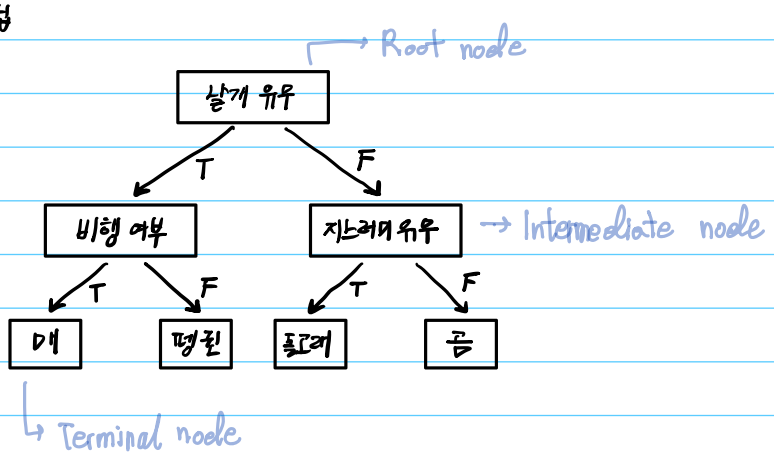
Decision Tree

Unit 1 의사결정 나무란?

- Decision Tree : 의사결정규칙을 나무구조로 나타내어 전체 데이터를 소집단으로 분류하거나 예측하는 방법



- 좋은 Decision Tree : 정확도가 높으면서, Simple한 것. 각 노드가 최대한 한가지 클래스만 가지도록이 좋음.
  - ↳ 노드를 나누는 방법 : ID3 & CART 알고리즘
  - ↳ 기준 : 불순도. (측정지표 : entropy, Gini index)

Unit 2. ID3 알고리즘

* Entropy : 무질서도를 정량화한 값. 데이터의 불확실성 의미.
  $Entropy(A) = -\sum_{k=1}^{m} p_k \log_2(p_k)$

- 불순도 지표 : Entropy
- ID3 알고리즘 : Entropy 지수를 통해 Information Gain 도출
  - ↳ 상위노드의 entropy - 하위노드의 entropy
  - ↳ Information Gain이 크게나오는 변수 A를 기준으로 선택

$$Gain(S,A) = E(S) - I(S,A) = E(S) - \sum_i \frac{|S_i|}{|S|} \cdot E(S_i)$$

(S : 주어진 데이터들의 집합
|S| : S의 데이터 개수)

Unit 3. CART 알고리즘

- 지니지수 : 데이터의 통계적 분산정도를 정량화해서 표현한 값.
  - ↳ 지니지수 ↓ = 불순도 ↓
  - ↳ $Gini(A) = \sum_{j=1}^{2} \frac{|D_j|}{|D|} \cdot Gini(D_j)$, $Gini(D_i) = 1 - \sum_{j=1}^{2} P_j^2$

• CART 알고리즘 : Gini index를 이용한 알고리즘, Binary split을 전제로 분석.

Unit 5. 가지치기

• Full tree : 모든 terminal node의 순도가 100%인 상태. 이런경우 분기가 너무 많아 과적합 위험↑

◦ 가지치기 : 의사결정나무에서 과적합을 방지하기 위해 적절한 수준에서 terminal node를 결합해주는 것

- Pre pruning (사전 가지치기) : 트리의 최대 depth나 분기점의 최소 개수를 미리 지정
- Post pruning (사후 가지치기) : 트리를 만든 후 데이터 포인트가 적은 노드를 삭제, 병합