머신러닝(Machine Learning)

의사결정트리, 앙상블 기법, 랜덤 포레스트

목 차

01 머신러닝(Machine Learning)

- 머신러닝(Machine Learning)은 지도학습과 비지도학습으로 나누어진다.
- 지도학습은 예측하려는 값이 존재하는 것이고, 비지도학습은 존재하지 않는다.
- 지도학습은 다시 회귀(regression)과 분류(classification)으로 나누어진다.

02 결정트리(decision tree)

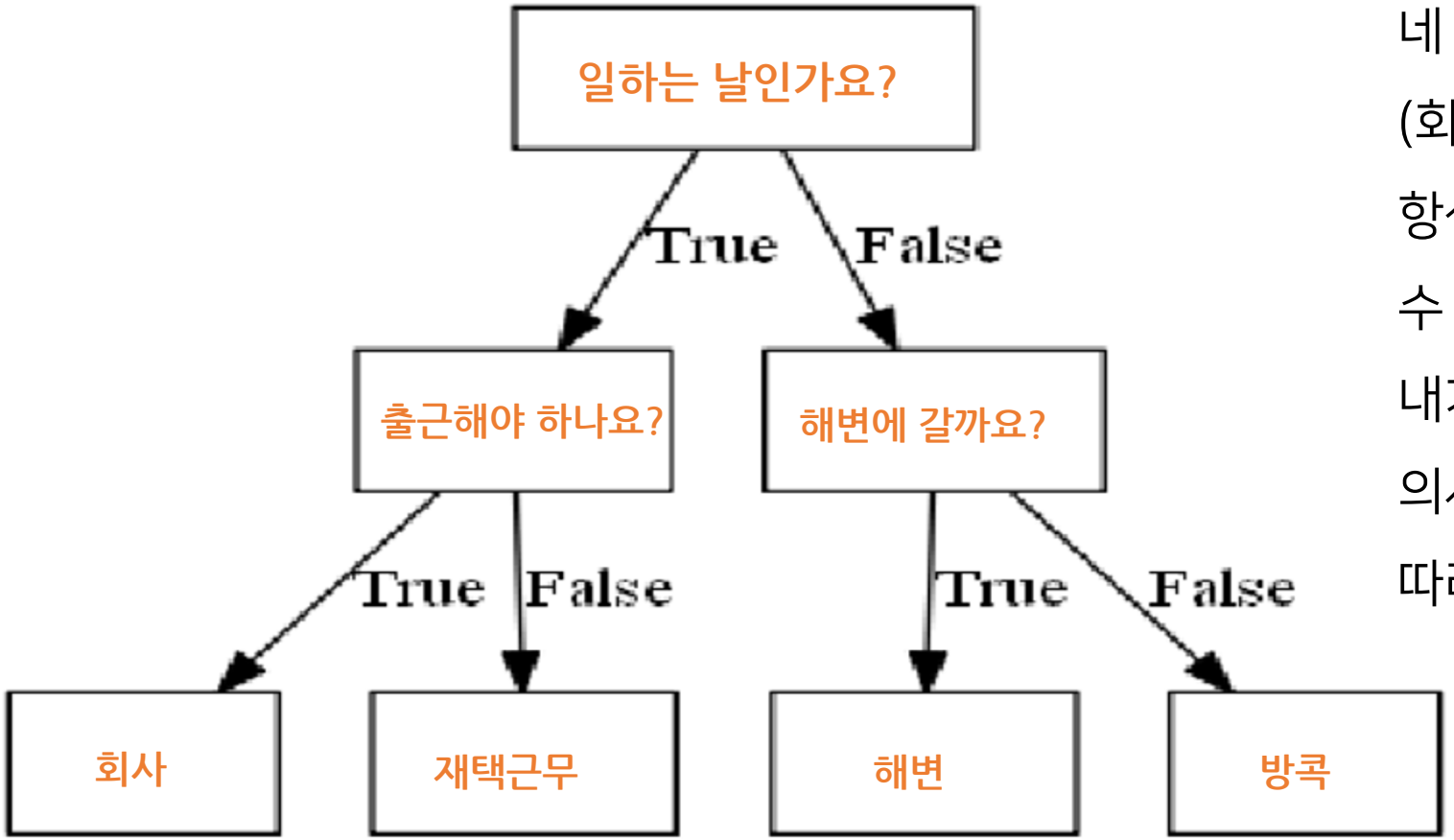
- **결정 트리는 분류와 회귀 문제에 널리 사용한다.**
- **결정트리는 과적합이 일어나기 쉬워, 이를 위해 가지치기 기법을 사용합니다.**

  (A) 사전 가지 치기

  (B) 사후 가지 치기

02 결정트리(decision tree)

