

규칙 결정 방법

- 규칙을 정하는 방법은 정보노드와 자식노드 간의 엔트로피를 가장 낮게 만드는 최상의 독립변수와 기준값을 찾는 것.

→ 이러한 기준을 정량화한 것이, 정보 획득량(Information Gain)이다.

→ 기본 아이디어, 모든 독립변수와 모든 가능한 기준값에 대해 정보 획득량을

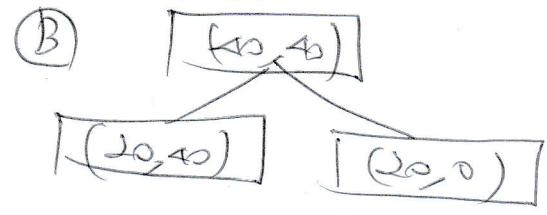
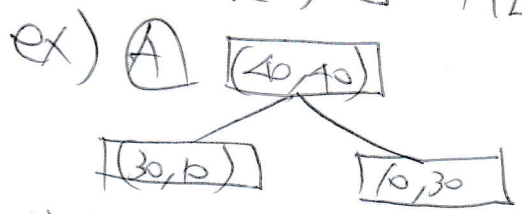
구하여 "가장 정보 획득량이 큰 독립변수와 기준값을 선택"

정보 획득량 (Information Gain)

→ Definition: X라는 조건에 의해 확률 변수 Y의 엔트로피가 얼마나 감소하였는가를 나타내는 값.

따라서 Y의 엔트로피에서 X에 대한 Y의 조건부 엔트로피를 빼 값을 얻어

$$IG[Y, X] = H[Y] - H[Y|X]$$



i) 정보노드 엔트로피 계산.

$$H[Y] = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} = \frac{1}{2} + \frac{1}{2} = 1$$

ii) A의 IG (Information Gain)

a) $H[Y|X=X_1] = -\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} = 0.81$

b) $H[Y|X=X_2] = -\frac{1}{4} \log_2 \frac{1}{4} - \frac{3}{4} \log_2 \frac{3}{4} = 0.81$

$$\Rightarrow H[Y|X] = \frac{1}{2} H[Y|X=X_1] + \frac{1}{2} H[Y|X=X_2] = 0.81$$

$$\Rightarrow IG = H[Y] - H[Y|X] = 1 - 0.81 = 0.19$$

iii) B의 IG (Information Gain)

a) $H[Y|X=X_1] = -\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3} = 0.92$

b) $H[Y|X=X_2] = 0$

$$\Rightarrow H[Y|X] = \frac{3}{4} \times H[Y|X=X_1] + \frac{1}{4} H[Y|X=X_2] = 0.69$$

$$\Rightarrow IG = H[Y] - H[Y|X] = 1 - 0.69 = 0.31$$

→ 결론 B 방법의 IG가 높으니, B 방법을 써 4은 방법