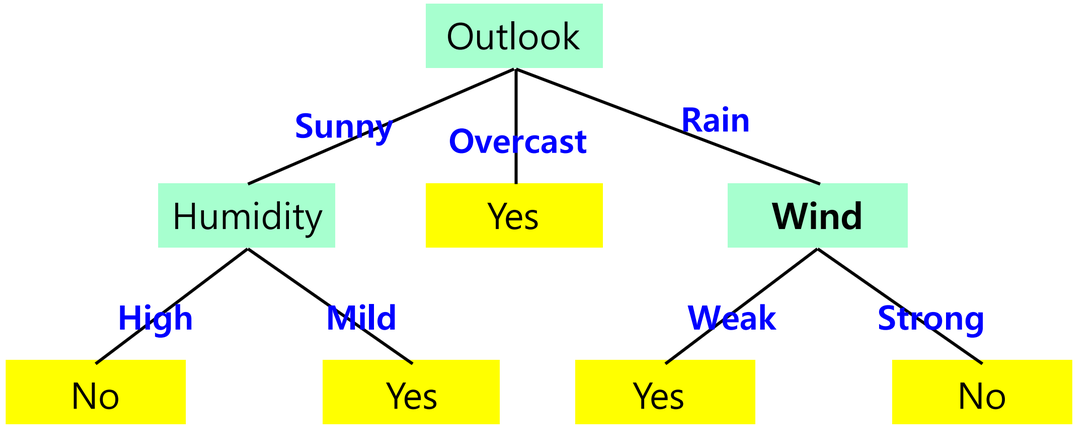
Decision Tree

Jin Hyun Kim

https://images.app.goo.gl/QJ7vz7xo9wj6kNCt7

Decision Tree 결정나무

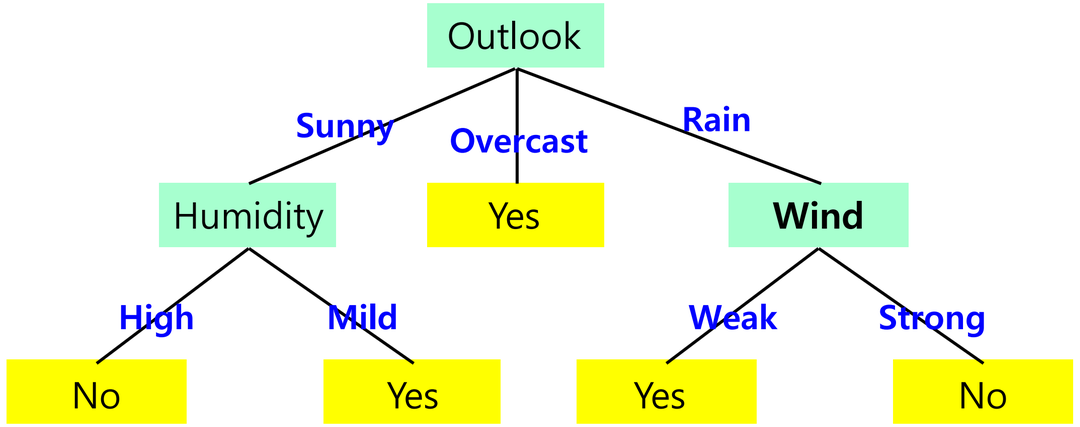
- 결정나무(decision tree)
  - 트리 형태로 의사결정 지식을 표현한 것
    - 내부 노드(internal node) : 비교 속성
    - 간선(edge) : 속성 값
    - 단말 노드(terminal node) : 부류(class), 대표값



