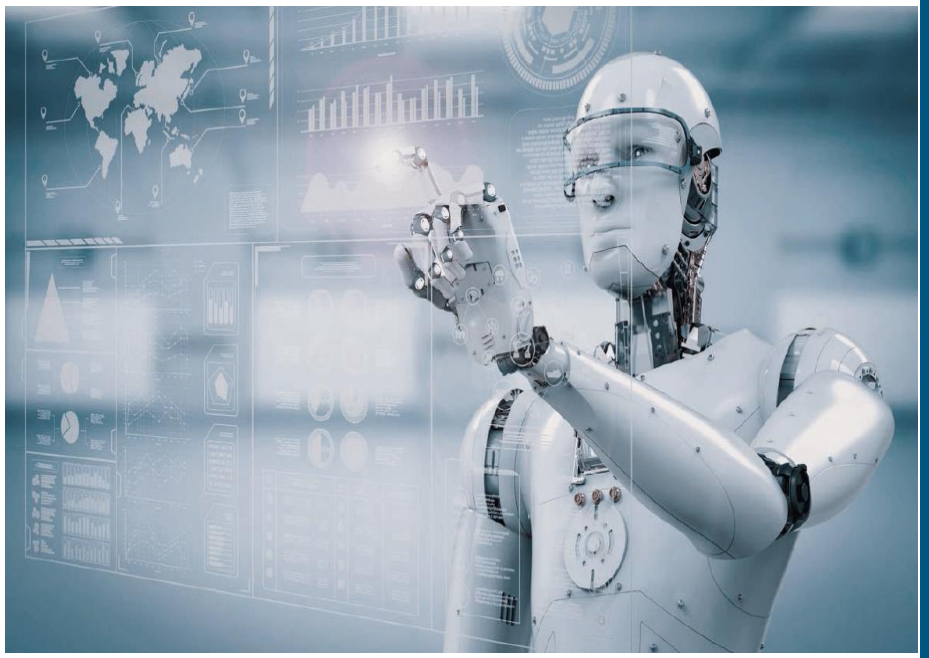


기계학습개론

- 트리 알고리즘 -

교수 이홍로

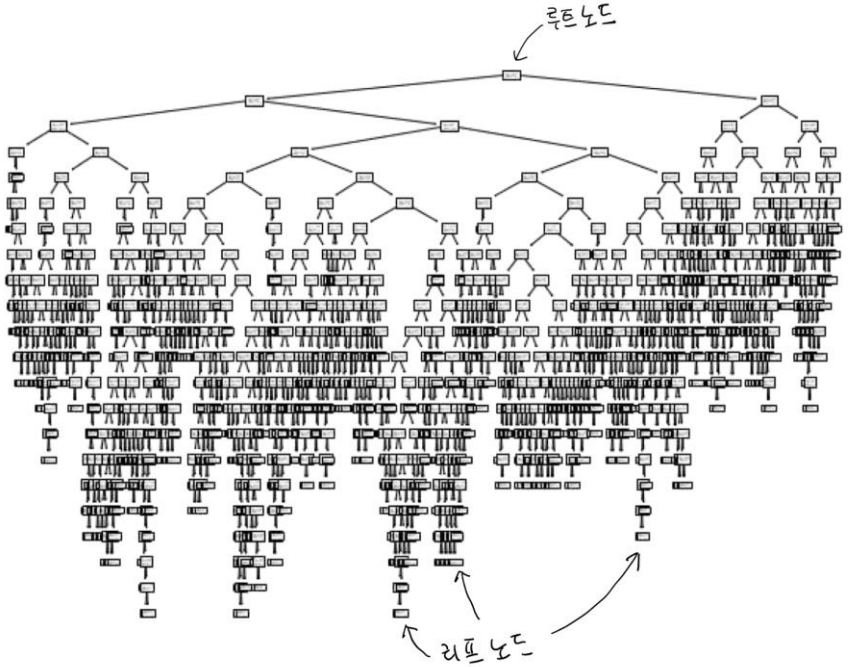
MP : 010-6611-3896
E-mail : hrlee@cnu.ac.kr
강의 홈페이지 : https://cyber.hanbat.ac.kr/

