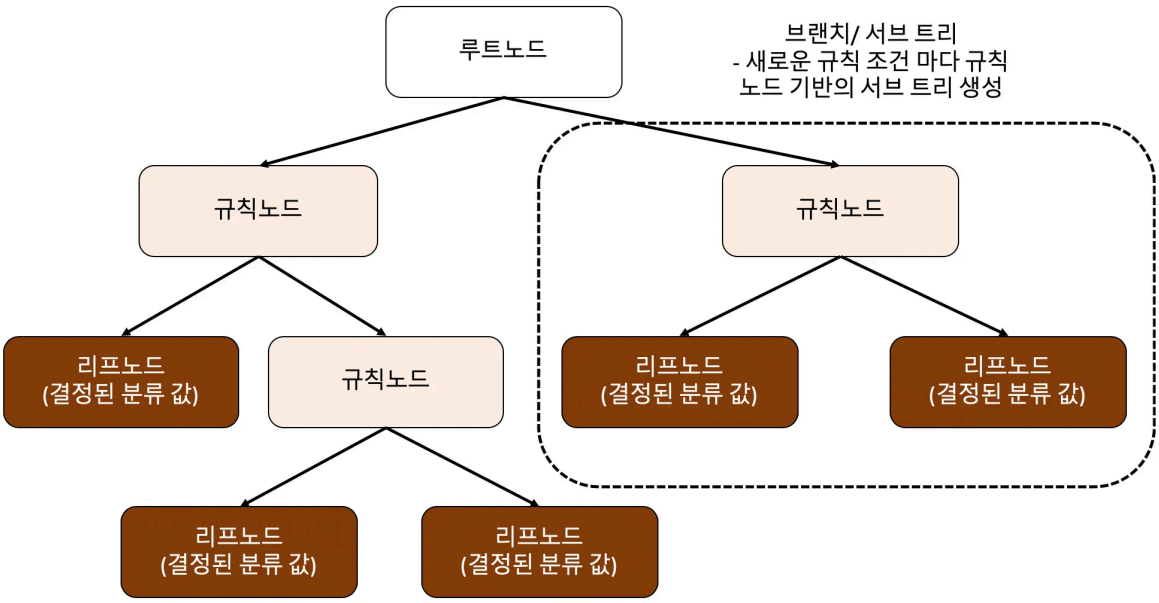
Ch 6. 의사결정트리

📍의사결정트리

-**목적**: 데이터를 나누는 기준을 찾아서 정답을 더 잘 예측할 수 있도록 함

-**종류**: 지도학습(학습 데이터에 레이블 존재)

-분류/회귀 가능



📍의사결정트리의 형성과정

-분리규칙

-정지규칙

-가지치기

-예측값 할당

📍분리규칙

-CART 훈련 알고리즘 사용

- m, m_{left}, m_{right}: 각각 부모와 양쪽 자식 노드에 속한 샘플 수
- G_{left}, G_{right}: 두 자식 노드의 지니 불순도

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

-**불순도**: 데이터가 같은 종류로만 이루어져있을 수록 불순도가 낮아짐

-**엔트로피 지수**: 엔트로피 지수가 높으면 불순도가 높음을 의미 *머신러닝에서 많이 사용하는 지수

-**지니 계수**: 지니계수가 높으면 불순도가 높음을 의미

📍정지규칙

-더 이상 불순도가 줄어들지 않을 때 정지

-각 자식마디에 있는 샘플의 수가 너무 적을 때 정지

-규제 파라미터에 도달했을 때 정지

***결정트리 규제**

-비파라미터 모델은 자유도에 제한이 없어 과대적합이 일어날 가능성이 높기 때문에 파라미터 규제를 시행

-DecisionTreeClassifier 하이퍼파라미터(max_: 값을 감소시켜야 규제 강화, min_: 값을 증가시켜야 규제 강화)

- max_depth
- min_samples_split: 이 값보다 작은 리프는 더이상 분할하지 않음
- min_samples_leaf: 분할 결과 리프에 이 값보다 더 작은 샘플이 있으면 안됨
- max_leaf_nodes: 최대 리프 노드 개수
- max_features: 분할에 사용되는 특성의 개수

📍가지치기

-**필요성**: depth가 깊어질수록 과대적합의 위험성이 높아짐

📍의사결정트리의 장단점 및 활용

-**장점**: 데이터 전처리가 따로 필요하지 않음(스케일 등에 영향을 받지 않기 때문)

-**단점**: 과대적합의 위험성이 큼, 모델의 분산이 큼(훈련데이터에만 의존하기 때문)

-**활용**: 앙상블에 주로 사용됨