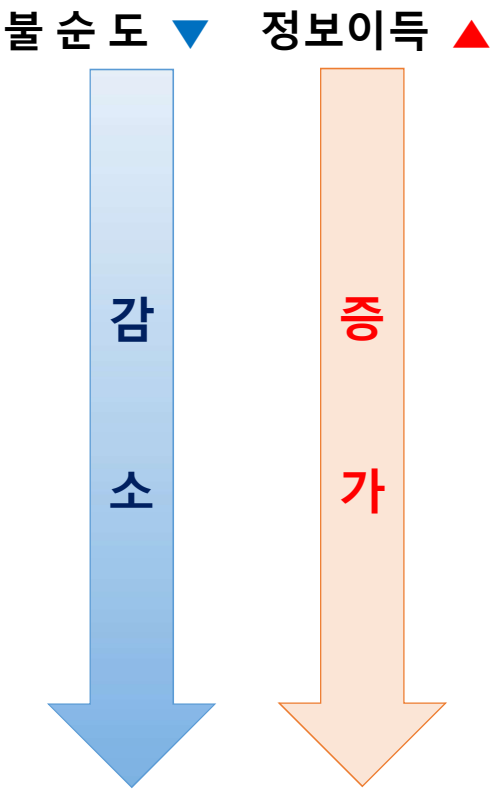


DECISION TREE

◆ 결정 트리

➢ 분할 속성(Spliting Attribute) 결정





DECISION TREE

◆ 결정 트리

➢ 불순도(Impurity) 수치화 - 엔트로피(Entropy)

정보이득(IG) 측정값으로 불순도를 수치화

값의 범위 : 0 ~ log(nc) **0에 가까울수록 좋음**

- 9 □ (사각형)
- 5 Δ (삼각형)
- 분류별 확률(Class Probability)

$$p(\square) = \frac{9}{14} \quad p(\Delta) = \frac{5}{14}$$

- 엔트로피 $I = -\sum_{c} p(c) \log_2 p(c)$

$$I = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.940 \text{ bits}$$

DECISION TREE

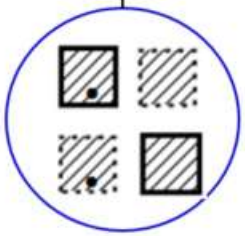
◆ 결정 트리

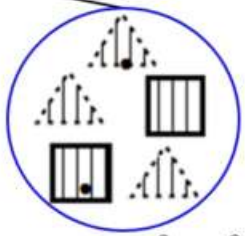
➢ 불순도(Impurity) 수치화 - 엔트로피(Entropy)

$$I = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.940$$

Pattern

수평 | 대각선 | 수직





$$I_{horizontal} = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.971 \quad I_{diagonal} = -\frac{0}{4}\log_2\frac{0}{4} - \frac{4}{4}\log_2\frac{4}{4} = 0.0 \quad I_{vertical} = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.971$$

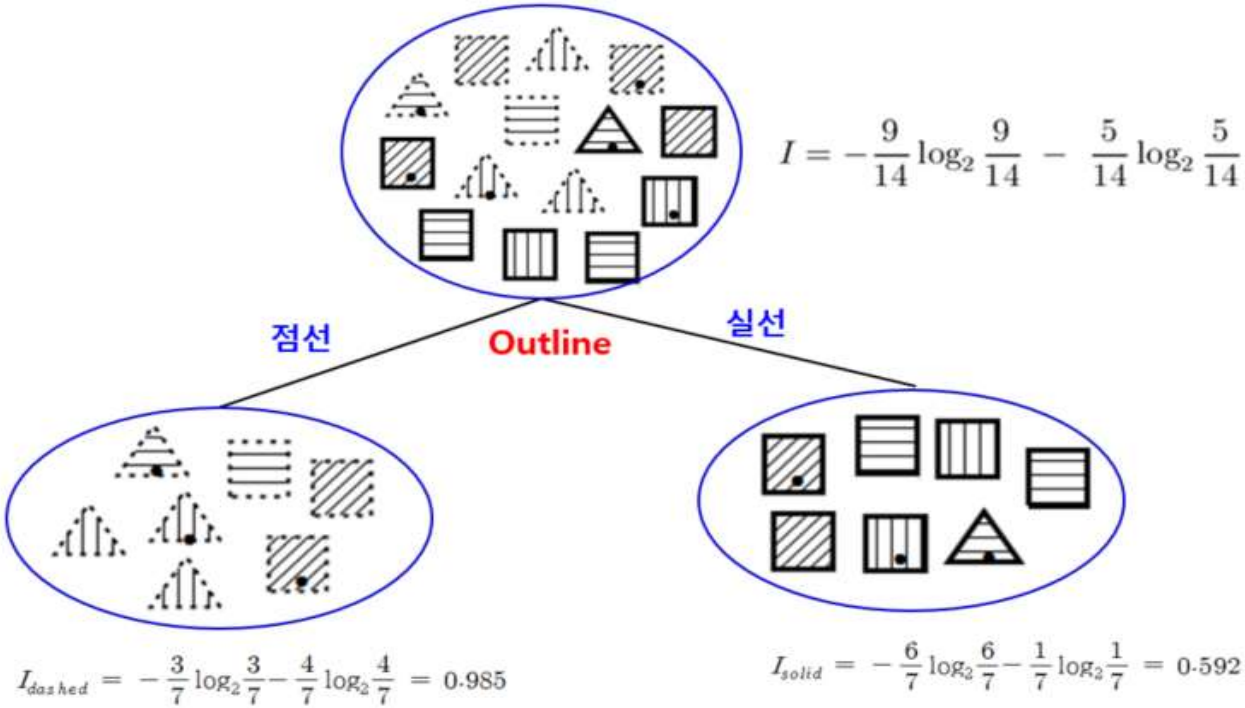
$$I_{res}(Pattern) = \sum p(v)I(v) = \frac{5}{14}\cdot 0.971 + \frac{4}{14}\cdot 0 + \frac{5}{14}\cdot 0.971 = 0.694$$