IF Outlook = Sunny **AND** Humidity = High **THEN** Answer = No

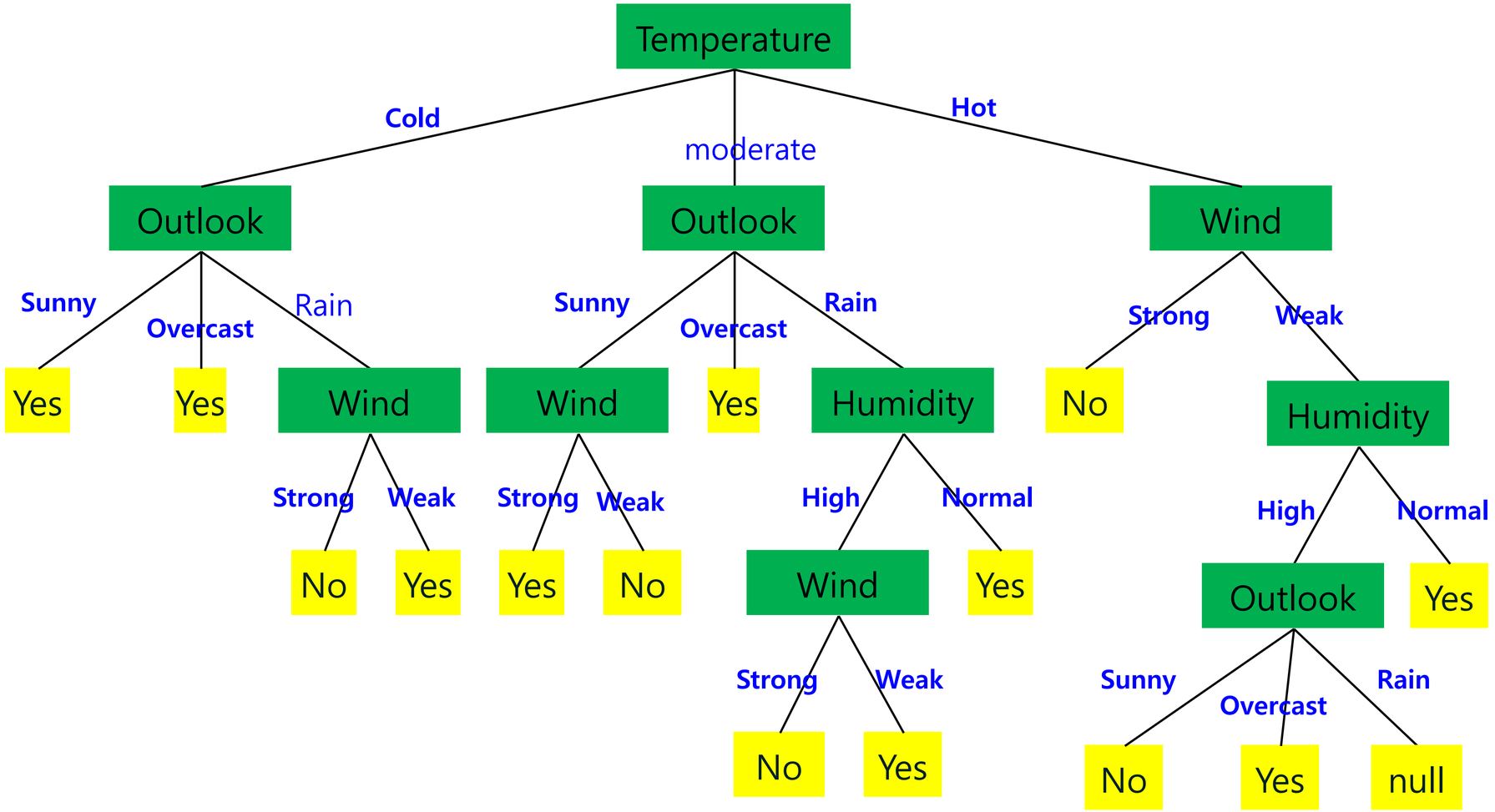
Decision Tree 결정나무

Day 날짜	Outlook 조망	Temperature 기온	Humidity 습도	Wind 바람	PlayTennis 테니스 여부
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No



