**의사결정나무(DecisionTree)**

**의사결정나무(DecisionTree)**

**Theory**

의사결정나무는 분류함수를 의사결정 규칙으로 이뤄진 나무 모양으로 그리는 방법입니다. 계산 결과가 의사결정나무에 직접 나타나기 떄문에 해석이 간편하며, if/else 기반으로 최대한 많은 데이터 세트가 해당 분류에 속할 수 있도록 결정 노드의 규칙이 정해지는 방법입니다.

**정보이득**

정보이득은 엔트로피라는 개념을 기반으로 합니다. 엔트로피는 주어진 데이터 집합의 혼잡도를 의미하는데, 서로 다른 값이 섞여 있으면 엔트로피가 높고, 같은 값이 섞여 있으면 엔트로피가 낮습니다. **정보 이득 지수는 1에서 엔트로피를 뺀 값입니다.** 결정 트리는 이 정보 이득 지수로 분할 기준을 정합니다.

**특징**

정보의 균일도만 신경 쓰면 되는 알고리즘이기 때문에 특별한 경우를 제외하고는 각 피처의 스케일링과 정규화 같은 전처리 작업이 필요 없습니다. 반면에 결정 트리의 가장 큰 단점은 모델의 높은 자유도로 인해 과적합이 발생할 여지가 크다는 점입니다. 또한, 하이퍼파라미터 튜닝을 할 인자가 너무 많다는 단점도 있습니다.

**하이퍼파라미터튜닝**

결정트리는 규칙 생성 로직을 미리 제어하지 않으면 완벽하게 클래스 값을 구별해내기 위해 트리 노드를 계속해서 만들어 갑니다. 따라서 결정트리 알고리즘을 제어하는 대부분의 하이퍼 파라미터는 복잡한 트리가 생성되는 것을 막기 위한 용도입니다.

{'ccp\_alpha': 0.0,

'class\_weight': None,

'criterion': 'gini',

'max\_depth': 3,

'max\_features': 0.9,

'max\_leaf\_nodes': None,

'min\_impurity\_decrease': 0.0,

'min\_impurity\_split': None,

'min\_samples\_leaf': 1,

'min\_samples\_split': 2,

'min\_weight\_fraction\_leaf': 0.0,

'random\_state': None,

'splitter': 'best'}