$$IG(Pattern) = I - I_{res}(Pattern) = 0.940 - 0.694 = 0.246$$

DECISION TREE

◆ 결정 트리

➢ 불순도(Impurity) 수치화 - 엔트로피(Entropy)



$$I = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.940$$

$$I_{dashed} = -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7} = 0.985$$

$$I_{solid} = -\frac{6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7} = 0.592$$

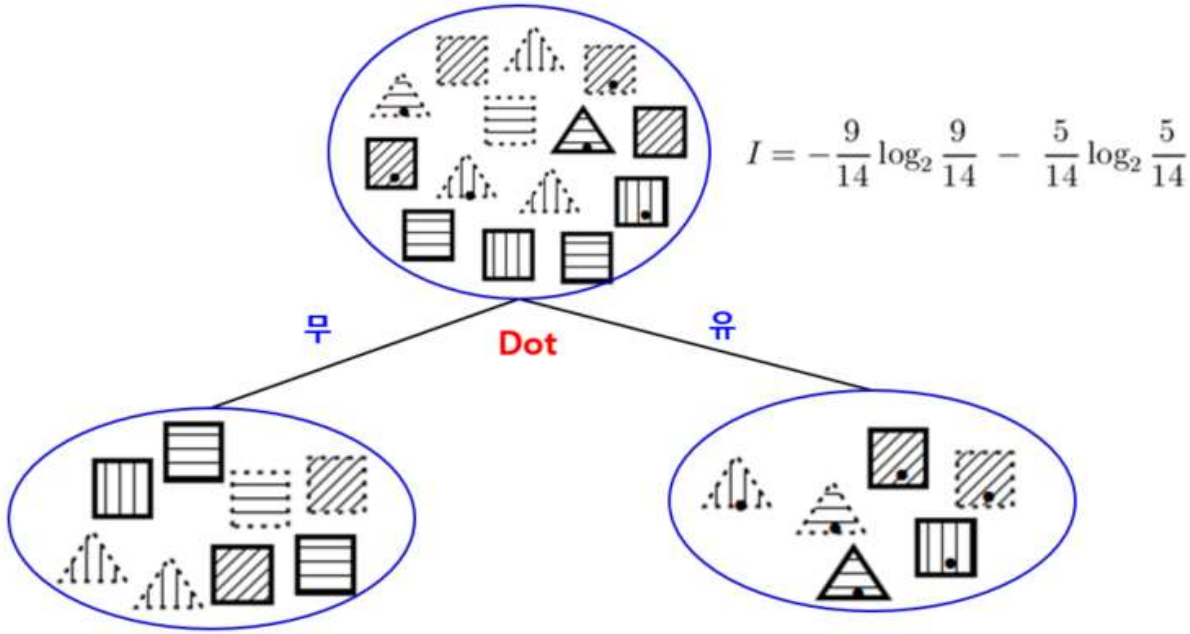
$$I_{res}(Outline) = \sum p(v)I(v) = \frac{7}{14} \cdot 0.985 + \frac{7}{14} \cdot 0.592 = 0.789$$

$$IG(Outline) = I - I_{res}(Outline) = 0.940 - 0.789 = 0.151$$

DECISION TREE

◆ 결정 트리

➢ 불순도(Impurity) 수치화 - 엔트로피(Entropy)