▶ 결정 트리(decision tree)는 분류와 회귀 문제에 널리 사용하는 모델

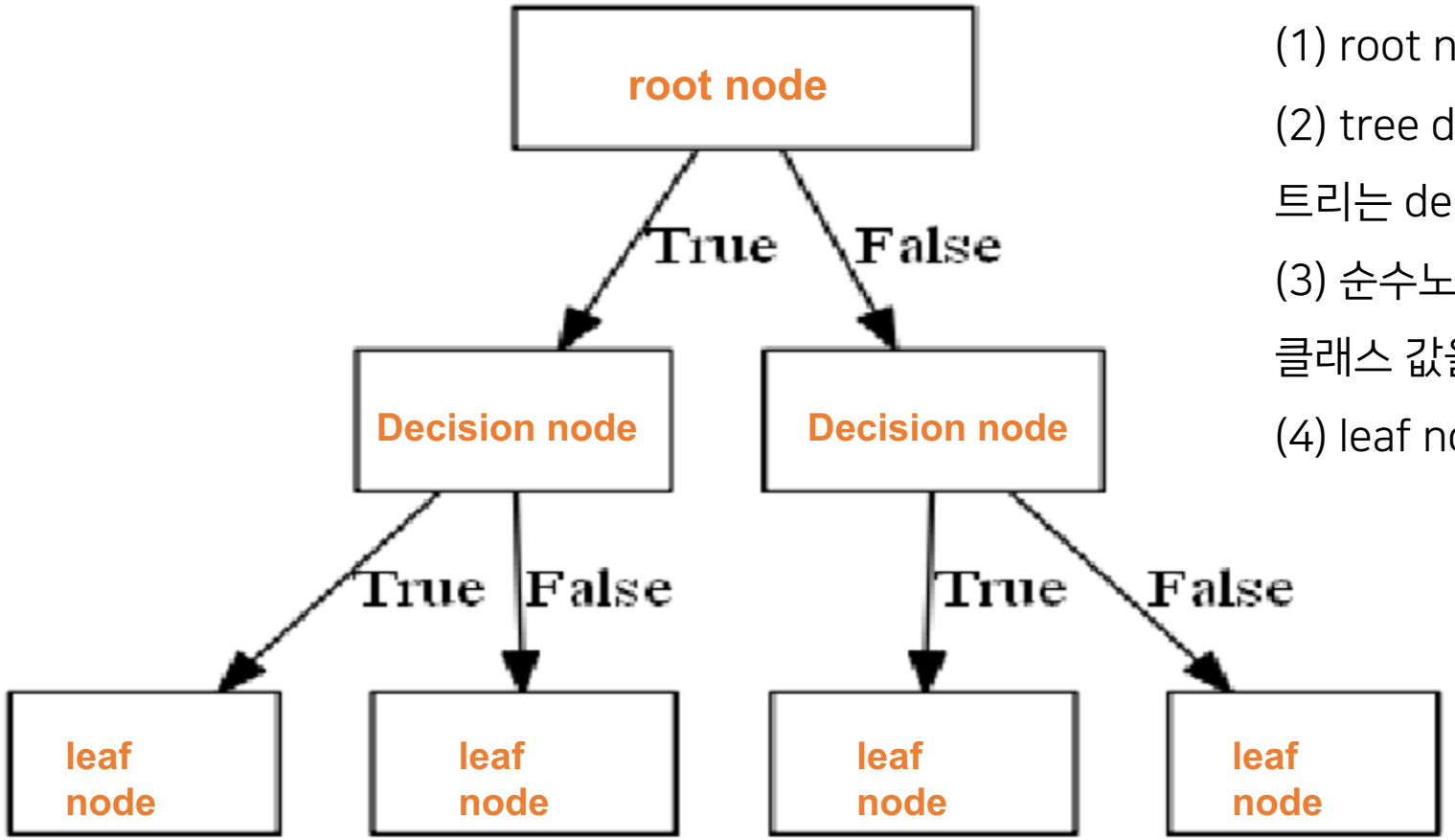


(그림 1-1) 분류의 문제에 대한 의사결정트리

네 개는 4개의 선택권이 있다.
(회사, 재택근무, 해변, 방콕)
항상 같은 고민을 매번 한다. 이를 해결할 수 있도록 빠른 결정을 위해 그동안의 내가 했던 데이터를 토대로 의사결정트리를 만들고, 이후부터는 따라서 행동해 보자.

02 결정트리(decision tree)

▶ 기본 용어 이해



(1) root node : 맨 위의 노드

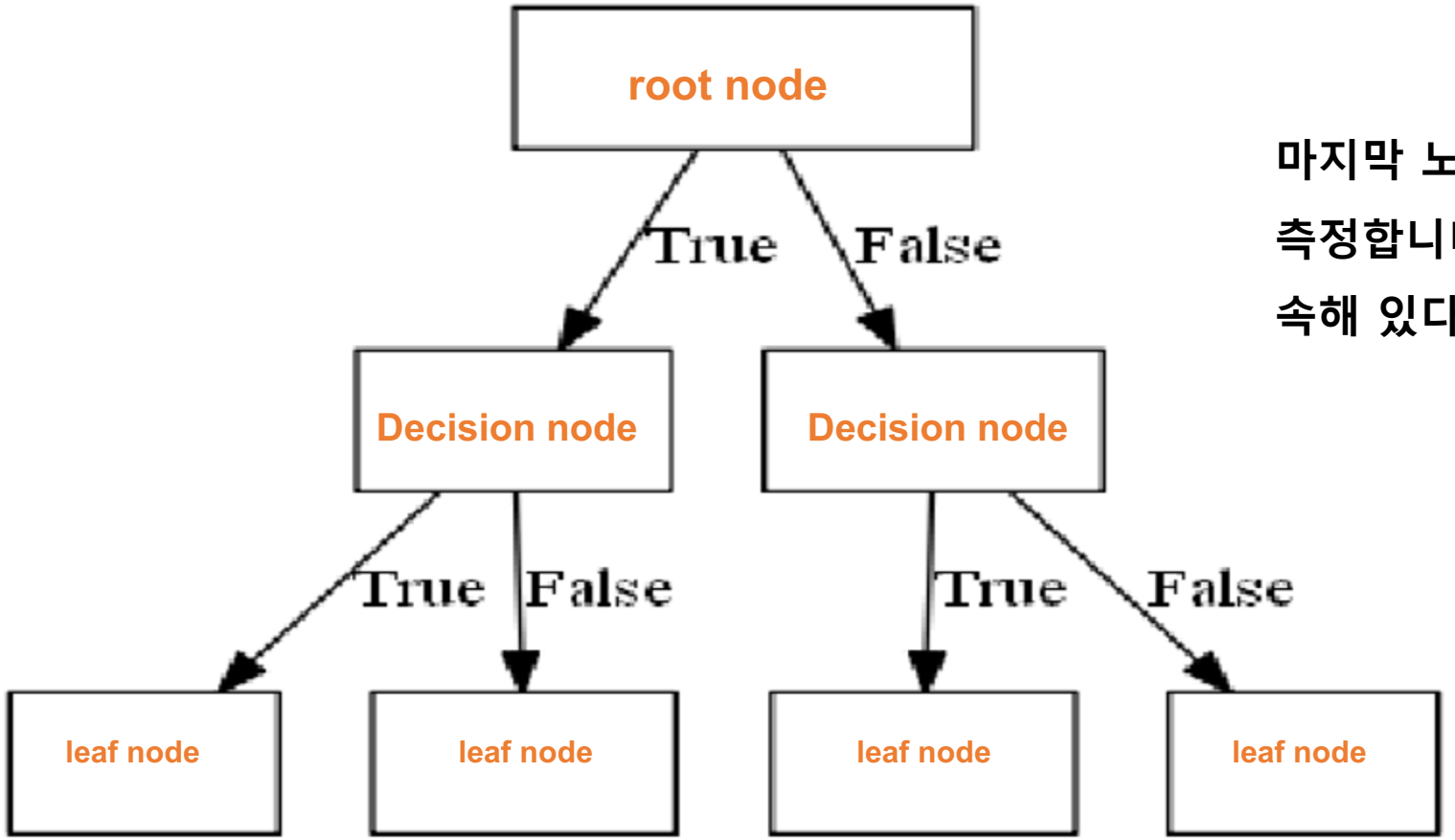
(2) tree depth : 분기를 최종 몇 단계까지 하는가? (옆의 트리는 depth가 2, 최상단은 0으로 시작.)