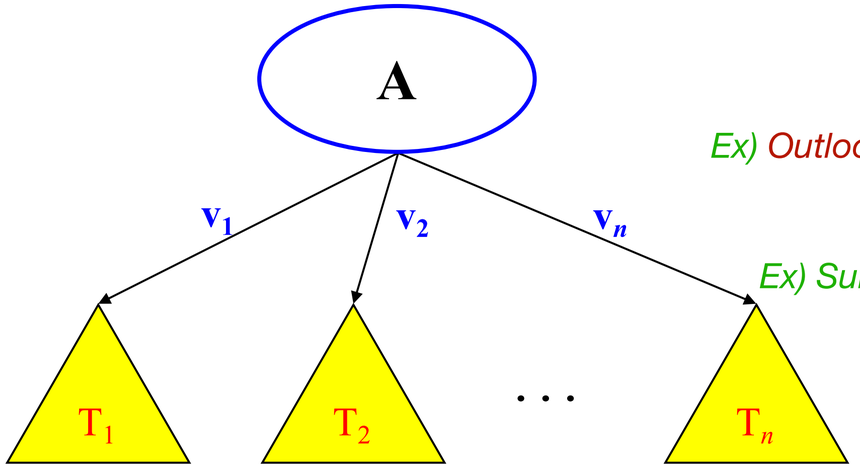
Outlook 조망	**Temperature** 기온	**Humidity** 습도	**Wind** 바람	**PlayTennis** 테니스 여부
Sunny	Hot	Mild	Weak	?
Rain	Hot	High	Weak	?

Decision Tree 결정나무



Decision Tree Learning Algorithm

- 모든 데이터를 포함한 하나의 노드로 구성된 트리에서 시작
- 반복적인 노드 분할 과정
  1. 분할 속성(spliting attribute)을 선택
  2. 속성값에 따라 서브트리(subtree)를 생성
  3. 데이터를 속성값에 따라 분배



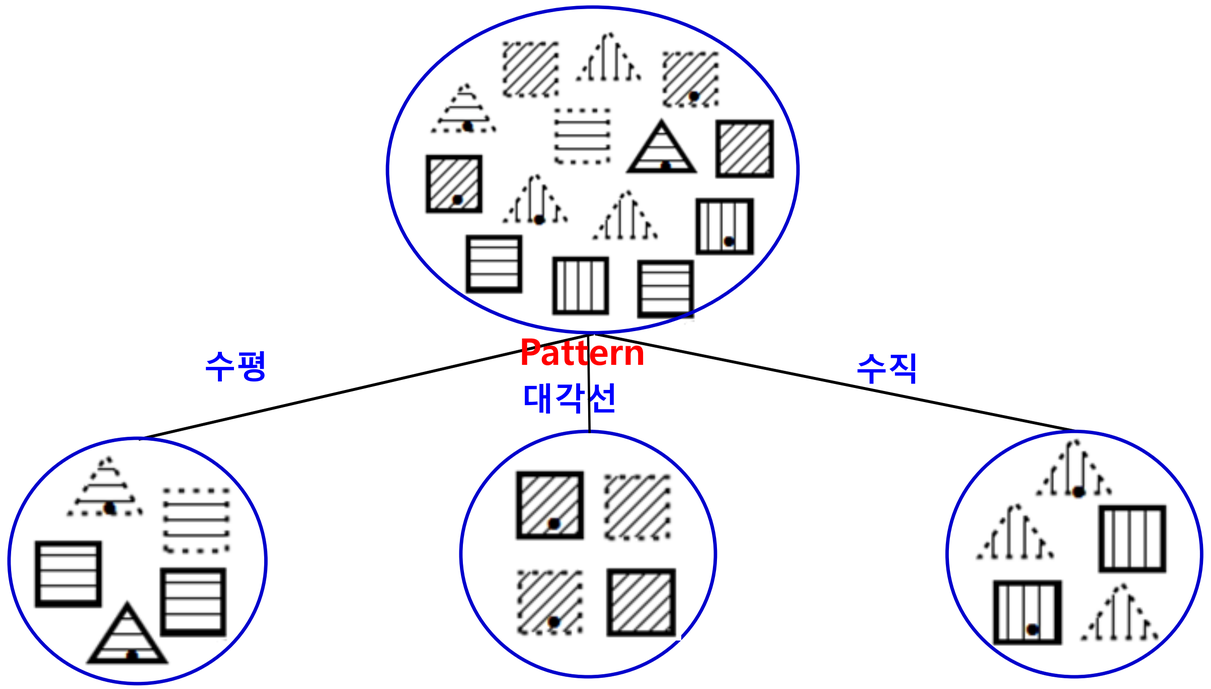
분할속성 (splitting attribute)

