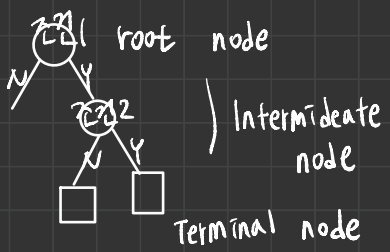
Decision Tree

- 데이터 사이에 존재하는 패턴을 예측가능한 규칙들의 조합으로 나타냄 = 나무



V 좋은 Decision Tree

└ 각각의 노드를 최대한 한가지의 클래스만 가지길 실어함 "simple"

V 기준 - 불순도

불순도 측정 지표

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

+) Entropy 감소 = 불순도 감소 = 순도 증가

→ ID3 알고리즘 : Entropy 지수 이용한 알고리즘

└ Information Gain 이 크게 나오는 변수 A 기준으로 선택.

+) Information Gain = 상위 노드 Entropy - 하위노드 Entropy (값 클수록, 엔트로피↓)

Gini Index 지니계수 $Gini(A_i) = \sum_{j=1}^{2} \frac{|D_j|}{|D|} * \underbrace{Gini(D_j)}_{=1-\sum_{j=1}^{k} P_j^2}$

+) Gini index 감소 = 불순도 감소 = 순도증가

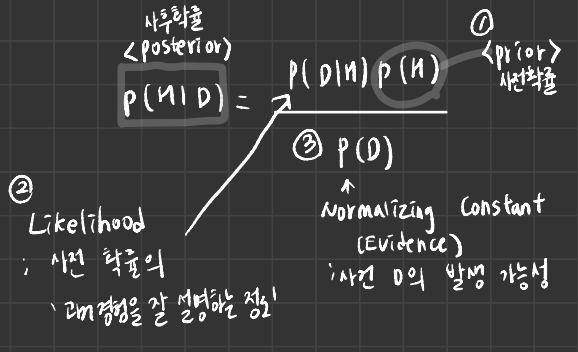
→ CART 알고리즘 : Gini index 기준 알고리즘

└ Binary split 전제로 분석함.

Naive Bayes

v 베이즈 정리

; 두 확률변수의 사전 확률과 사후 확률 사이의 관계를 나타내는 정리
(사전 확률 prior, 사후 확률 posterior)



But,, 계산량 AA

⇉ 조건부 독립 가정!

- 가정) 종속변수 Y가 주어졌을 때, 입력변수들이 독립적이다.

$$P(A \cap B \mid C) = P(A \mid C)\, P(B \mid C)$$

- 라플라스 스무딩

; Likelihood가 0이 되는 것을 방지하기 위해 최소한의 확률 정해주기!

$$P_{LAP} = \frac{C(x) + 1}{\sum x [C(x) + 1]} \qquad P_{(x|C)} = \frac{count(x, c) + 1}{\sum_{x \in V} count(x, c) + V}$$

- 나이브 베이즈

장점 : 텍스트에서 강점, 입력 공간 차원 많을 때 유리.

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