



DecisionTree

의사결정나무

15기 분석 이윤정

CONTENTS

01 머신러닝?!

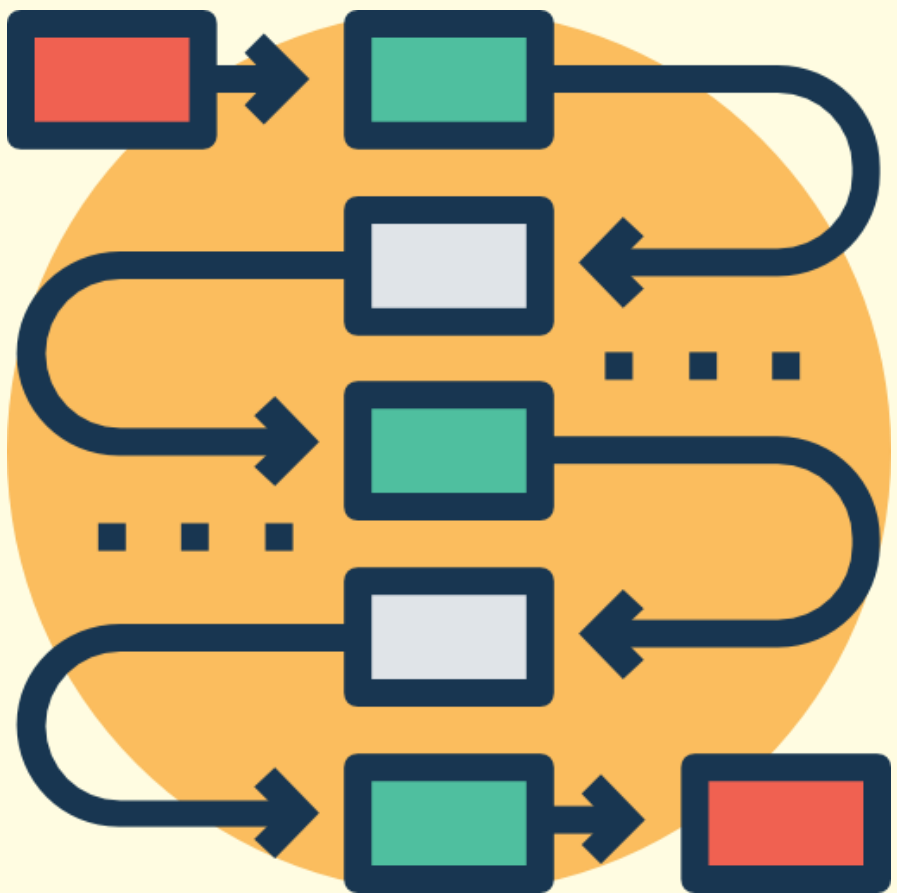
02 의사결정나무란?

03 의사결정나무 모델 개요

04 의사결정나무의 형성 과정

05 의사결정나무의 장/단점

01 머신러닝 ?!



01 머신러닝 ?!

01 사람이 감독하는가 ?

지도, 비지도, 준지도, 강화학습

02 실시간으로 점진적인 학습을 하는가?

온라인, 배치

03 예측 모델을 만드는가?

사례기반, 학습 모델 기반

01 머신러닝 ?!

01 사람이 감독하는가 ?

지도, 비지도, 준지도, 강화학습

02 실시간으로 점진적인 학습을 하는가?

온라인, 배치

03 예측 모델을 만드는가?

사례기반, 학습 모델 기반

01

머신러닝 ?!

01

사람이 감독하는가 ?

지도, 비지도, 준지도, 강화학습



02 의사결정나무란 ?

나무구조로 도표화하여 분류 및 예측을 수행하는 머신러닝 알고리즘



02 의사결정나무란 ?

특징

1. 분류와 회귀가 모두 가능한 머신러닝 알고리즘

분류) DecisionTreeClassifier


회귀) DecisionTreeRegressor

2. 질문을 던져서 맞고 틀리는 것에 따라 대상을 좁혀나감.

즉, 스무고개 놀이와 그 원리가 유사함.

