Decision Tree

6.6-6.9

Entropy

정보 이론

: 데이터를 정량화하기 위한 응용 수학 분야 중 하나

: 정보량이 높다 = 어떤 일이 일어날 확률이 낮다. 불확실하다.

: 예시) 동전 던져 앞면 나올 확률 = 1, 주사위 던져서 6 나올 확률 = 2.5849
=> 주사위 던져서 6 나올 확률이 덜 발생함으로 높은 정보량을 갖는다.

섀넌 엔트로피

$$I(x) = -\log P(x)$$

엔트로피

: 정보량의 평균

: 분자의 무질서함을 측정하는 개념

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

Gini vs Entropy

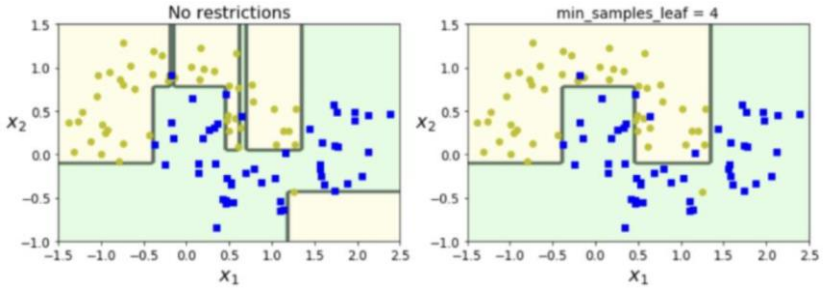
- 계산 빠르기 : Gini > Entropy
- 균형은 트리 생성 : Gini < Entropy

규제 매개변수

정지조건

1. 더 이상 분리해도 불순도가 줄어들지 않을 때
2. 자식 마디에 남은 sample 수가 너무 적을 때
3. 분석자가 지정한 규제 매개변수에 도달했을 때

예시



왜 사용할까?

Overfitting의 위험성이 크기 때문

종류

Max_depth : 최대 깊이 설정

Min_samples_split : 분할되기 위해 노드가 가져야하는 최소 샘플 수

Min_samples_leaf : 리프 노드가 가지고 있어야하는 최소 샘플 수

Max_leaf_nodes : 리프 노드의 최대 수

Max_features : 각 노드에서 분할에 사용할 특성의 최대 수

회귀

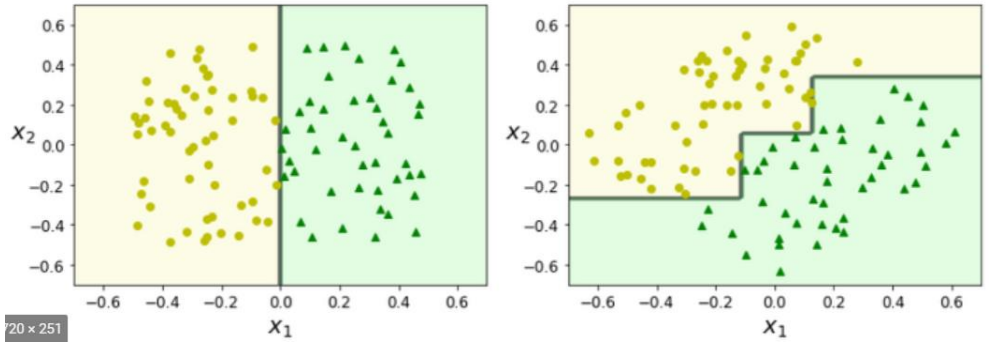
- 의사 결정 분류 트리와 거의 동일
- 불순도가 아닌 MSE를 최소화 하도록 분할

$$MSE = \frac{1}{n}\Sigma\underbrace{\left(y - \hat{y}\right)^2}_{\text{The square of the difference between actual and predicted}}$$

$$J(k,t_k) = \frac{m_{left}}{m} MSE_{left} + \frac{m_{right}}{m} MSE_{right}$$

불안정성

- 결정 트리는 계단 모양의 결정 경계를 만듦
- 따라서 회전에 민감
- PCA 기법을 사용하여 더 좋은 방향으로 회전을 시키면서 모델을 일반화 할 수 있음



데이터 셋을 45도 회전

정리

장점

- 직관적
- 이상치, 노이즈에 큰 영향 x
- 높은 모델 해석력
- 연속형 데이터, 범주형 데이터 모두 처리 가능
- 균일도에만 초점 가능(scaling 불필요)

단점

- 일반화가 어려움(불안정성)
: 학습데이터에 따른 차이가 큼
= 모델 variance가 높음
- Overfitting 가능성 매우 높음