강의 목표

- **결정트리**
- **교차검증과 그리드 서치**
- **트리의 앙상블**
  - 랜덤 포레스트

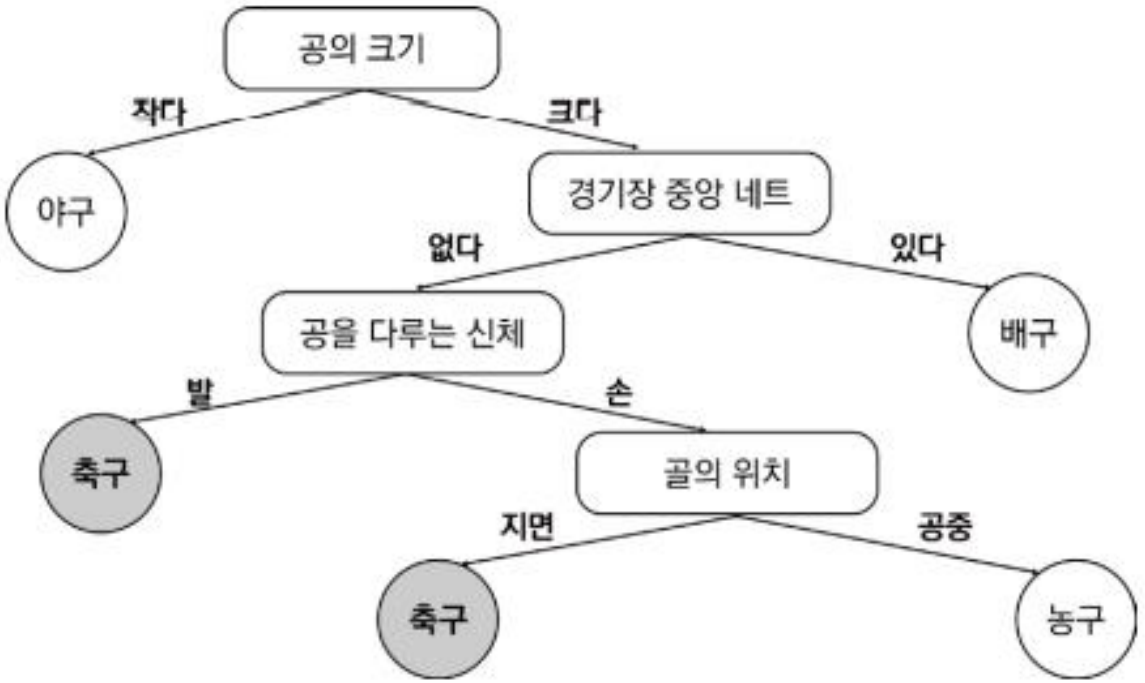
결정트리

- **결정트리**
  - 트리 형태로 의사결정 지식을 표현하여 데이터를 분류하는 기법
  - 관측값과 목표값을 연결하는 예측 모델로써, 분류와 회귀 작업, 다중출력 작업이 가능한 머신러닝 알고리즘
  - '스무고개' 문답처럼 선택 방법으로 진행
    - 주택이나 자동차 구입비용 등의 **예측에 활용**
    - 타이타닉호 탑승객의 **생존 여부**를 나타내는 결정 트리
    - 운동경기 진행 적합 여부 판단
    - 붓꽃 종류 분류



결정트리

- 데이터들을 트리 구조의 루트root에서 시작하여 차례로 중간 노드internal node들을 거쳐 단말 노드leaf node에 배정하는 기능을 수행