3. RandomForest의 기본 구성 요소





안녕하세요, 아키네이터입
니다

실제, 소설, 만화 인물을 생
각하십시오.
맞추어 보겠습니다

akinator®

03 의사결정나무 모델 개요



03 의사결정나무 모델 개요



03 의사결정나무 모델 개요



데이터가 균일해지도록 분할

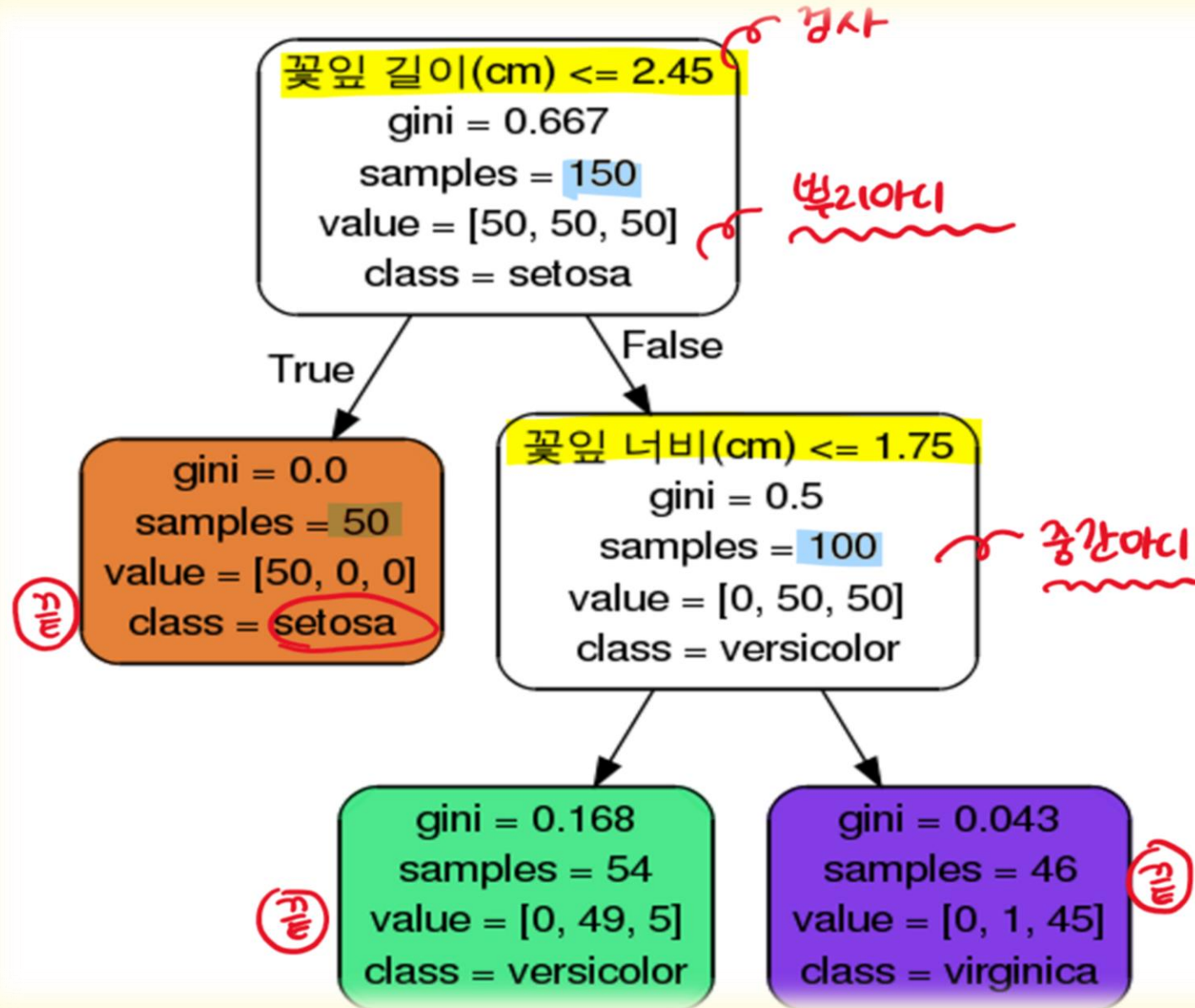
분류 비슷한 범주를 가지고 있는 관측치끼리 균일해지도록

회귀 비슷한 수치를 가지고 있는 관측치끼리 균일해지도록

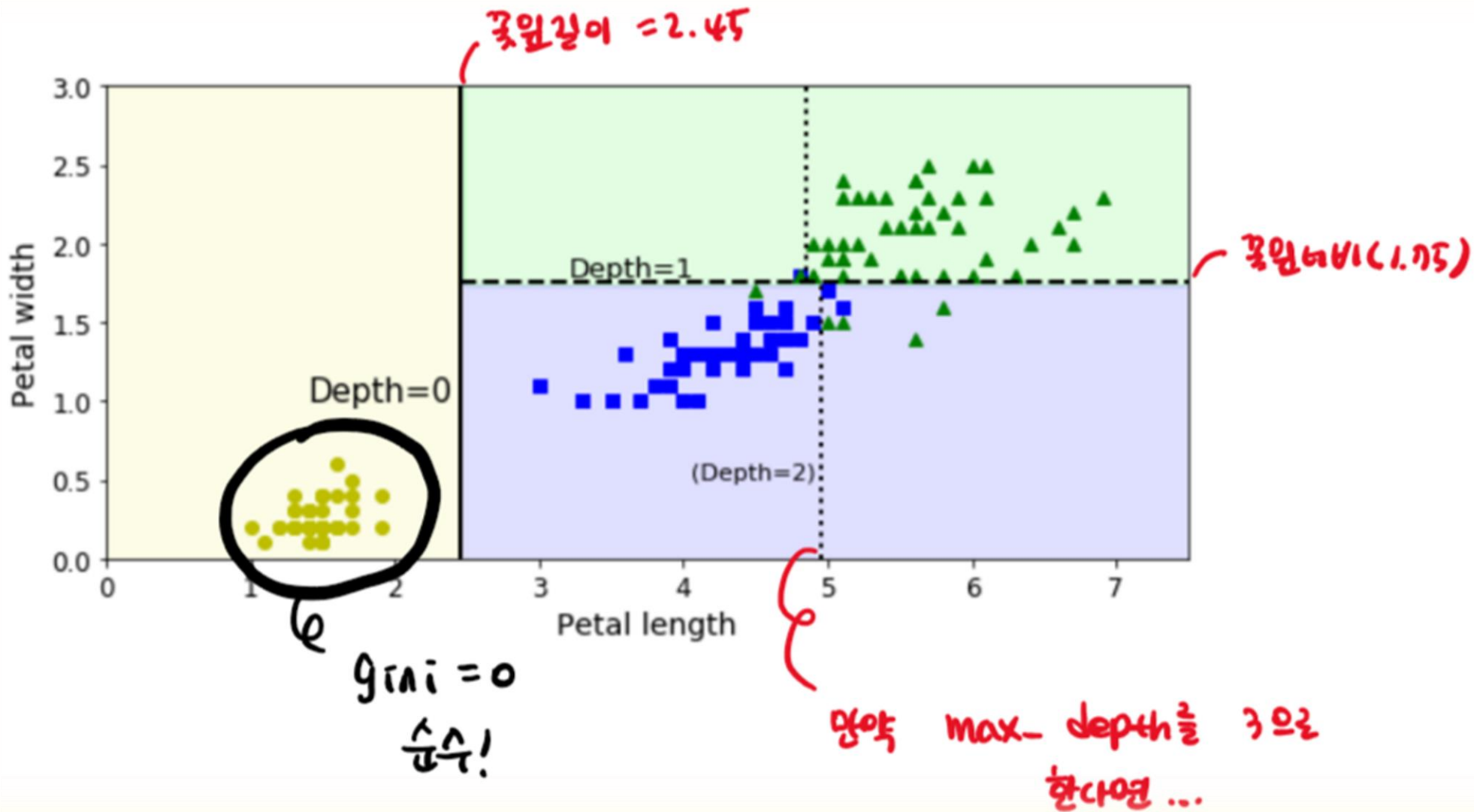
03 의사결정나무 모델 개요



03 의사결정나무 모델 개요



03 의사결정나무 모델 개요



04 의사결정나무의 형성 과정

적절한 ‘분할규칙’ 과 ‘정지규칙’ 을 지정하고 ‘예측값’ 을 할당

분리규칙

정지규칙

가지치기

예측값
할당

04 의사결정나무의 형성 과정



분리규칙

정지규칙

가지치기

예측값
할당

- 부모마디에서 자식 마디를 생성하는 기준
- 순도 / 불순도에 의해 목표 변수 구별

03 의사결정나무 모델 개요

CART 훈련 알고리즘

classification and regression tree

[기본원리]

훈련 세트를 하나의 특성 k 의 임계값 t_k 을 사용해 서브셋으로 나누기

[분류에 대한 CART 비용 함수]