- 분할속성 (splitting attribute) 결정
  - 누가 Top node 의 분할 속성인가?
  - 어떤 속성을 선택하는 것이 효율적인가?
    - 분할 경과가 가능하면 동질인 것으로 만드는 속성을 선택, 즉 동질인 정도를 측정

Day 날짜	Outlook 조망	Temperature 기온	Humidity 습도	Wind 바람	PlayTennis 테니스 여부

Decision Tree Learning Algorithm

- 학습의 목표
  - 순도(homogeneity)가 증가
  - 불순도(impurity) 혹은 불확실성 (uncertainty)이 최대한 감소

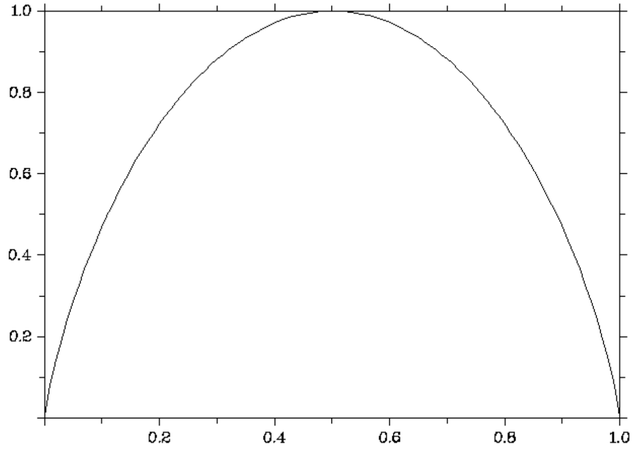


엔트로피 (Entropy)

- 동질인 정도 측정 가능 척도 (Impurity)
- 원래 정보량 (amount of information) 측정 목적

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

  - $p(c)$: c 부류에 속하는 것의 비율



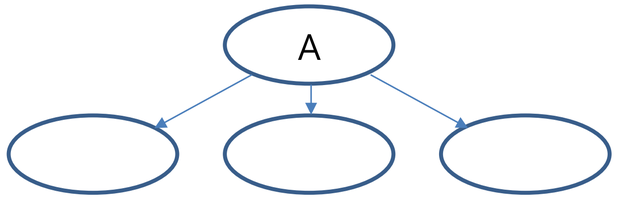
2개의 부류가 고르게 분포되어 있는 경우의 엔트로피

ID3 - Decision Tree Learning Algorithm

- 분할 속성을 결정할 때, 엔트로피(I) 를 이용하여 계산되는, information gain 정보 이득 (IG)을 사용
- 가중평균 (I_{res}) - 특정 속성으로 분할한 후의 각 부분집합의 정보량의 가중평균

$$I_{res} = -\sum_{v} p(v) \sum_{c} p(c|v) \log_2 p(c|v)$$

$$IG = I - I_{res}(A) = -\sum_{c} p(c) \log_2 p(c) + \sum_{v} p(v) \sum_{c} p(c|v) \log_2 p(c|v)$$



- 정보이득이 클 수록 우수한 분할 속성

분할속성 방법

부류(class) 정보가 있는 데이터

	속성			부류
	Pattern	*Outline*	*Dot*	*Shape*
1	수직	점선	무	삼각형
2	수직	점선	유	삼각형
3	대각선	점선	무	사각형
4	수평	점선	무	사각형
5	수평	실선	무	사각형
6	수평	실선	유	삼각형
7	수직	실선	무	사각형
8	수직	점선	무	삼각형
9	대각선	실선	유	사각형
10	수평	실선	무	사각형
11	수직	실선	유	사각형
12	대각선	점선	유	사각형
13	대각선	실선	무	사각형
14	수평	점선	유	삼각형

분할속성- Which of Pattern, Outline, Dot is the best splitting attribute?