(3) 순수노드 : 데이터의 분할하여, 해당 데이터가 한 개의 클래스 값을 가질 때까지 반복

(4) leaf node : 줄기가 없는 맨 마지막 노드

02 결정트리(decision tree)

▶ 기본 용어 이해



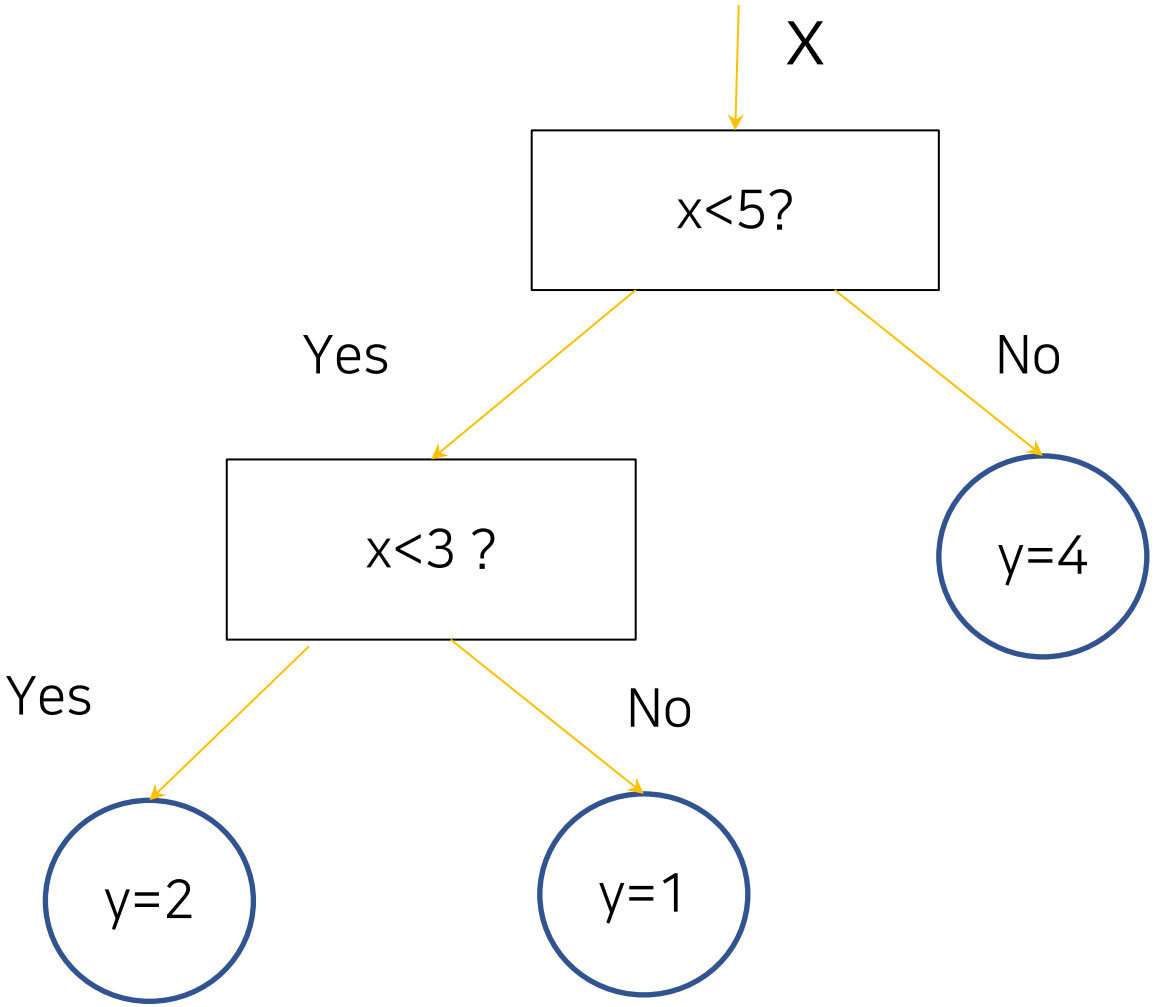
마지막 노드의 gini 속성은 불순도(impurity)를 측정합니다. 한 노드의 모든 샘플이 같은 클래스에 속해 있다면 이 노드를 순수(gini=0)하다고 합니다.

