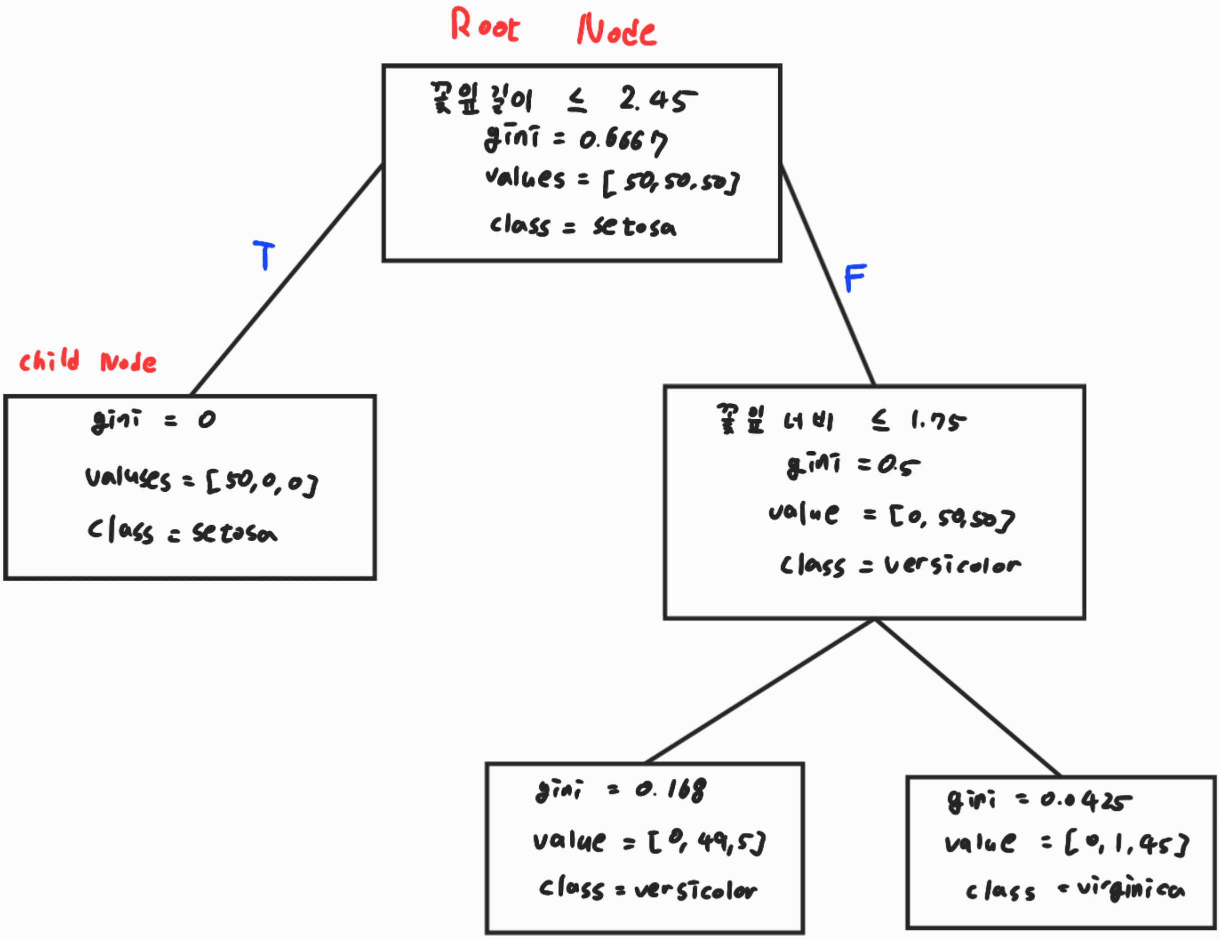

gini 계수 : 불순도(impurity) 측정

$$G_i = 1 - \sum_{k=1}^{n} P_{i,k}^2$$ ($P_{i,k}$ = i번째 노드에 있는 훈련 샘플 중 k에 속한 비율)

max_depth 를 통해 최대 깊이 조절 가능

※ sklearn 은 이진트리만 만드는 CART 알고리즘 사용

모델 해석 : white box & black box

- white box : 매우 직관적, 결정방식 이해 쉬움
- Black Box : 어떻게 예측 결과가 나오는지 모름

클래스 확률 추정

: k 클래스일 확률 반환 ⇒ 조건에 만족하는 샘플 / 전체 샘플(리프노드 안에)