(ball, net, hand, goal) – 종목

(작다, 없다, 0, 없다) – 야구
(크다, 없다, 0, 지면) – 축구
(크다, 없다, X, 지면) – 축구
(크다, 있다, 0, 없다) – 배구
(크다, 없다, 0, 공중) – 농구

결정트리

- **속성의 선택**
  - 데이터는 여러 가지 속성 중에서 어떤 것이 가장 중요한 것인지 판단하기 위해 정보 이득information gain이라는 개념이 사용
  - **정보이득 :** 특정한 속성이 원하는 분류 방식에 부합하게 데이터를 나누는지를 측정할 수 있는 척도
  - **엔트로피 :** 원래 정보량을 측정하기 위해 고안된 것으로, 이 값이 크면 많은 정보가 담겨 있다는 것

$$H(S) = -p^{+}\lg p^{+} - p^{-}\lg p^{-}$$

p^{+} : 전체 데이터 표본 S에서 차지하는 양성 데이터의 비율
p^{-} : 음성인 데이터 비율
H(S) : 엔트로피

결정트리

- **ID3 알고리즘과 지니 불순도 측정 방법이 있음**
- **ID3 알고리즘**
  - 가장 엔트로피가 높은 경우
    - 두 부류가 균등하게 섞여 있는 경우로 각각의 비율이 1/2인 경우

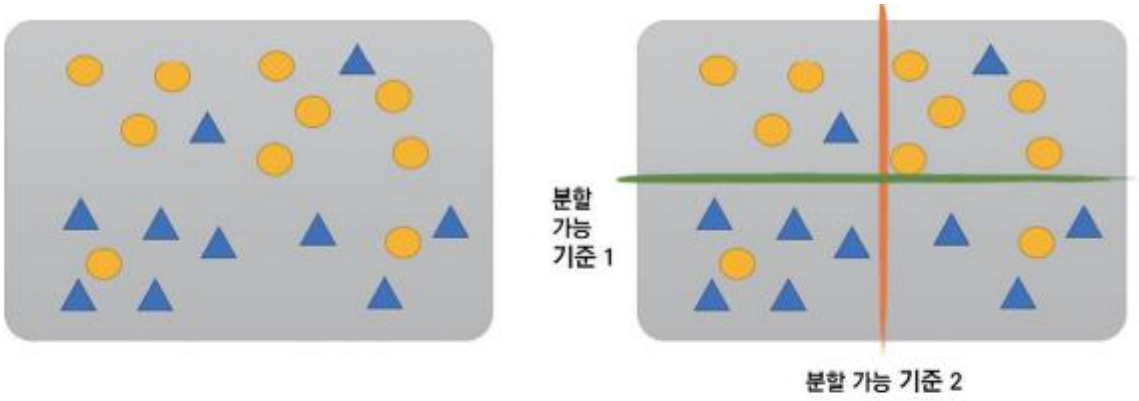
$$H(p^+ = 0.5, p^- = 0.5) = -\frac{1}{2}\lg 2^{-1} - \frac{1}{2}\lg 2^{-1} = 1$$

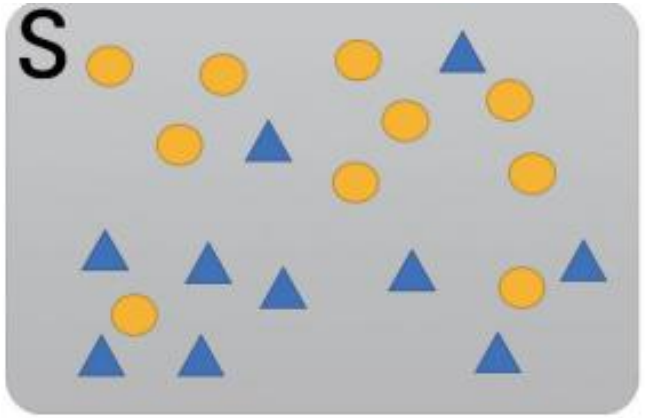
  - **정보 이득**
    - 특정한 속성에 따라 데이터를 나누었을 때 줄어든 엔트로피로 정의
    - 어떤 속성 $\mathbb{A}$가 가지는 모든 값들의 집합을 A라고 할 때, 데이터 표본 S를 나누어서 얻는 정보 이득 $G(S, A)$