분할속성 방법

엔트로피 계산

$$I = -\sum_c p(c) \log_2 p(c)$$

$$I_{res} = -\sum_v p(v) \sum_c p(c|v) \log_2 p(c|v)$$

$$IG = I - I_{res}(A) = -\sum_c p(c)\log_2 p(c) + \sum_v p(v) \sum_c p(c|v)\log_2 p(c|v)$$

- 부류 - 9 □ (사각형) / 5 △ (삼각형)
- 부류별 확률(class probability)

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

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

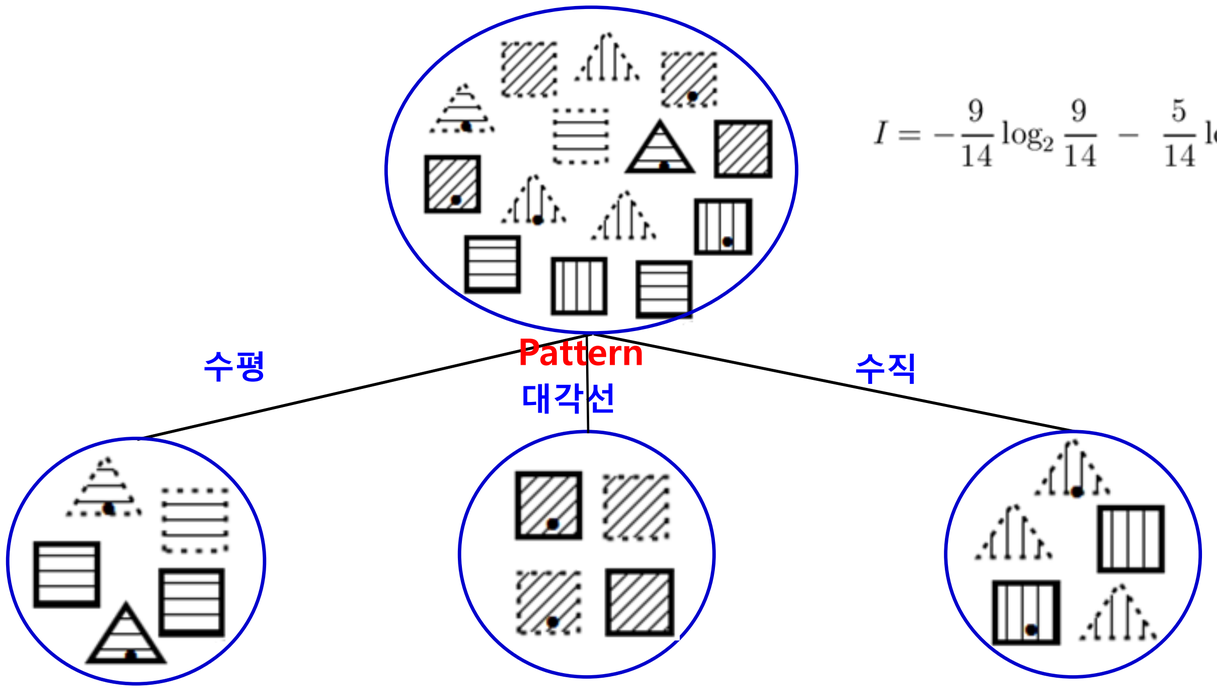
- 엔트로피(entropy)

