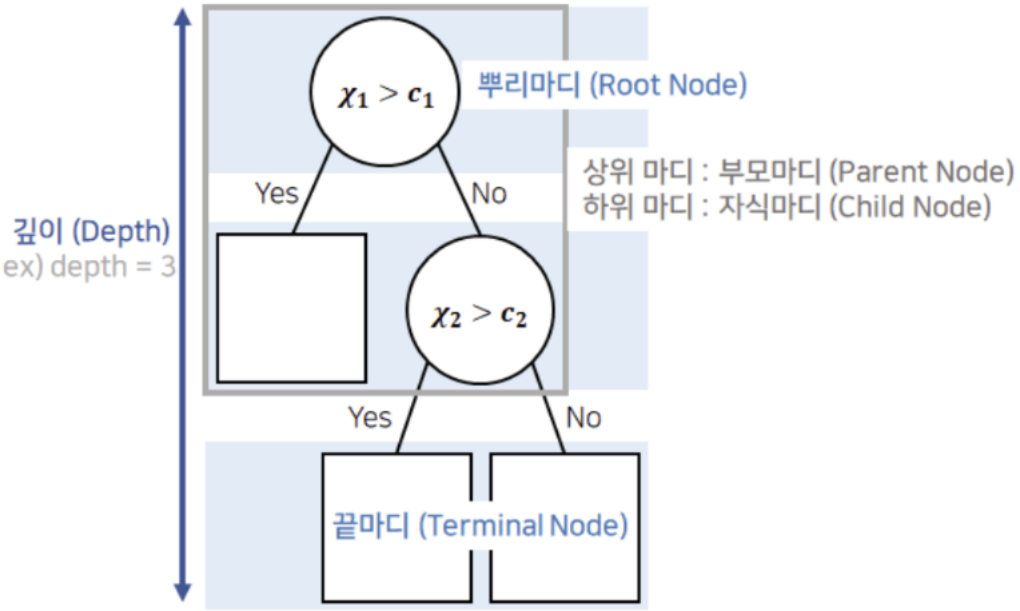
머신러닝 세션 6주차 과제

2023170832 고지원

1. 의사결정트리(Decision Tree)

- 데이터 속성 패턴을 예측 가능한 규칙 조합으로 나타냄
- 분류, 회귀 모두 가능 (범주형, 연속형 모두 예측 가능)
- **구조**



- **끝마디 반환값**

  ① 분류 : 가장 빈도 높은 범주로 새로운 데이터 분류
  ② 회귀 : 종속변수 평균의 예측값 반환 (예측값 종류 = 끝마디 개수)

- **수행과정**
  ① 의사결정나무 형성
    ○ 분리기준 : 순수도, 불순도
    ○ 정지규칙 : 깊이나 끝마디의 개수 설정
  ② 가지치기
    ○ 부적절한 추론 규칙 가지고 있는 가지 제거
    ○ 과적합 방지
  ③ 타당성 평가 & 해석 및 예측
- **장점** : 해석 용이, 교호작용 해석, 비모수적 모형
- **단점** : 비연속형, 선형성/주효과 결여 / 비안정성(훈련셋 회전 민감, 높은 분산)

2. 결정트리 훈련과 활용

- 데이터 전처리 불필요
- 길이와 너비 기준으로 분류
- export_graphviz() : 시각화
- predict_proba() : 지정된 샘플의 범주별 추정 확률 계산
- predict() : 품종 범주 예측, 가장 높은 추정 확률 갖는 품종 지정

3. CART 알고리즘 (Classification and Regression Tree)

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

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

- 비용함수 작을수록 불순도 낮은 두 개의 부분집합으로 분할됨

4. 지니 불순도

$$G_i = 1 - \sum_{k=0}^{K-1} (p_{i,k})^2$$

G_i : i번째 노드의 지니 불순도
$P_{i,k}$: i번째 노드에 있는 훈련 샘플 중 범주 k에 속한 샘플 비율
k : 범주 개수

지니 불순도 대신 엔트로피 사용 -> 큰 차이 없음, 빠르게 훈련됨

5. 결정트리 규제

- 파라미터 모델 : 훈련 시작 전 파라미터 수 규정. 과대적합 가능성 감소
- 비파라미터 모델 : 자유도 제한 X, 과대적합 가능성 높음