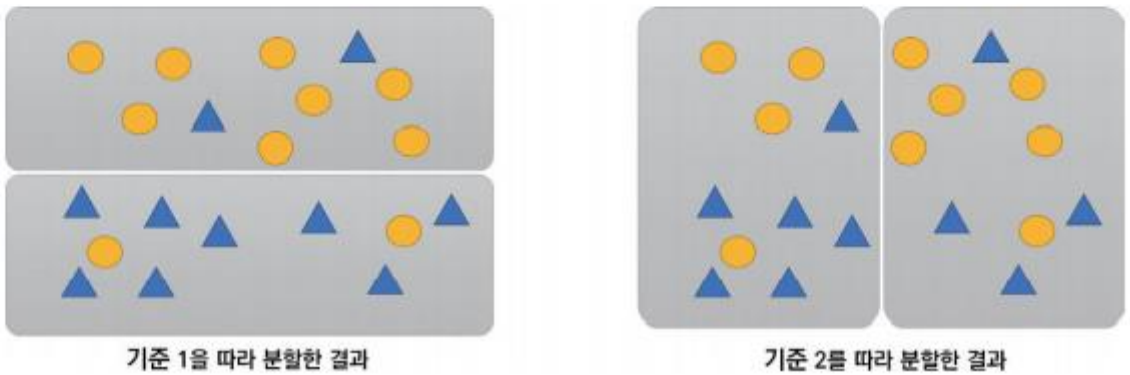
$$G(S, A) = H(S) - \sum_{a \in \mathbb{A}} \frac{|S_a|}{|S|} H(S_a)$$