CART 훈련 알고리즘(Classification And Regression Tree)

$$J(k, t_k) = \frac{m_{left}}{m} G_{left} + \frac{m_{Right}}{m} G_{Right}$$ (특성 : k, 특성 k의 임계값 : t_k, 불순도 : G, m : 서브셋 샘플 수)

→ 최소화해야 하는 CART 비용함수

⇒ 서브셋 나누는 것 반복 ⟶ 최대 깊이 or 불순도 줄이는 분할 못찾을 때 중지

CART = Greedy Algorithm ⇒ 최적의 분할 찾기위해 반복
But, 최적 보장 X , 종종 납득할 만한 솔루션 도출

최적트리 계산 = NP-완전

시간복잡도 = $O(\exp(m))$

계산 복잡도

: 결정트리 탐색 $O(\log_2 m)$ ⇒ ∴ 특성수와 무관하게 시간복잡도 $O(\log_2 m)$

⇒ But 훈련 알고리즘은 모든 훈련샘플의 특성비교

⇒ $O(n \times m \log(m))$ sklearn presort = True 옵션 사용하여 정렬후 훈련

⇒ 훈련시간 감소

지니 불순도 또는 엔트로피

- 엔트로피 : 분자가 안정되면 즉, 손실 X 면 Entropy = 0

$H_i = -\sum_{k=1}^{n} p_{i,k} \log_2(p_{i,k})$ (H_i : i번째노드 불순도, k : 특성, $p_{i,k}$ = k일 확률)

규제 매개변수

- 비파라미터 모델(nonparametric model) : 훈련되기 전에 파라미터 수가 결정되지 않는 모델   ex.결정트리
- 파라미터 모델(parametric model) : 미리 정의된 모델 파라미터 수를 가지므로 자유도가 제한되고 과대적합될 위험이 줄어듦 ex. 선형 모델

규제 매개변수 : 과대적합을 막기 위해 매개변수 설정(결정트리의 자유도 및 max_depth)

- Decision Tre Classifier의 규제 매개변수

1. min_samples_leaf : 리프 노드가 가지고 있어야 할 최소 샘플 수
2. min_weight_fraction_leaf : 가중치가 부여된 전체 샘플 수에서의 비율
3. max_leaf_nodes : 리프 노드의 최대 수
4. max_features : 각 노드에서 분할에 사용할 특성의 최대 수

※ 참고

사후가지치기 = 제한없이 결정트리를 가지치기(제거)하는 알고리즘

⇒ χ^2 검정을 통해 p-값이 높다면 삭제

DecisionTree Regressor

⇒ 분류트리와 비슷하지만 MSE (Mean Squared Error) 사용

회귀를 위한 CART 비용함수

$$J(k, t_k) = \frac{m_{left}}{m} MSE_{left} + \frac{m_{right}}{m} MSE_{right} \quad \left(MSE_{node} = \sum (\hat{y} - y^{i})^2 \;,\; \hat{y}_{node} = \frac{1}{m_{node}} \sum y^{(i)} \right)$$

불안정성

- 결정 트리는 계단 모양의 결정 경계 생성
- 이상치에 매우 민감
- 일반화를 위해 PCA 기법을 통해 훈련 데이터를 더 좋은 방향으로 회전
- 확률적이기 때문에 같은 훈련 데이터에서도 다른 모델 얻을 수 있음