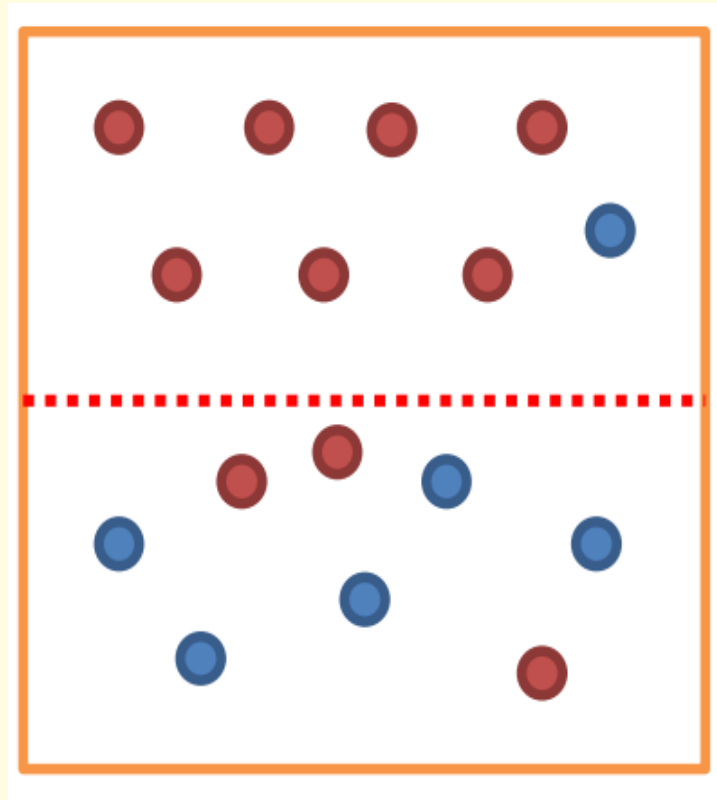
$$J(k, t_k) = \frac{m_{\text{left}}}{m} G_{\text{left}} + \frac{m_{\text{right}}}{m} G_{\text{right}}$$

m : 전체 샘플 수 (m_{left} : 왼쪽 서브셋의 불순도, m_{right} : 오른쪽 서브셋의 불순도)

G : 불순도 (G_{left} : 왼쪽 샘플 수, G_{right} : 오른쪽 샘플 수)

03 의사결정나무 모델 개요

순도 / 불순도 / 지니계수



불순도가 낮음

순도가 높음

지니계수 낮음 (0에 가까움)

불순도가 높음

순도가 낮음

지니계수 높음



불순도를 최소화하는 방향으로 학습을 진행

03 의사결정나무 모델 개요

엔트로피

[정보이론]

- : 데이터를 정량화하기 위한 응용수학의 분야 중 하나
- : 정보량이 높다 = 어떤 일이 일어날 확률이 낮다. 불확실하다.
- : 이때, 정보량이 높은 문장이 맞을수록 해당 정보의 중요도는 높아짐.

$$I(x) = -\log P(x)$$

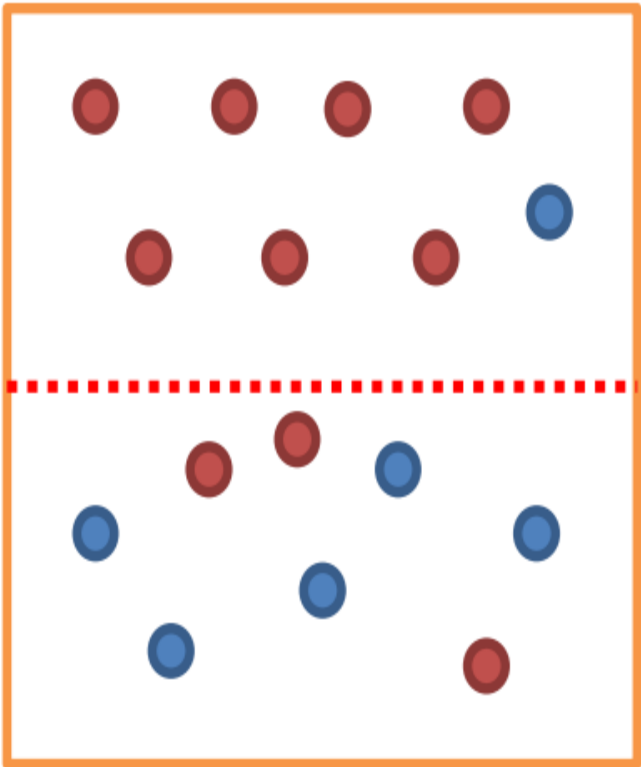
[엔트로피]

- : 정보량의 평균
- : 분자의 무질서함을 측정하는 개념

$$Entropy(A) = - \sum_{k=1}^m p_k \log_2 (p_k)$$

03 의사결정나무 모델 개요

엔트로피



$$Entropy(A) = - \sum_{k=1}^m p_k \log_2 (p_k)$$

$$Entropy(A) = -\frac{10}{16} \log_2 \left(\frac{10}{16} \right) - \frac{6}{16} \log_2 \left(\frac{6}{16} \right) \approx 0.95$$