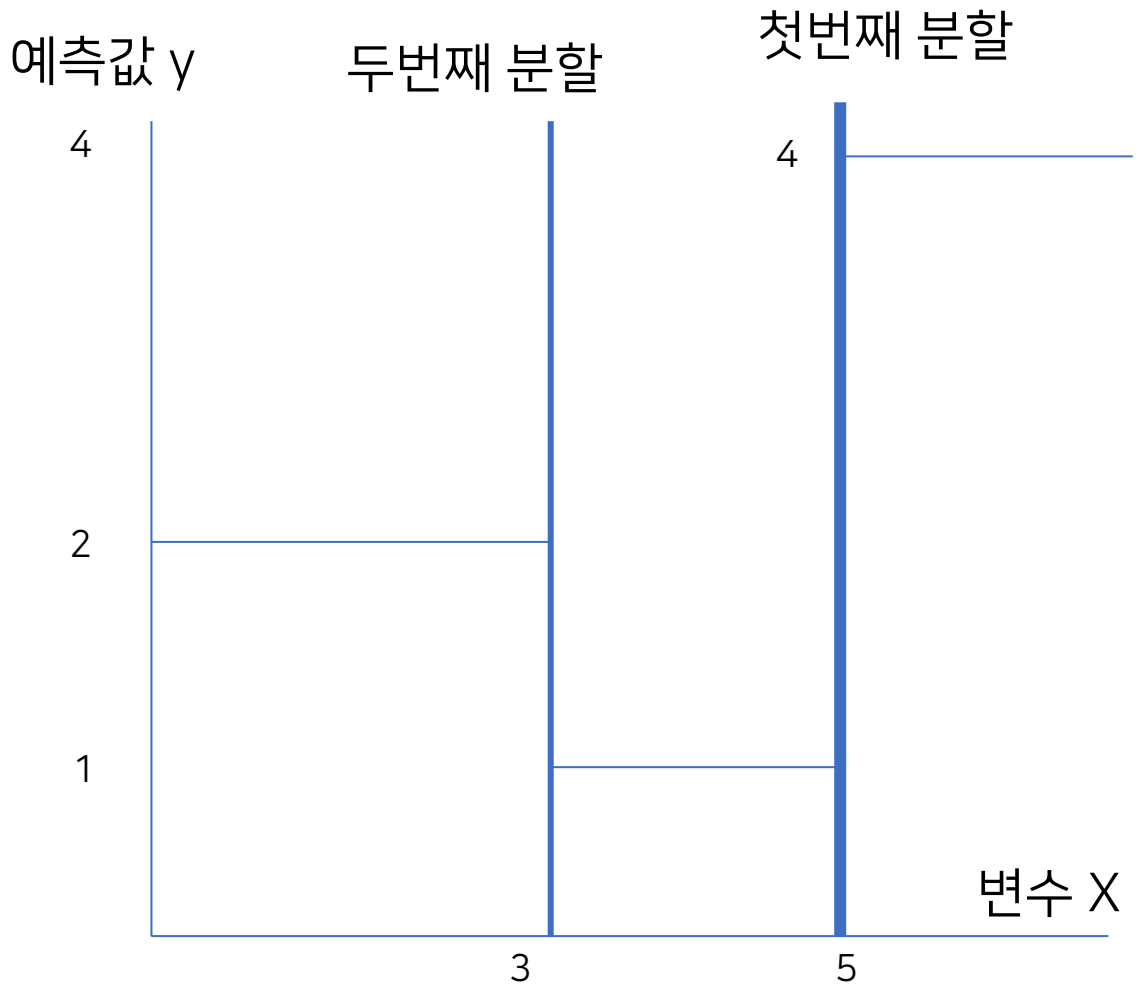
02 결정트리(decision tree)

▶ 이진분류의 결정 트리(decision tree)를 학습한다는 것은 정답에 가장 빨리 도달하는 예/아니오 질문 목록을 학습한다는 뜻이다.

