머신러닝 6주차 과제

2024170928 노연경

**1. 의사결정트리**

: 각 데이터들이 가진 속성들로부터 패턴을 예측 가능한 규칙들의 조합으로 나타내며, 이를 바탕으로 분류를 수행할 수 있도록 하는 지도학습 모델

- 분류와 회귀 모두 가능

1) 분류 : 새로운 데이터가 속하는 해당 끝마디에서 가장 빈도가 높은 범주로 새로운 데이터를 분류

2) 회귀 : 해당 끝마디의 종속변수의 평균을 반환

**2. 의사결정트리 수행과정**

1) 형성

분리기준과 정지규칙을 지정하여 의사결정나무 모형을 생성

부모마디의 순수도에 비해 자식마디들의 순수도가 증가하도록 자식마디를 형성

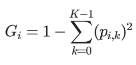
2) 가지치기

분류오류를 크게 할 위험이 높거나 부적절한 추론 규칙을 가지고 있는 가지를 제거

3) 타당성 평가 & 해석 및 예측

**3. 결정트리 훈련과 활용**

\* 지니 불순도 : 결정 트리 알고리즘에서 노드의 순수도를 측정하는 지표



\* CART : Classification and Regression Tree

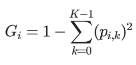
분류와 회귀 문제 모두에 사용할 수 있는 결정트리 알고리즘

비용함수가 작을수록 불순도가 낮은 두 개의 부분집합으로 분할됨

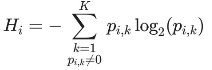
**4. 지니 불순도 vs 엔트로피**

- 지니 불순도와 엔트로피 모두 목적은 같음 ; 불순도가 낮을수록 좋음

- 지니 불순도 계산



- 엔트로피 계산



- 둘의 성능 차이는 거의 없다

**5. 결정트리 규제**

- 비파라미터 모델 : 데이터에 대해 가정을 하지 않고, 노드로 분할할 때 제한이 가해지지 않으며, 노드를 분할할 때마다 새로운 파라미터가 학습되기 때문에 학습되어야 하는 파라미터 개수를 미리 알 수 없음, 과대적합 가능성이 높음

- 파라미터 모델 : 파라미터 수가 훈련 시작 전에 규정됨, 과대적합 가능성이 상대적으로 적음

- 결정트리는 비파라미터 모델

- 사이킷런 DecisionTreeClassifier 규제 하이퍼파라미터 규제 강화방법

1) min\_접두사 사용 규제 : 값을 키울 것

2) max\_접두사 사용 규제 : 값을 감소시킬 것

**6. 결정트리 단점**

1) 훈련셋 회전 민감도 : 결정트리는 항상 축에 수직인 분할을 사용해서 조금만 회전을 가해도 결정 경계가 많이 달라짐

2) 높은 분산 : 훈련 데이터의 작은 변화에도 매우 민감함. (random\_state를 지정하지 않으면서 동일한 모델을 훈련시키면 다른 결과가 나옴) 🡪 여러 개의 결정트리를 동시에 훈련시킨 후 평균값을 활용하는 랜덤 포레스트 모델을 이용하면 해결 가능