정보이득이 가장 큰 속성을 선택

결정트리





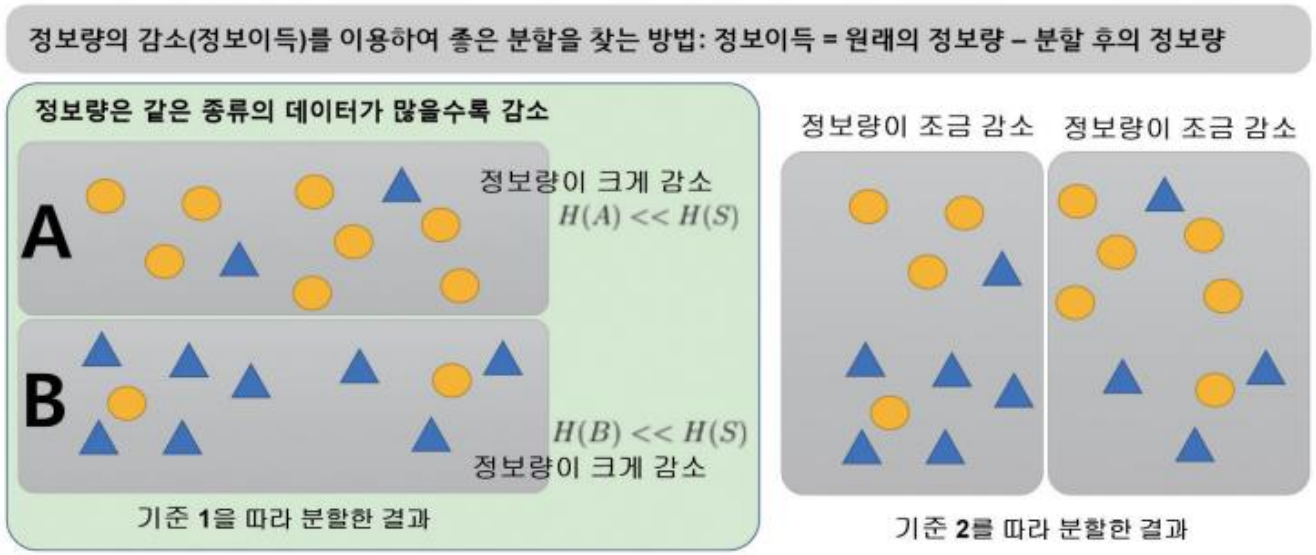


정보량 = 엔트로피

$$H(S) = -\left(\frac{1}{2}\lg\frac{1}{2} + \frac{1}{2}\lg\frac{1}{2}\right) = 1$$

결정트리

- **좋은 분할**
  - 정보가 잘 나누어져 있어야 함
  - 분할된 노드 각각에 한 종류의 데이터 비율이 높아져서 정보량이 줄어들어야 한다는 것.
  - 원래의 정보량에서 분할 후의 정보량을 뺀 값이 커질수록 좋다는 것이며 이것을 정보이득information gain이라고 정의



결정트리

- **데이터 표본의 엔트로피 측정**
  - 야구, 배구, 농구의 비율 : 1/5
  - 축구 비율 2/5

$$H(S) = -p^{+}\lg p^{+} - p^{-}\lg p^{-} \qquad H(S) = -\frac{2}{5}\lg\frac{2}{5} - 3\cdot\frac{1}{5}\lg\frac{1}{5}$$

데이터 표본 S
특성 – (ball, net, hand, goal) – 종목
데이터 1 – (작다, 없다, 0, 없다) – 야구
데이터 2 – (크다, 없다, 0, 지면) – 축구
데이터 3 – (크다, 없다, X, 지면) – 축구
데이터 4 – (크다, 있다, 0, 없다) – 배구
데이터 5 – (크다, 없다, 0, 공중) – 농구

- **정보이득 계산**

$$G(S, ball) = H(S) - \frac{4}{5}\left(-\frac{2}{4}\lg\frac{2}{4} - 2\cdot\frac{1}{4}\lg\frac{1}{4}\right) = 1.92 - 1.20 = 0.72$$

$$G(S, hand) = H(S) - \frac{4}{5}\left(-4\cdot\frac{1}{4}\lg\frac{1}{4}\right) = 1.92 - 1.60 = 0.32$$

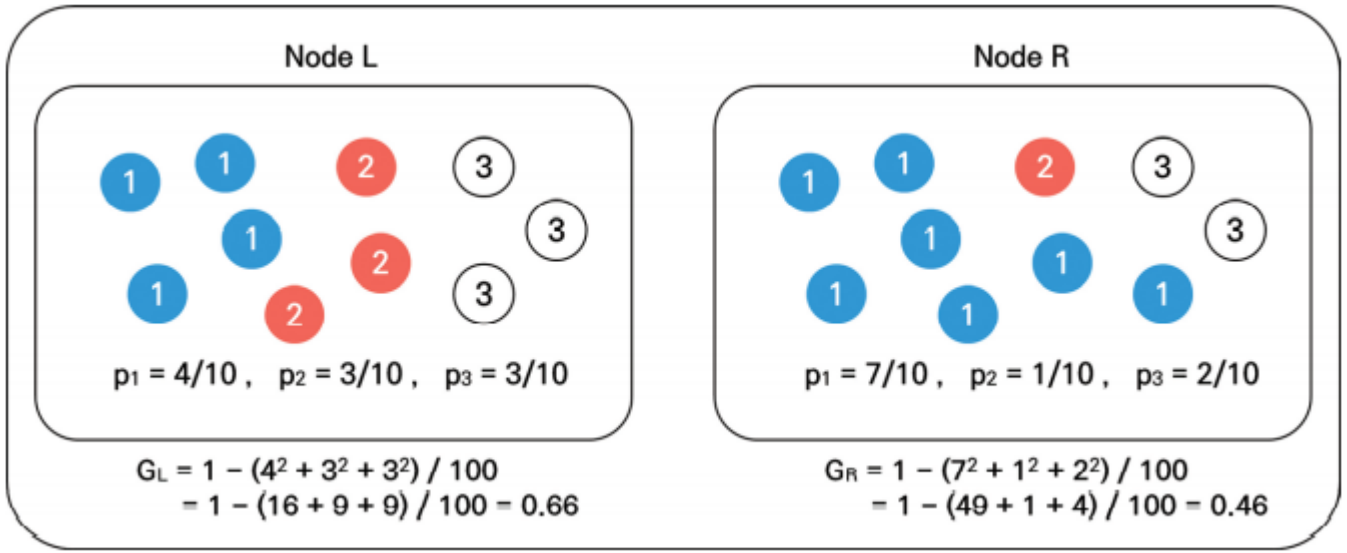
$$G(S, goal) = H(S) - \frac{2}{5}\left(-2\cdot\frac{1}{2}\lg\frac{1}{2}\right) = 1.92 - 0.4 = 1.52$$