02 결정트리(decision tree)-연속형

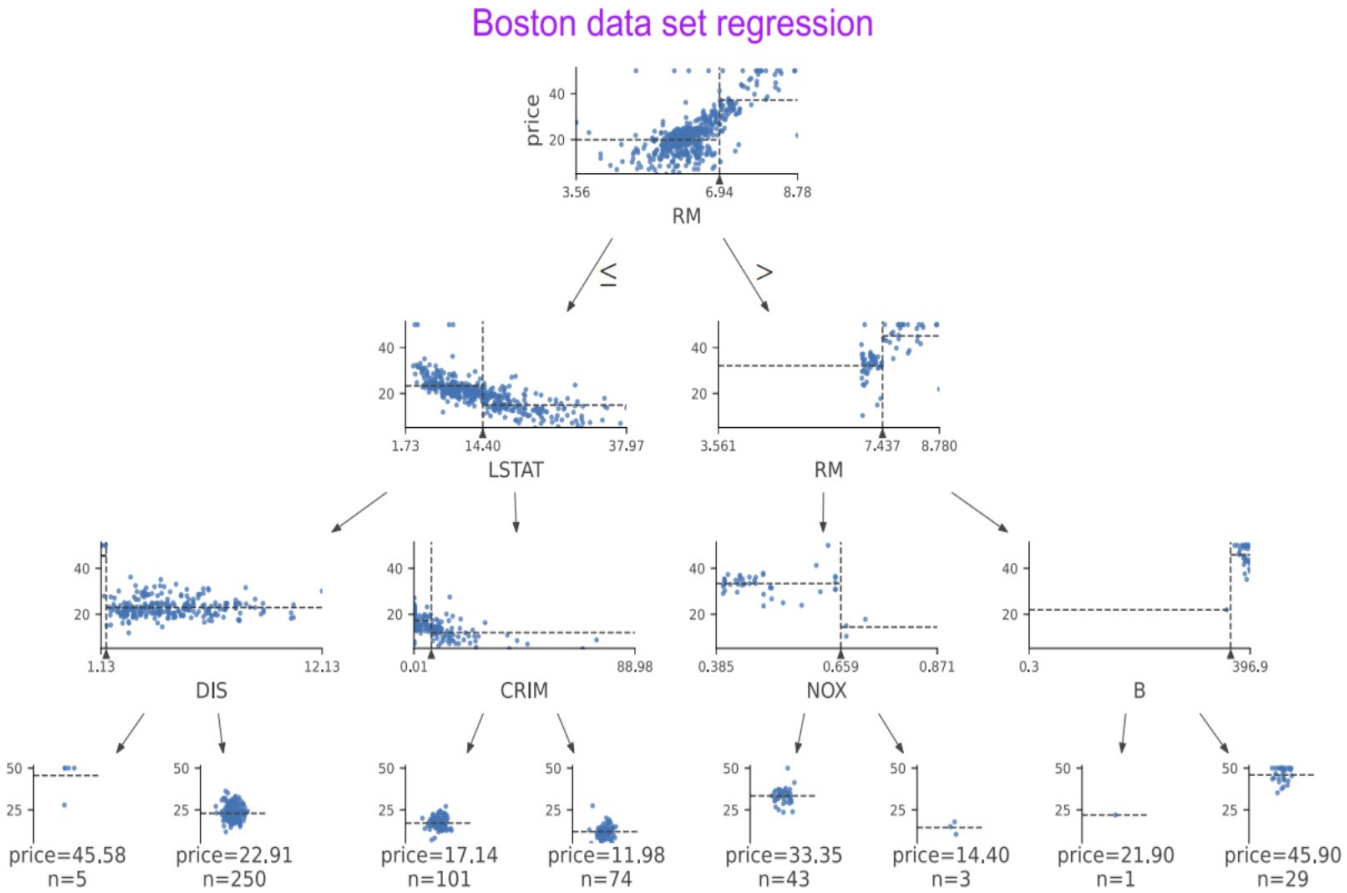
▶ **우리가 예측하려는 값이 연속형 값일 경우,**





02 결정트리(decision tree)-연속형

▶ 우리가 예측하려는 값이 연속형 값일 경우,

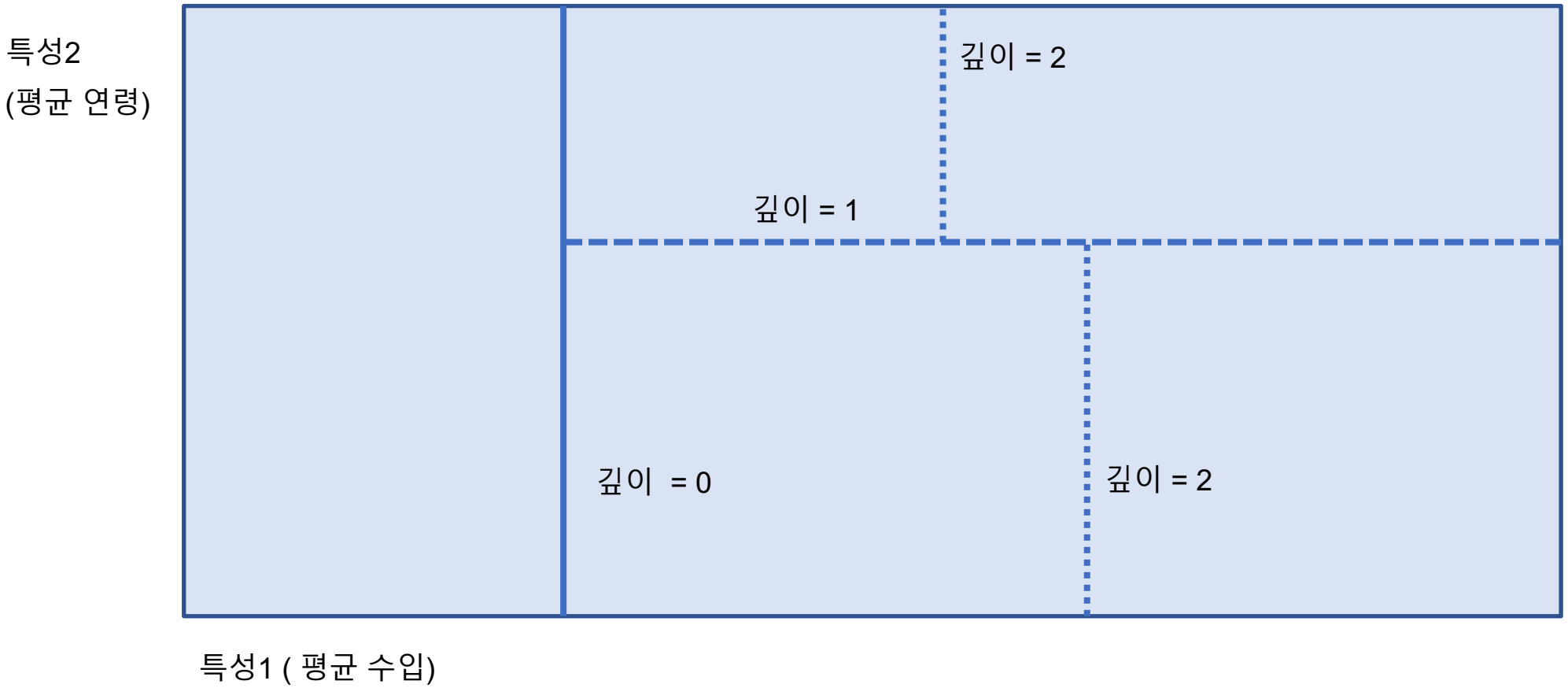


▶ 노드를 나누는 기준은 MSE값이 된다.

▶ leaf node의 데이터가 가르키는 target의 평균으로 예측을 하게 된다.

03 결정 트리의 결정 경계(decision boundary)

▶ 결정 경계(decision boundary)



04 지니 계수 및 엔트로피

▶ 데이터를 나누는 기준으로 기본적으로 지니 계수와 MSE가 사용된다.

▶ 머신러닝의 불순도의 측정 방법으로 지니계수, 엔트로피가 많이 사용됨.

▶ criterion 매개변수를 'entropy', 'gini'로 지정하여 엔트로피 및 지니계수로 불순도 지정 가능.

▶ 지니계수는 불순도를 측정하는 지표로서, 데이터의 통계적 분산 정도를 정량화하여 표현한 값.

▶ 엔트로피는 분자의 무질서함을 측정하는 것으로 열역학의 개념 분자가 안정되고 질서 정연할 경우 엔트로피가 0에 가깝다.

