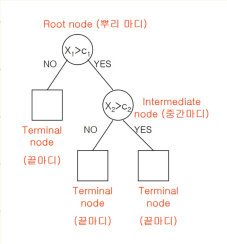
Desicion Tree : 의사 결정나무



1) 전체 데이터를 소집단으로 분류 및 예측

2) 통과하는 노드의 수가 늘어날수록 조건에 부합하는 데이터의 수 감소

3) Terminal node 데이터의 합은 root node의 데이터와 동일

좋은 Desicion Tree 가 되기 위해서는 같은 정확도에서 가장 Simple 한 것

+ 각각의 노드가 한쪽에 몰려있는 것

〈 불순도 〉

: 좋은 Desicion Tree 를 만들기 위한 기준

→ 불순도를 측정하는 지표 : Entropy, Gini index

→ 노드의 위치를 정하는 기준 : ID3, CART (연결)

① 엔트로피

- 데이터의 불확실성 , 즉 엔트로피가 높을수록 그 집단의 특징을 찾는 것이 어렵다.
- Entropy 감소 = 불순도 감소 = 순도 증가 (순도 최대 - 엔트로피 0, 순도 최소 - 엔트로피 1)

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

② ID3 알고리즘

- Entropy 지수를 이용한 알고리즘

→ Entropy 지수를 통해 Information Gain 도출한 후 Information Gain 이 크게 나오는 변수 A를 기준으로 선택하는 알고리즘

→ 엔트로피로 단일집합이 아닌 전체에서의 품질을 계산하기 위해서는 '가중치를 고려한 평균' 이용

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

* Information Gain

→ 상위노드의 Entropy 에서 하위노드의 Entropy 를 뺀 값

Information Gain 이 클수록 엔트로피가 작아진 것을 의미

③ Gini index (지니지수)

- 데이터의 통계적 분산정도를 정량화 해서 표현한 값
- Gini index 감소 = 불순도 감소 = 순도 증가

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

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

④ CART 알고리즘

- Gini index 를 이용한 알고리즘
- 데이터를 split 했을 때의 불순한 정도
- 무조건 Binary split !
- 데이터의 대상 속성을 얼마나 잘못 분류할지 계산

* feature 가 연속형일 때

step1) 각 Feature 에 대해 오름차순으로 정렬

step2) Label의 class가 변하는 지점을 찾기

step3) 경계의 평균값을 기준값으로 잡기

step4) 각 기준점에 대해 분할 후, Gini index 혹은 Entropy 계산

이 과정을 반복하면 최종 Decision Tree model을 만들 수 있음

* 가지치기

→ 모든 terminal node의 순도가 100% 인 상태를 Full tree 라고 하는데,
이 경우 분기가 너무 많아 일반화 능력이 떨어지고 과적합 위험이 생긴다.

- 이를 방지하기 위해 적절한 수준에서 terminal node를 결합해 주는 것

종류 : 사전 가지치기 , 사후 가지치기

* 앙상블

→ 의사결정 나무의 단점 해결을 위해 만들어진 방안