트리 기반의 feature importane는 어떻게 계산이 될까?

- 의사 결정 트리 학습 알고리즘 → 트리가 루트 노드를 기준으로 하여 학습을 진행할때 어떠한 기준이나 근거를 가지고 학습이 되는지를 정하는 알고리즘이다.
- 의사 결정 트리 학습 알고리즘의 종류 : ID3, C4.5, CART
- 이중에서 Scikit-learn에서 주로 사용하는 CART 알고리즘에 대해 알아보겠습니다.

CART 알고리즘

- CART(Classification and Regression Tree) : 각 노드가 2개의 child node를 가지는 binary tree를 적절한 불순도(Impurity)를 기준으로 생성해 나가는 알고리즘이다.
- 목적(분류, 회귀)에 따라 불순도 지표가 달라진다
- 분류 : 지니(Gini), 엔트로피(Entrophy를 이용한다.
- 회귀 : MSE(Mean Squared Error)등을 이용하여 분산이 감소되는 방향으로 학습이 진행된다.
- 즉 불순도를 가장 크게 감소시키는 변수의 중요도가 가장 크게 계산이 된다.

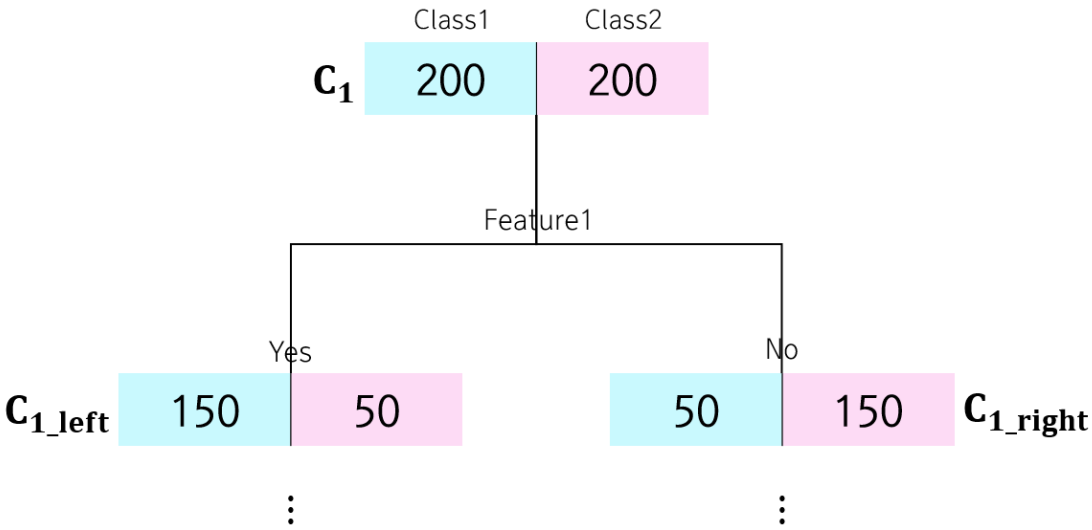
Gini importance

- Scikit-learn에서는 지니 중요도(Gini Importance)를 이용해서 각 featre의 중요도를 측정합니다.
- 지니 불순도(Gini Impurity) : Class가 총 K개가 있고, 각 샘플이 해당 class에 속할 확률을 각각 p_i, i = 1, ... , K개라고 할 때, 특정 노드 N_j에서 지니 불순도는 아래와 같다.

$$G(N_j) = \sum p_i(1 - p_i) = 1 = \sum p_i^2$$

- 위와 같은 식으로 지니 불순도가 계산되기 때문에, 해당 노드에서 샘플들이 이질적으로 구성되어 있을시, 다시 말해 Class에 골고루 분포 되어 있을 수록 지니 불순도는 높게 계산되어 진다.

Gini importnace 예시



- 위와 같은 간단한 예시의 지니 불순도를 구해보면 아래와 같다(계산과정 생략)
- G(C_1) = 0.5 , G(C_1_left) = 0.25, G(C_1_right) = 0.25
- 결과를 보면 직관적으로 아래로 갈수록 두 클래스가 골고루 섞이지 않고 특정 클래스에 치중된 것을 확인 할 수 있다. → 순도(homogeneity)가 증가했다고도 볼 수 있다.
- 여기서 순도란건 Class 적인 관점에서 접근해야한다. Class별 분포가 균등하다는 개념과 혼동하지 않도록 주의하자!
-