$$I = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.940$$

$$I_{no} = -\frac{2}{8}\log_2\frac{2}{8} - \frac{6}{8}\log_2\frac{6}{8} = 0.811$$

$$I_{yes} = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} = 1.000$$

$$I_{res}(Dot) = \sum p(v)I(v) = \frac{8}{14}\cdot 0.811 + \frac{6}{14}\cdot 1.000 = 0.892$$

$$IG(Dot) = I - I_{res}(Dot) = 0.940 - 0.892 = 0.048$$

DECISION TREE

◆ 결정 트리

➢ 불순도(Impurity) 수치화 - 엔트로피(Entropy)

- 속성별 정보 이득

  IG(Pattern) = **0.246**

  IG(Outline) = 0.151

  IG(Dot) = 0.048

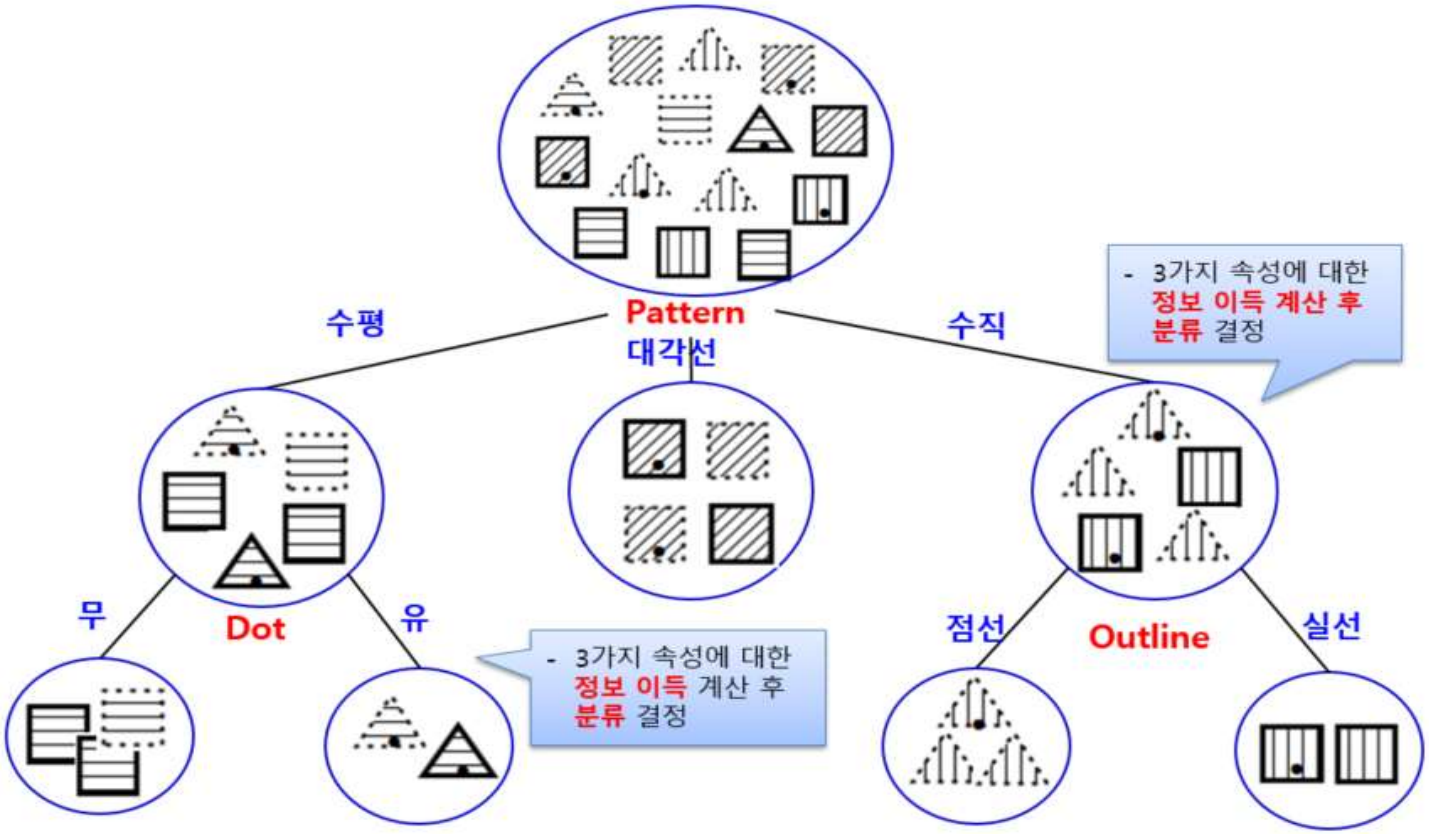
분할속성 선택

정보이득이 큰 것 선택 **"Pattern" 선택**

DECISION TREE

◆ 결정 트리

➢ 불순도(Impurity) 수치화 - 엔트로피(Entropy)



