6주차 숙제

의사결정트리

정의: 데이터의 속성으로부터 패턴을 학습해 규칙을 생성하는 지도학습 모델.

분류, 회귀 모두 가능

용어: Root node: 시작점

Parent/Child Node: 상위/하위 노드

Leaf Node: 분류/예측이 끝나는 지점

Depth: 트리의 깊이

수행 과정: 분리기준(불순도 감소)과 정지규칙(깊이 제한 등) 설정 -> 가지치기: 과적합 방지를 위해 불필요한 분기 제거 -> 타당성 평가: 모델 평가 및 해석

장점: 해석 용이, 변수 간 상호작용 해석 가능, 비모수적

단점: 경계 근처 예측 오류, 불안정성, 선형성 및 주효과 해석 어려움

지니 불순도: 노드의 순수도를 측정하는 기준. 노드 안에 여러 클래스가 섞여 있을 때, 해당 노드가 얼마나 불순한지를 수치로 나타냄.

폰트, 그래픽, 텍스트, 상징이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다. K: 클래스의 개수 | pi,k: 노드 i에 속한 샘플 중 클래스 k에 속한 비율.

0에 가까울수록 순수한 노드이며 두 클래스의 비율이 1:1일 때 불순도가 가장 큼. 자식 노드의 지니 불순도를 최소화하도록 부모 노드를 분할

CART 알고리즘: 지니 불순도를 활용해 가장 순수한 자식 노드를 만들기 위해 쓰이는 알고리즘 지니 불순도 말고도 엔트로피, MSE도 사용가능

결정트리 규제: 모델을 조절해서 일반화 성능을 높이기 위해 필요.

| **파라미터** | **의미** | **값을 어떻게 하면 규제가 강해지는가** |
| --- | --- | --- |
| max\_depth | 트리의 최대 깊이 | **작게 설정** |
| min\_samples\_split | 노드를 분할하기 위한 최소 샘플 수 | **크게 설정** |
| min\_samples\_leaf | 리프 노드가 가져야 할 최소 샘플 수 | **크게 설정** |
| max\_leaf\_nodes | 최대 리프 노드 수 제한 | **작게 설정** |
| max\_features | 분할 시 고려할 최대 특성 수 | **작게 설정** |