Decision tree

Point 영향력이 큰 특징을 상위 노드로, 영향력이 작은 특징은 하위 노드로!

데이터가 주어졌을 때 특징별 영향력의 크고 작음을 비교하는 방법

* Entropy

정보이론에서 불확실성을 수치적으로 표현한 값

$Gain(T, X) = Entropy(T) - Entropy(T, X)$

확률 $Entropy = \sum_{i=1}^{m} -p_i \log_2 p_i$

특징 $Entropy = \sum_{c \in X} P(c) E(c)$

X : 선택된 특징

C : 선택된 특징에 의해 생성되는 하위 노드

P(C) : 선택된 특징에 의해 생성된 하위노드에 데이터가 속할 확률

E(C) : 선택된 특징에 의해 생성된 하위 노드의 엔트로피

* Gini coefficient

특징에 의한 분리가 이진 분류로 나타날 경우 사용

지니 계수의 특징

1. 특징이 항상 이진분류로 나뉠 때 사용됨
2. 지니 계수가 높을수록 순도가 높음

순도가 높다 → 한 그룹에 모여있는 데이터들의 속성들이 많이 일치한다.

불순도가 높다 → 한 그룹에 여러 속성의 데이터가 많이 섞여있다.

decision tree 알고리즘은 지니 계수가 높은 특징으로 트리의 노드를 결정한다.

지니계수를 통해 트리의 노드를 결정하는 순서

1. 특징으로 분리된 두 노드의 지니 계수를 구함 ($p^2 + q^2$)
2. 특징에 대한 지니 계수를 구함

Decision tree의 장단점

장점 1. 수학적 지식이 없어도 결과를 해석하고 이해하기 쉽다

2. 수치 데이터, 범주 데이터에 모두 사용가능하다

단점 과대적합의 위험이 높다.

학습 시 적절한 리프노드의 샘플 수와 트리의 깊이에 제한을 두어 모델이 치우치지 않게 주의