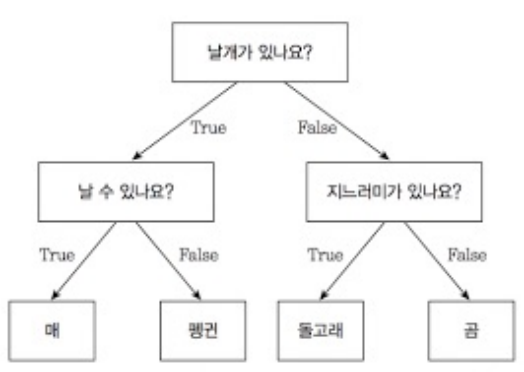
Decision Tree



의사결정 나무는
데이터를 분석하여
이들 사이에 존재하는
패턴을 예측 가능한
규칙들의 조합으로
나타내며, 그 모양이
'나무'와 같다고해서
의사결정 나무라고 불립니다.

좋은 의사결정 트리를 만들기 위해서는
좋은 기준을 잡아야 한다.
How? 불순도를 통해서

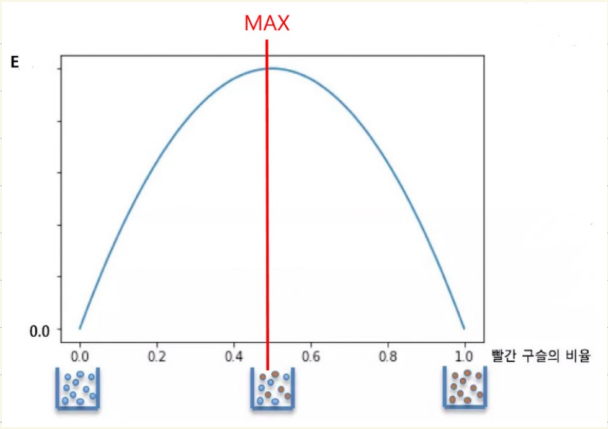
불순도

다양한 범주들의 개체들이 얼마나 포함되어 있는가를 의미합니다. 이 불순도를 수치화 한 지표로 Entropy, Gini index 등이 있습니다. 불순도를 엔트로피로 계산하는 알고리즘은 ID3, CART 알고리즘이 있다.

불순도	알고리즘
Entropy	ID3
Gini Index	CART

Entropy

$$Entropy(A) = -\sum_{k=1}^{m} P_k \log_2(P_k)$$



무질서도를 정량화 해서 표현한 값
Entropy ↑ → 분류가 어려움
Entropy ↓ → 분류가 쉬움
⇒ Entropy를 감소시키는 방향으로 분류

ID3 알고리즘

상위 노드의 Entropy 에서 하위노드의 Entropy를 뺀 값을 Information Gain 이라고 한다.

Information Gain ↑ ⇒ 엔트로피가 작아졌다.

$$\uparrow Gain(S,A) = E(S) - I(S,A)$$

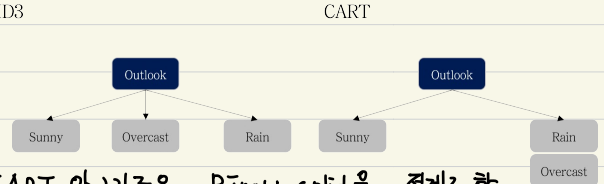
$$\downarrow I(S,A) = \sum \frac{|S_i|}{|S|} \cdot E(S_i)$$

Gini Index

데이터의 분산정도를 정량화 해서 표현한 값
Gini index를 감소시키는 방향으로 분류하기

$$Gini(D_i) = 1 - \sum_{j=1}^{x} P_j^2$$

CART 알고리즘



CART 알고리즘은 Binary split을 전제로 함

$$Gini(A) = \sum_{j=1}^{2} \frac{|D_j|}{|D|} * Gini(D_j)$$