의사결정 나무 (Decision Tree)

→ 여러가지 규칙을 순차적으로 적용하면서 독립변수 공간을 분할하는 분류모형.
i) classification ii) Regression 에 모두 사용가능해서 CART (Classification And Regression Tree)라고도 불린다.

의사 결정 나무를 이용한 분류

- 의사결정 나무를 이용한 분류법은 다음과 같다.

① 여러가지 독립변수 중 하나의 독립변수를 선택하고, 그 독립변수에 대한 기준값 (threshold)를 정한다. ⇒ 이를 "분류규칙"이라고 함.

② 전체 학습 데이터 집합(최상위 노드)을 해당 독립변수의 값이 기준값보다 작은 데이터 그룹(노드1)과 해당 독립변수의 값이 기준값보다 큰 데이터 그룹(자식노드2)로 나눈다.

③ 각각의 자식노드에 대해 1,2 단계를 반복하여 하위의 자식 노드를 만든다. 단, 자식노드에 한 가지 클래스의 Data만 존재한다면, 더이상 자식노드를 나누지 않고 종지한다.

⇒ 이렇게 자식 노드 나누기를 연속적으로 적용하면 노드가 계속 증가하는 나무와 같은 형태로 표현할 수 있다.

의사 결정 나무를 사용한 분류 예측

- 의사 결정 나무에 전체 training data를 모두 적용해보면 각 Data는 특정한 Node를 타고 내려가게 된다. 각 노드는 그 노드를 선택한 데이터 집합을 가진다.
- 이 때, 노드에 속한 data의 클래스의 비율을 구하여 이를 그 노드의 조건부 확률 표호 $P(Y=k|X)_{node}$ 라고 정의한다.

$$P(Y=k|X)_{node} \approx \frac{N_{node,k}}{N_{node}}$$

- Test Data (X_{test})의 클래스를 예측할 때는, 가장 상위의 노드를 시작으로 분류 규칙을 차례대로 적용하여, 마지막으로 도달하는 노드의 조건부 확률 표호를 이용하여 클래스를 예측한다.

$$\hat{y} = \arg\max_k P(Y=k|X_{test})_{last\ node}$$