[의사결정 나무]

장점 : 결과 해석 용이, 비모수적 모형, 별도 가정 필요 x, 데이터 가공 필요 거의 x

단점 : 연속형 변수 비연속적 값으로 취급, 경계점 부근에서 예측 오류 가능성, 특정 변수에 수직/수평적으로 구분되지 못하면 성능 떨어짐, 과적합 쉬움, 오류 계속 전파, 노이즈 크게 영향 받음

단점 해결 방안 => 앙상블

[의사결정 나무 구조]

초기 지점- root node | 중간 node들- intermediate node | 끝 지점- terminal node

SUM(terminal node data)=root node data

[좋은 의사결정 나무]

좋은 decision tree는 최대한 simple하고 각각의 노드가 최대한 한가지 클래스만을 가진 것

기준: 불순도

1. Entropy = $-\sum_{k=1}^{m} p_k log_2(p_k)$ -> 엔트로피 감소 = 불순도 감소
   - 무질서도, 불확실성, 특정 집단 특징 찾기 어려움
   - ID3 알고리즘
     - Information Gain이 큰, 즉, Entropy를 가장 많이 줄인 변수 선택
     - Information Gain = 상위 노드 Entropy- 하위 노드 Entropy = $E(S) - \sum_i \frac{|S_i|}{|S|} E(S_i)$ (S: 주어진 데이터들의 집합, |S|: 주어진 데이터들의 집합의 데이터 개수)
2. Gini index = $\sum_{j=1}^{2} \frac{|D_j|}{|D|}(1 - \sum_{j=1}^{x} P_j^2)$ -> Gini index 감소 = 불순도 감소
   - 분산 정도
   - CART 알고리즘
     - 데이터를 split 했을 때의 불순한 정도 (Binary split을 전제)
     - 가장 작은 Gini index 값을 가지는 변수로 최초 split됨
     - Ex. Min[Gini(age=youth), Gini(age=middle_aged), Gini(age=senior)] = Gini(age=middle_aged) = Min(Gini(age)) 이고 이것이 Min(Gini(income)), Min(Gini(credit)), Min(Gini(student)) 보다 작다면 Age에서 Middle_aged와 youth, senior로 최초 split

[연속형 변수일 때]

1. 각 Feature 정렬 후 Label의 class가 바뀌는 지점을 찾고 그 경계의 평균값을 기준값으로 잡는다.
2. 각 기준점에 대해 분할 후 Gini index나 Entropy 계산

[가지치기]

Full tree = 모든 terminal node의 순도가 100%인 상태

분기가 너무 많아 일반화 low, 과적합 위험 high -> 이를 방지하기 위해 적절한 수준에서 terminal node 결합해줌

사전 가지치기 – 트리의 최대 depth나 분기점의 최소 개수 미리 지정

사후 가지치기 – 트리 만든 후 데이터 포인트가 적은 노드 삭제/ 병합