04 지니 계수 및 엔트로피

▶ 지니계수

$$G(S) = 1 - \sum_{i=1}^{c} {p_i}^2$$
$$p_{i,k} \neq 0$$

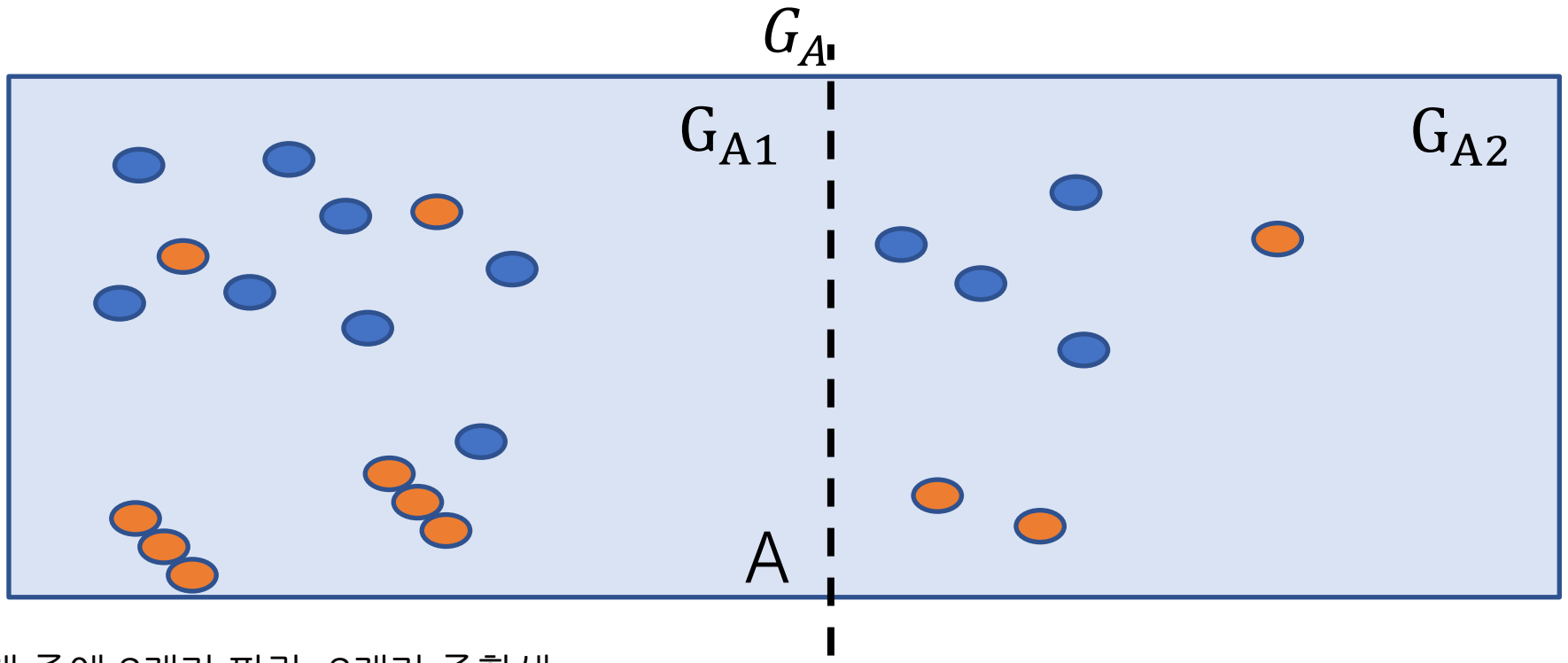
- s : 이미 발생한 사건 모음
- c : 사건의 갯수

Gini Index가 높을수록 데이터가 분산되어 있음을 의미.

▶ 엔트로피 식

$$H_i = -\sum_{k=1}^{n} p_{i,k} \log_2(p_{i,k})$$
$$p_{i,k} \neq 0$$

04 지니 계수 및 엔트로피



G_{A1}은 16개 중에 8개가 파란, 8개가 주황색

$$G_{A1} = 1 - (\frac{8}{16})^2 - (\frac{8}{16})^2 = 0.5$$

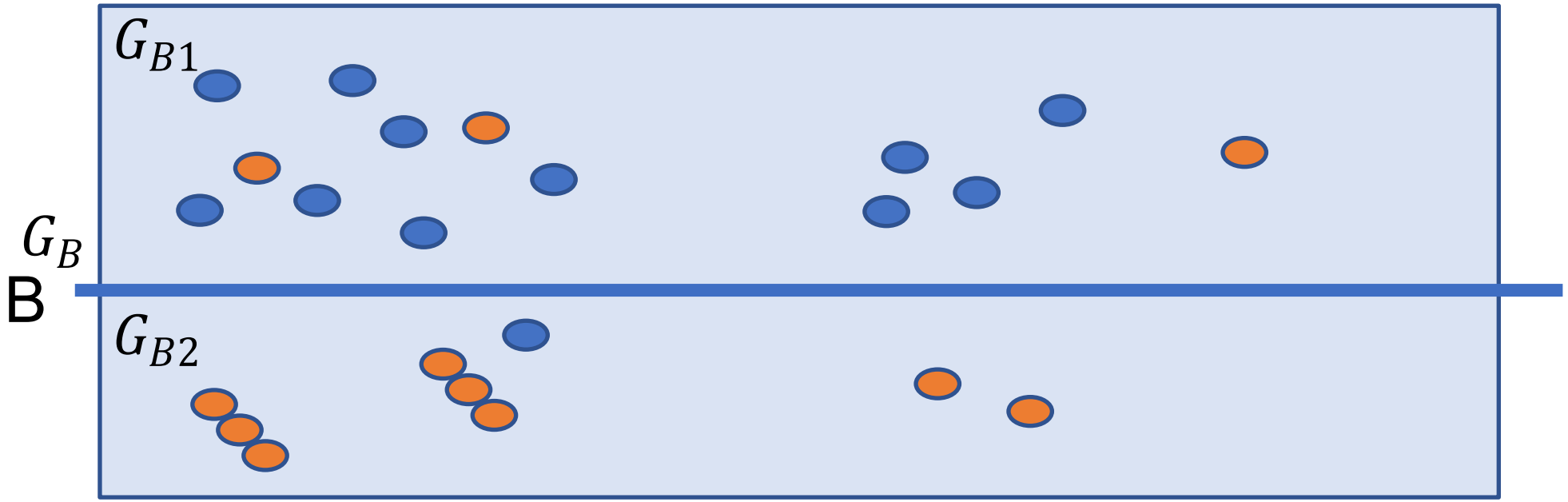
G_{A2}은 7개 중에 4개가 파란, 3개가 주황색

$$G_{A2} = 1 - (\frac{4}{7})^2 - (\frac{3}{7})^2 = 0.49$$

G_A는 다음과 같이 계산할 수 있다.

