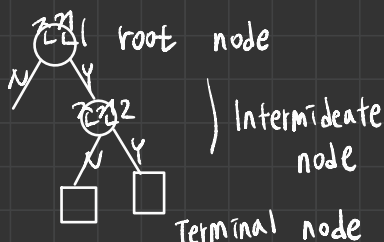


Decision Tree

- 데이터 사이에 존재하는 패턴을 예측가능한 규칙들의 집합으로 나타냄
= 나무



V 종류 Decision Tree

- L 각각의 노드는 최대한 한가지의 클래스만 가지길 싶어함 "simple"

V 기준 - 불순도

불순도 측정 지표

Entropy 엔트로피

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

1) Entropy 감소 = 불순도 감소 = 순도 증가

→ ID3 알고리즘 : Entropy 지수 이용한 알고리즘

- L Information Gain 이 크게 나오는 변수 A 기준으로 선택.

1) Information Gain = 상위 노드 Entropy - 하위 노드 Entropy (값 클수록, 엔트로피 ↓)

Gini Index 지니계수

$$\begin{aligned} \text{Gini}(A_i) &= \sum_{j=1}^2 \frac{|D_i|}{|D|} * \text{Gini}(D_i) \\ &= 1 - \sum_{j=1}^2 p_{ij}^2 \end{aligned}$$

1) Gini Index 감소 = 불순도 감소 = 순도 증가

→ CART 알고리즘 : Gini Index 기준 알고리즘

- L Binary split 전제로 완벽함.

Naive Bayes

v 베이즈 정리

두 확률 변수의 사전 확률과 사후 확률 사이의 관계를 나타내는 정리

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

① <prior> 사후확률

② Likelihood
: 사전 확률의
'관계를 잘 설명한 정도'

③ P(D)
↑
Normalizing Constant
(Evidence)
: 사건 D의 발생 가능성

But, 계산량 많아

⇒ 조건부 독립 가정!

- 가정) 종속변수가 주어졌을 때,

입력변수들이 독립적이다.

$$P(A \cap B | C) = P(A | C) P(B | C)$$

- 라플라스 스무딩

Likelihood가 0이 되는 것을 방지하도록 최소한의 확률 정해주기!

$$P_{LAP} = \frac{c(x) + 1}{\sum x [c(x) + 1]}$$

$$P_{(x|c)} = \frac{\text{count}(x, c) + 1}{\sum_{x \in V} (\text{count}(x, c) + 1)}$$

- 나이브 베이즈

장점 : 텍스트에서 강점, 입력 공간 차원 높을 때 유리.

단점 : 희귀한 확률 나왔을 때, 조건부 독립 자체가 바뀔 것
(라플라스 스무딩)