$$I = -\sum_c p(c) \log_2 p(c)$$

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

결정트리 학습 알고리즘
정보이득

- 데이터 집합 분할과 정보이득 - Pattern 기분 분할



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

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

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

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

결정트리 학습 알고리즘

정보이득비

- 정보이득비 - 속성값이 많을 수록 불이익을 받도록 정보이득 측도를 개선

$$GainRatio(A) = \frac{IG(A)}{I(A)} = \frac{I - I_{res}(A)}{I(A)}$$

- *I*(*A*) 는 속성 A의 속성값을 부류인 것처럼 간주, 속성값의 개수가 많아 질 수록 값이 커짐

$$I(A) = -\sum_{v} p(v) log_2 p(v)$$

결정트리 학습 알고리즘

정보이득비

- *I*(*Pattern*) - 속성 Pattern에 대한 엔트로피

$$I(Pattern) = -\frac{horizental}{\square+\triangle}log_2\frac{horizental}{\square+\triangle} - \frac{vertical}{\square+\triangle}log_2\frac{vertical}{\square+\triangle} - \frac{diagonal}{\square+\triangle}log_2\frac{diagonal}{\square+\triangle}$$

$$= -\frac{5}{14}log_2\frac{5}{14} - \frac{5}{14}log_2\frac{5}{14} - \frac{4}{14}log_2\frac{4}{14} = 1.58$$

결정트리 학습 알고리즘

정보이득비

- 속성 Pattern 정보이득비

$$GainRatio(Pattern) = \frac{IG(A)}{I(Pattern)} = \frac{I - I_{res}(PatternA)}{I(Pattern)} = \frac{0.94 - 0.694}{1.58}$$

정보이득 vs 정보이득 비

속성	속성의 개수	정보이득	정보이득비
Pattern	3	0.247	0.156
Outline	2	0.152	0.152
Dot	2	0.048	0.049

지니 지수 (Gini Index)