중요속성은 goal

결정 트리

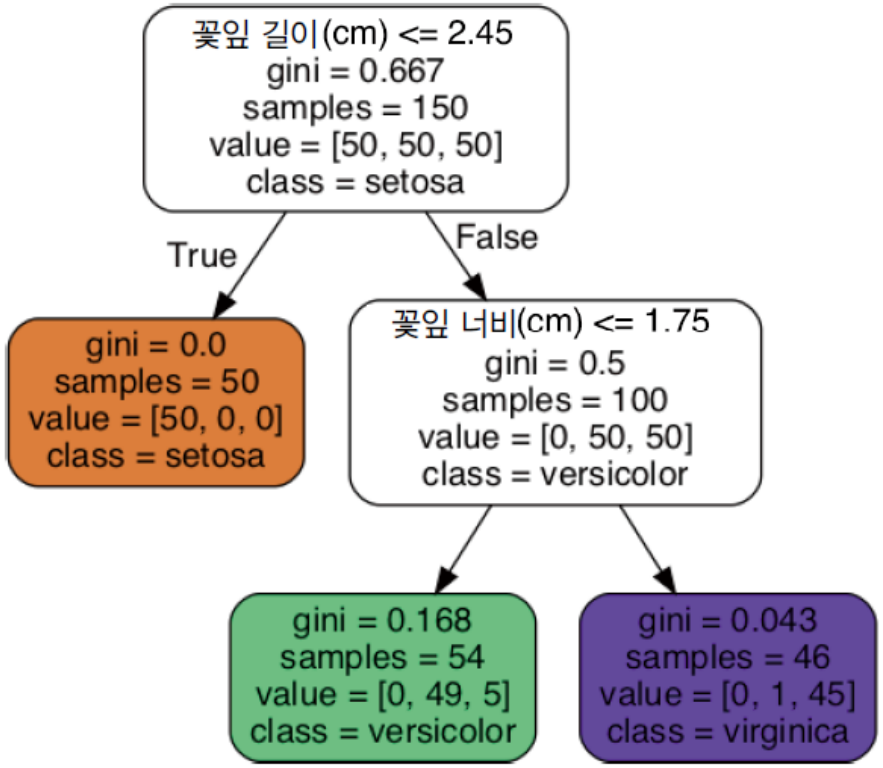
- **지니 불순도(CART 알고리즘)**
  - ID3 알고리즘은 엔트로피에 기반한 정보 이득 개념을 사용하는데, 이 개념은 다소 복잡한 수식과 계산을 요구
  - 이를 피하기 위해 많이 사용 하는 척도가 지니 불순도
  - 하나의 그룹 내에 섞여 있는 n개 종류의 레이블이 있고, 레이블이 i인 객체 비율이 p_i 라고 할 때 다음과 같은 값을 가짐