$$\text{지니계수 } G_A = (\frac{16}{23}) * 0.5 + (\frac{7}{23}) * 0.49 = 0.497$$

04 지니 계수 및 엔트로피



B를 기준으로 분할했을 때 지니계수

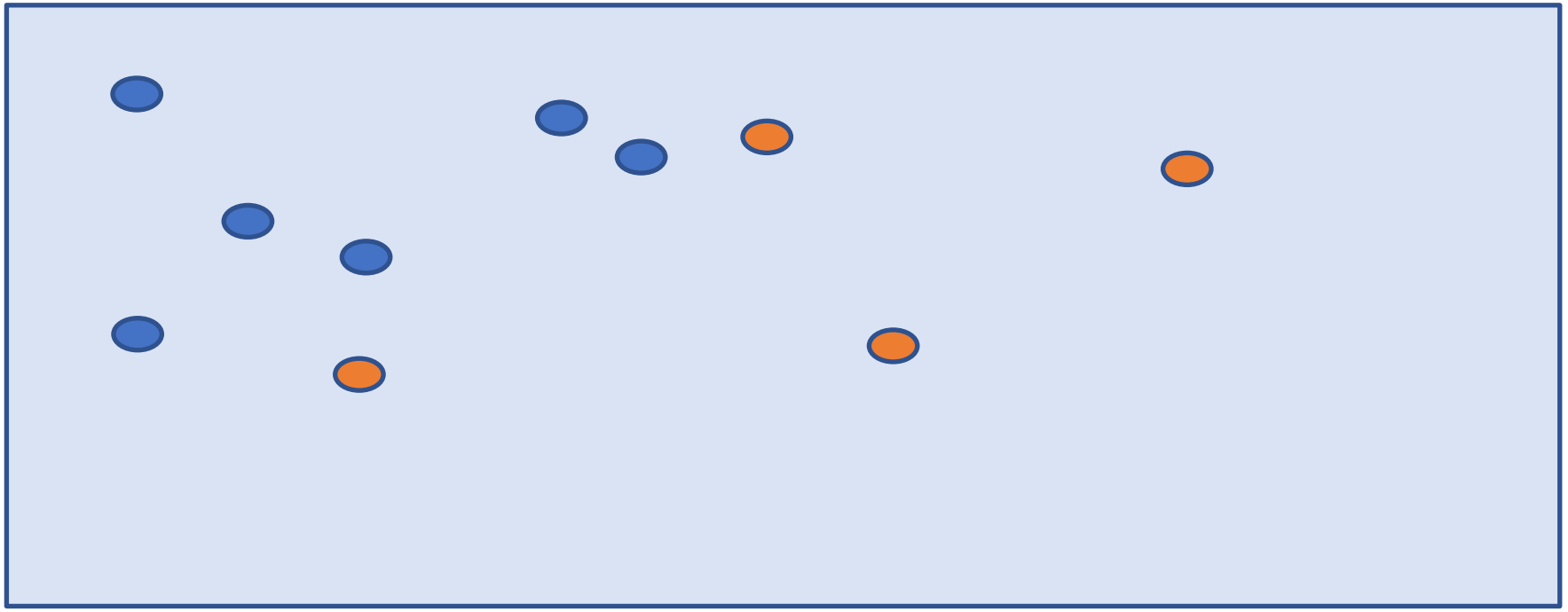
$G_{B1} = 1 - (\frac{11}{14})^2 - (\frac{3}{14})^2 = 0.34$

$G_{B2} = 1 - (\frac{1}{9})^2 - (\frac{8}{9})^2 = 0.2$

$G_B = (\frac{14}{23}) * 0.34 + (\frac{9}{23}) * 0.2 = 0.28$

04 지니 계수 및 엔트로피

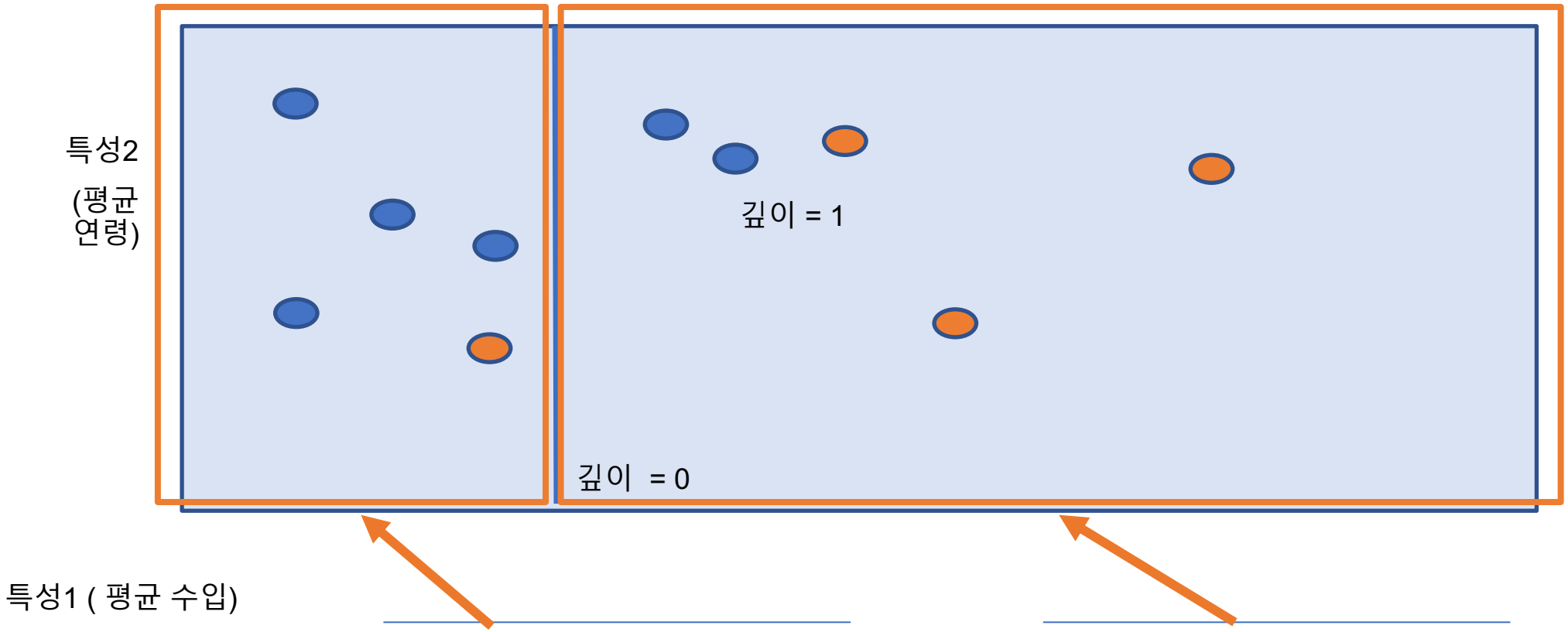
$$H_i = -\sum_{\substack{k=1 \\ p_{i,k} \neq 0}}^{n} p_{i,k} \log_2(p_{i,k})$$