- 데이터 집합에 대한 지니 값 - i, j 가 부류 (class) 를 나타낼 때

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

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

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

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

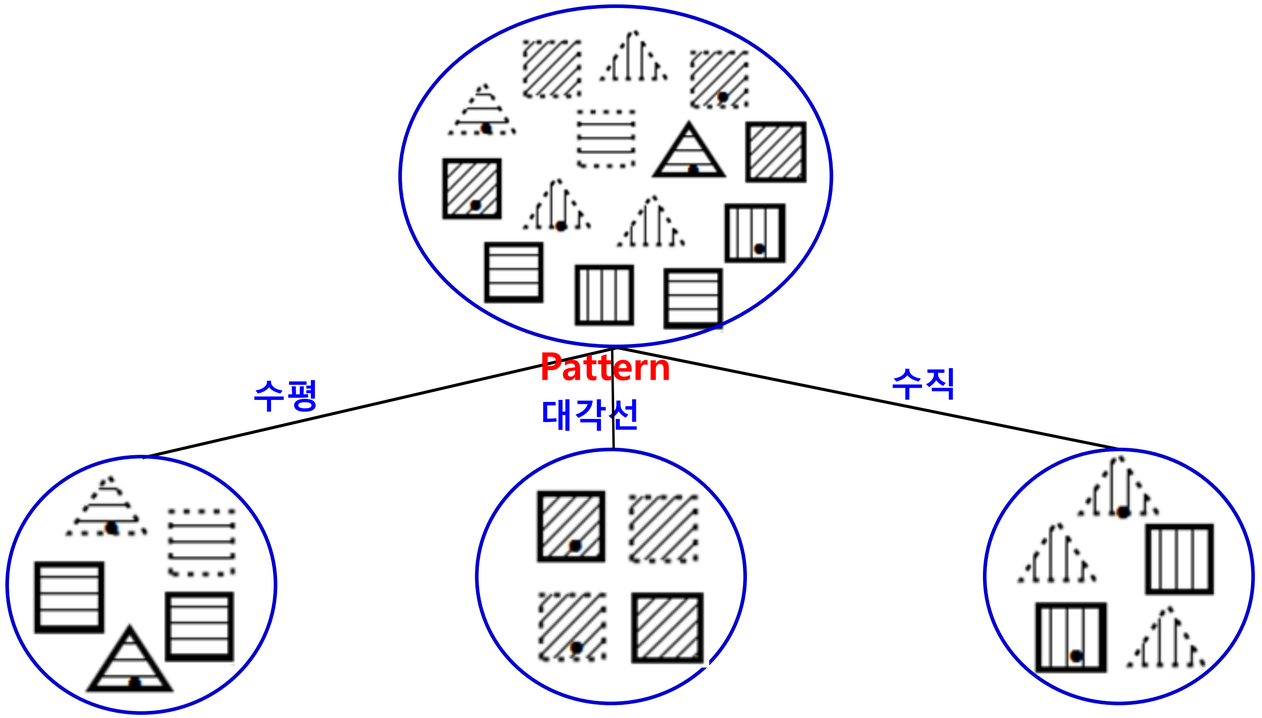
- 속성 A에 대한 지니 지수 값 가중 평균

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

- 지니 지수 이득 (Gini index gain)

$$GiniGain(A) = Gini - Gini(A)$$

지니 지수 (Gini Index)



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

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

$$Gini(Pattern) = \frac{5}{14} \times (\frac{3}{5} \times \frac{2}{5}) + \frac{5}{14} \times (\frac{2}{5} \times \frac{3}{5}) + \frac{4}{14} \times (\frac{4}{4} \times \frac{0}{4}) = 0.171$$

$$Gini\,Gain(Pattern) = 0.230 - 0.171 = 0.058$$

분할속성 평가 척도 비교

속성	정보이득	정보이득비	지니이득
Pattern	0.247	0.156	0.058
Outline	0.152	0.152	0.046
Dot	0.048	0.049	0.015

결정트리 학습 알고리즘

- 결정트리 알고리즘
  - ID3 알고리즘
    - 범주형(categorical) 속성값을 갖는 데이터에 대한 결정트리 학습
    - 예. PlayTennis, 삼각형/사각형 문제
  - C4.5 알고리즘
    - 범주형 속성값과 수치형 속성값을 갖는 데이터로 부터 결정트리 학습
    - ID3를 개선한 알고리즘
- C5.0 알고리즘
  - C4.5를 개선한 알고리즘
- CART 알고리즘
  - 수치형 속성을 갖는 데이터에 대해 적용