DECISION TREE

◆ 결정 트리

➢ 불순도(Impurity) 수치화 - 엔트로피(Entropy)

Entropy = 1

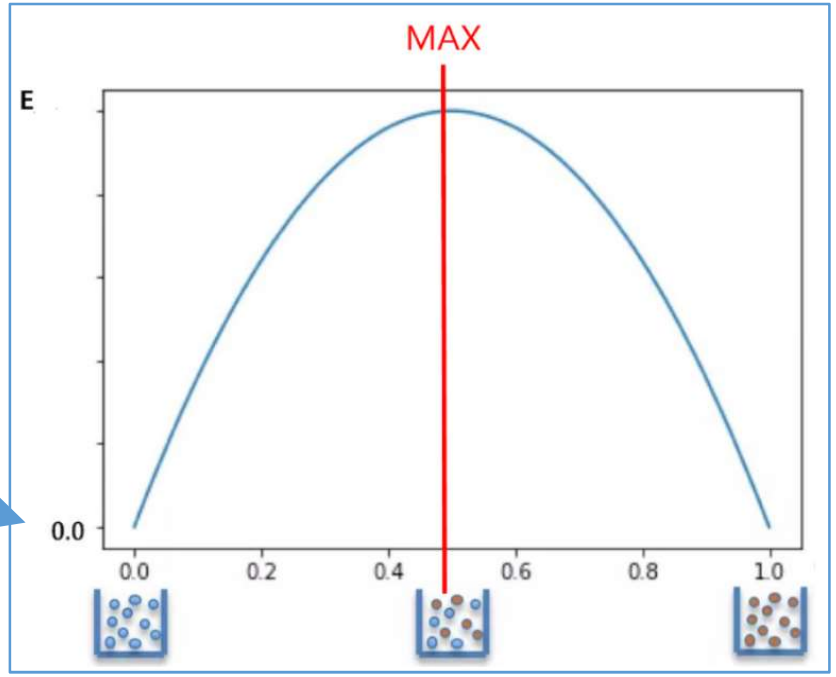
반반 있는 상태

분류 어려움

Entropy = 0

한 종류만 존재

분류하기 쉬움



DECISION TREE

◆ 결정 트리

➢ 불순도(Impurity) 수치화 - 지니(Gini) 계수

데이터의 통계적 분산정도 즉 균일도로 불순도를 수치화 함

$$Gini = \sum_{i \neq j} p(i)p(j)$$

$$p(\square) = \frac{9}{14}$$

$$p(\triangle) = \frac{5}{14}$$

$$Gini = \frac{9}{14} \otimes \frac{5}{14} = 0.230$$

DECISION TREE

◆ 결정 트리

➢ 가지치기(Pruning)

- **과대적합(Overfitting)을 방지**하기 위한 방법
- 트리에 **가지가 지나치게 많을 때** 나타남
- 최대 깊이나 Terminal Node의 최대 개수 제한
  - ➢ **min_sample_split :** 한 노드에 최소 데이터 수
    최소 데이터 수 아래로 분할하지 않음
  - ➢ **max_depth :** 최대 깊이 조정

DECISION TREE

◆ 결정 트리

➢ Scikit-Learn LIB

sklearn.tree.**DecisionTreeClassifier**

(*,

criterion='gini'	**: 분할 품질을 측정하는 기능**
splitter='best'	: 각 노드 분할 선택 방법 설정
max_depth=None	: 트리 최대 깊이 (값이 클수록 모델 복잡도 ▲)
min_samples_split=2	: 자식 노드 분할 위한 최소 샘플 수
min_samples_leaf=1	: 리프 노드에 있어야 할 최소 샘플 수
min_weight_fraction_leaf=0.0	: 가중치가 부여된 샘플 수에서의 비율
max_features=None	: 각 노드 분할에 사용할 특징 최대 수

DECISION TREE

◆ 결정 트리

➢ Scikit-Learn LIB

```
sklearn.tree.DecisionTreeClassifier
(
        random_state=None                : 난수 seed 설정
        max_leaf_nodes=None              : 리프 노드의 최대수
        min_impurity_decrease=0.0        : 최소 불순도
        class_weight=None                : 클래스 가중치
        ccp_alpha=0.0
)
```