$$G = 1 - \sum_{i=1}^{n} p_i^2$$



결정트리

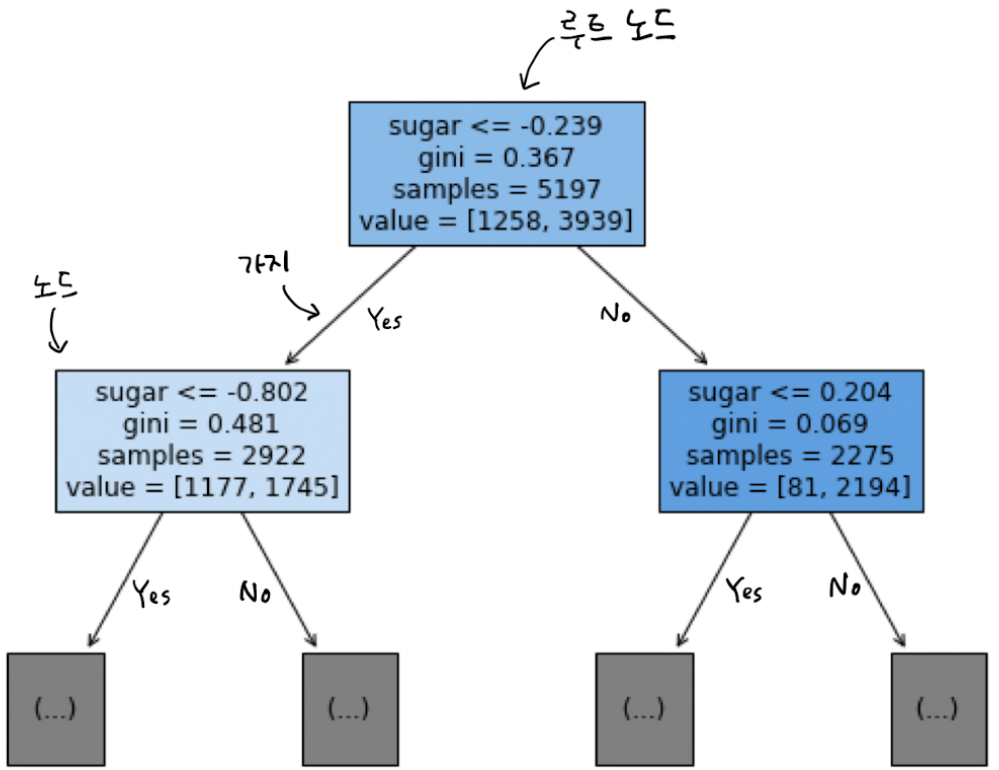
- **예시**
  - 꽃잎 길이, 꽃잎 너비에 따른 꽃 분류 문제



결정트리

- **예시**
  - RED or WHITE Wine 분류 문제

	alcohol	sugar	pH	class
count	6497.000000	6497.000000	6497.000000	6497.000000
mean	10.491801	5.443235	3.218501	0.753886
std	1.192712	4.757804	0.160787	0.430779
min	8.000000	0.600000	2.720000	0.000000
25%	9.500000	1.800000	3.110000	1.000000
50%	10.300000	3.000000	3.210000	1.000000
75%	11.300000	8.100000	3.320000	1.000000
max	14.900000	65.800000	4.010000	1.000000



결정트리

- **실습**
  - Github : https://github.com/hongrolee/Machine-Learning/blob/main/colab/%EC%99%80%EC%9D%B8%EB%B6%84%EB%A5%98_%EA%B2%B0%EC%A0%95%ED%8A%B8%EB%A6%AC.ipynb
  - Colab : https://colab.research.google.com/github/rickiepark/hg-mldl/blob/master/5-1.ipynb#scrollTo=Kt_biWBq_M-p

“천재는 노력하는 사람을 이길 수 없고, 노력하는 사람은 즐기는 사람을 이길 수 없다.”