결정트리를 이용한 회귀

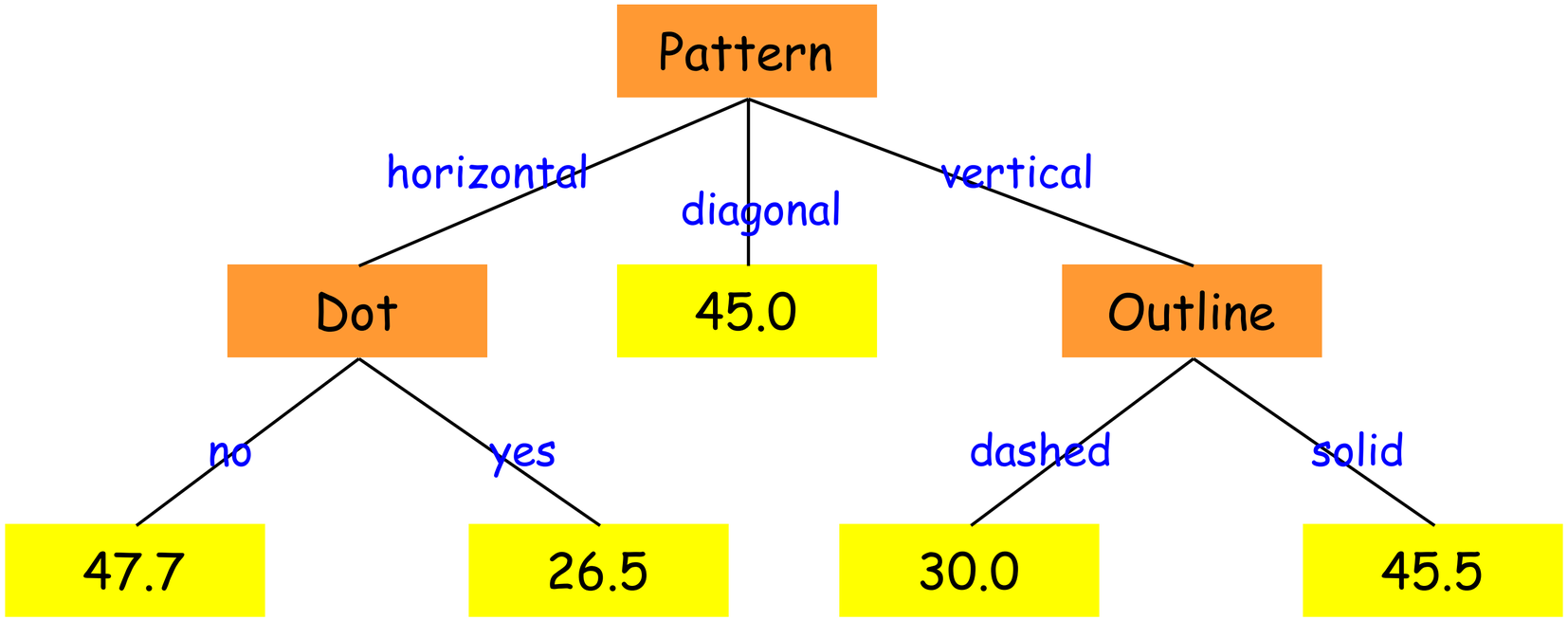
회귀(regression)를 위한 결정트리

- 출력값이 수치값

	속성			면적
	Pattern	*Outline*	*Dot*	*Area*
1	수직	점선	무	25
2	수직	점선	유	30
3	대각선	점선	무	46
4	수평	점선	무	45
5	수평	실선	무	52
6	수평	실선	유	23
7	수직	실선	무	43
8	수직	점선	무	35
9	대각선	실선	유	38
10	수평	실선	무	46
11	수직	실선	유	48
12	대각선	점선	유	52
13	대각선	실선	무	44
14	수평	점선	유	30

결정트리를 이용한 회귀

회귀(regression)를 위한 결정트리



결정트리를 이용한 회귀

회귀 (regression)를 위한 결정트리

- 분류를 위한 결정트리와 차이점
  - 단말노드가 부류(class)가 아닌 수치값(numerical value)임
  - 해당 조건을 만족하는 것들이 가지는 대표값
- 분할 속성 선택
  - 표준편차 축소(reduction of standard deviation) SDR를 최대로 하는 속성 선택

$$SDR\,(A) = SD - SD\,(A)$$

- **표준편차** $SD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - m)^2}$ **,** m**: 평균**
  - *SD(A)* - 속성 *A*를 기준으로 분할 후의 부분 집합별 표준표차의 가중평균

결정트리를 이용한 회귀

회귀(regression)를 위한 결정트리

Area
26
30
48
46
62
23
43
36
38
48
48
62
44
30

$SD = 9.67$

		Area의 표준편차	개수
Pattern	수평	12.15	5
	수직	9.36	5
	대각선	5.77	4
			14

$$SD(Pattern) = \frac{5}{14} \times 12.15 + \frac{5}{14} \times 9.36 + \frac{4}{14} \times 5.77 = 9.05$$

$$SDR(Pattern) = SD - SD(Pattern) = 9.67 - 9.05 = 0.61$$

결론

- 결정나무 분류
  - 엔트로피 - 집단 내의 클래스 분포의 정도 - 를 이용한 기계학습
  - 분류기준을 정하는 것이 핵심
    - 정보이득, 정보이득비, 지니 이득
  - 선형회귀에도 활용 가능