전체 엔트로피 = $-\frac{6}{10}\log_2\left(\frac{6}{10}\right) - \frac{4}{10}\log_2\left(\frac{4}{10}\right)$ = 0.44218 + 0.52877 = 0.97095

04 지니 계수 및 엔트로피

$$H_i = -\sum_{\substack{k=1 \\ p_{i,k} \neq 0}}^{n} p_{i,k} \log_2(p_{i,k})$$



분할 후, 엔트로피 = $0.5 * \left(-\frac{1}{5}\log_2\left(\frac{1}{5}\right) - \frac{4}{5}\log_2\left(\frac{4}{5}\right)\right) + 0.5 * \left(-\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)\right)$

= 0.36097 + 0.48548 = 0.84645

05 장단점과 매개변수(하이퍼 파라미터)

▶ 모델의 복잡도를 조절하는 매개변수(하이퍼 파라미터)

max_depth, max_leaf_nodes, min_samples_leaf

(가) max_depth : 최대 tree의 depth

(나) max_leaf_nodes ： leaf의 최대 노드 개수 제한

(다) min_samples_leaf : 노드 분할을 위한 데이터 최소 개수 지정

05 장단점과 매개변수

▶ 장점

(1) 만들어진 모델을 쉽게 시각화 할 수 있어, 비전문가도 이해하기 쉽다.

(2) 각 특성이 개별적으로 처리되어, 데이터 분할시에 데이터 스케일의 영향을 받지 않음.

=> 특성의 정규화와 표준화 같은 전 처리 과정이 필요 없음.

=> 특성의 스케일(범위)가 다르거나 이진 특성과 연속적인 특성이 혼합되어 있을 때, 잘 작동

▶ 단점

(1) 과대적합(Overfitting)이 되는 경향이 있다.

=> 대안으로서 앙상블 방법을 사용함.(랜덤 포레스트, 그래디언트 부스팅)

06 앙상블 기법

▶ 앙상블(ensemble)는 여러 머신러닝 모델을 연결하여 더 강력한 모델을 만드는 기법

▶ 랜덤 포레스트(Random Forest)와 그래디언트 부스팅(gradient boosting)

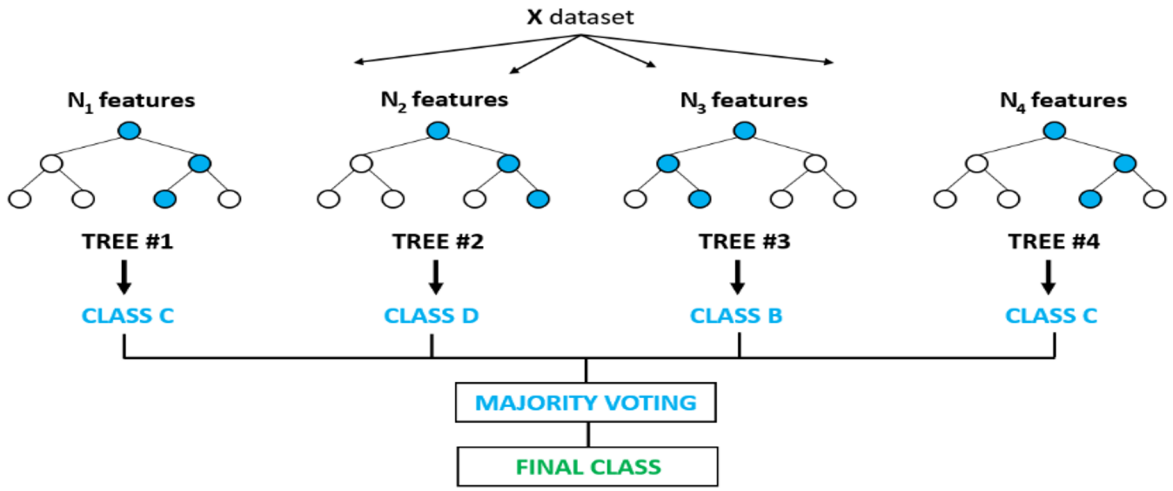
=> 둘 다 모델을 구성하는 기본 요소로 결정 트리를 사용.

06 앙상블 기법-랜덤 포레스트

▶ 결정 트리의 주요 단점 - 훈련 데이터에 과대 적합되는 경향이 있음.

=> 랜덤 포레스트 등장

▶ 아이디어 : 조금씩 다른 여러 결정 트리의 묶음.



http://bitly.kr/s2GOAgZm0 참조

06 앙상블 기법-랜덤 포레스트

▶ 결정 트리의 주요 원리

=> 잘 작동하되 서로 다른 데이터에 대해서 과대 적합된 트리를 많이 만들어 평균을 내면 과대적합이 줄어든다.

=> 수학적으로 증명됨.

(1) 타깃 예측을 잘 해야 함.

(2) 다른 트리와 구별됨.

=> A. 데이터 포인트를 무작위로 선택

=> B. feature(특성)을 무작위로 선택

07 랜덤 포레스트-장단점 매개변수

- **max_features : 전체 feature 중에 몇 개의 feature를 선택할지**

  => 기본값 분류 : sqrt(n_features), 회귀 : n_features

- **n_jobs : 멀티 코어 프로세서일 때는 사용할 코어 수를 지정**

- **n_estimators : 최대 몇 개의 트리를 사용할지.**

  => n_estimators는 클수록 좋다.

  (장점) 더 많은 트리를 평균하면 과대 적합을 줄여준다.

  (단점) 더 많은 트리는 더 많은 메모리와 긴 훈련시간

07 랜덤 포레스트-장단점

- (장점) 성능이 매우 뛰어나고 매개 변수 튜닝을 하지 않아도 잘 작동한다.
- (장점) 데이터의 스케일을 맞출 필요가 없음.
- (단점) 텍스트 데이터 같이 매우 차원이 높고 희소한 데이터는 잘 작동하지 않음.
- (단점) 선형 모델보다 많은 메모리를 사용한다.