분할 전



$$Entropy(A) = \sum_{i=1}^d R_i \left(- \sum_{k=1}^m p_k \log_2 (p_k) \right)$$

분할 후

$$Entropy(A) = 0.5 \times \left(-\frac{7}{8} \log_2 \left(\frac{7}{8} \right) - \frac{1}{8} \log_2 \left(\frac{1}{8} \right) \right) + 0.5 \times \left(-\frac{3}{8} \log_2 \left(\frac{3}{8} \right) - \frac{5}{8} \log_2 \left(\frac{5}{8} \right) \right) \approx 0.75$$

04 의사결정나무의 형성 과정

분리규칙

정지규칙

가지치기

예측값
할당

[더 이상 분리가 일어나지 않는 기준]

- 1) 더 이상 분리해도 불순도가 줄어들지 않을 때
- 2) 자식 마디에 남은 sample 수가 너무 적을 때
- 3) 분석자가 지정한 규제 매개변수에 도달했을 때

04 의사결정나무의 형성 과정



규제 매개변수란 ?

결정트리는 대체로 데이터에 대한 제약사항이 매우 적음.

하지만, 오버피팅의 위험성이 커서 규제 매개변수를 활용해 제약을 걸어주는 것이 좋음.

max_depth : 최대 깊이 설정

min_samples_split : 분할되기 위해 노드가 가져야하는 최소샘플 수

min_samples_leaf : 리프 노드가 가지고 있어야하는 최소 샘플 수

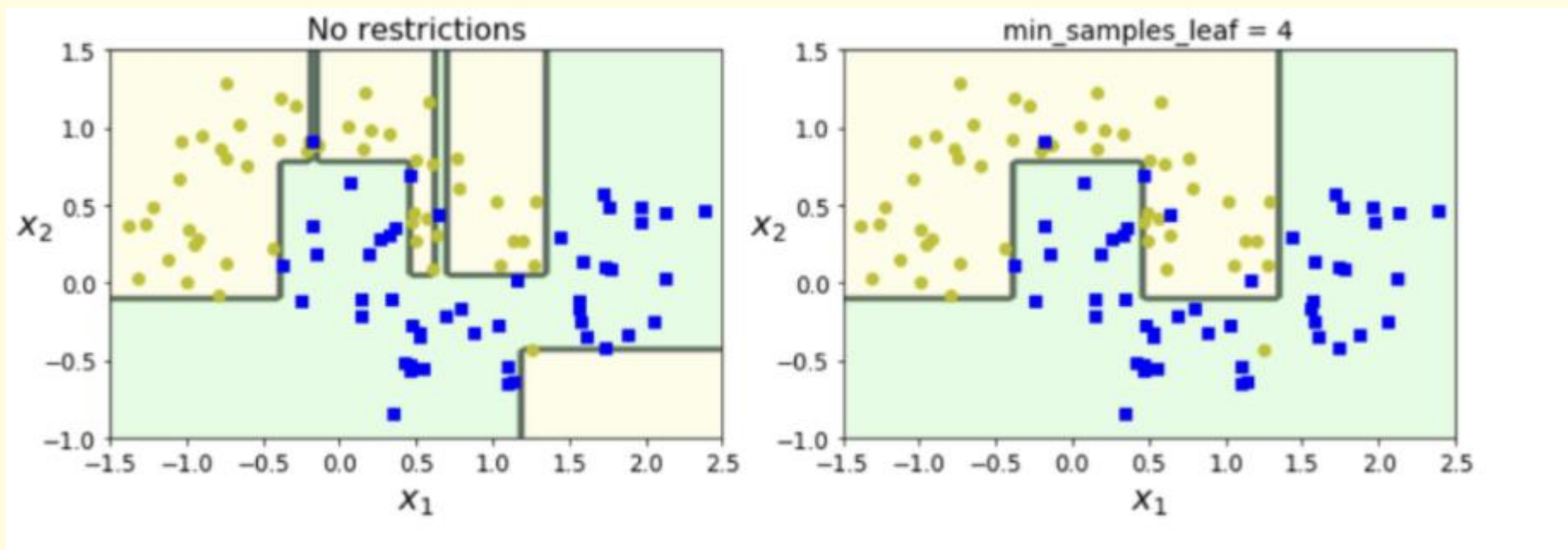
max_leaf_nodes : 리프 노드의 최대 수

max_features : 각 노드에서 분할에 사용할 특성의 최대 수

04 의사결정나무의 형성 과정



규제 매개변수란 ?



왼쪽은 규제 없는 상태. 오른쪽은 리프노드의 최소 샘플 수를 4개로 제한.

→ 오른쪽은 과적합을 피함.

04 의사결정나무의 형성 과정

분리규칙

정지규칙

가지치기

예측값
할당

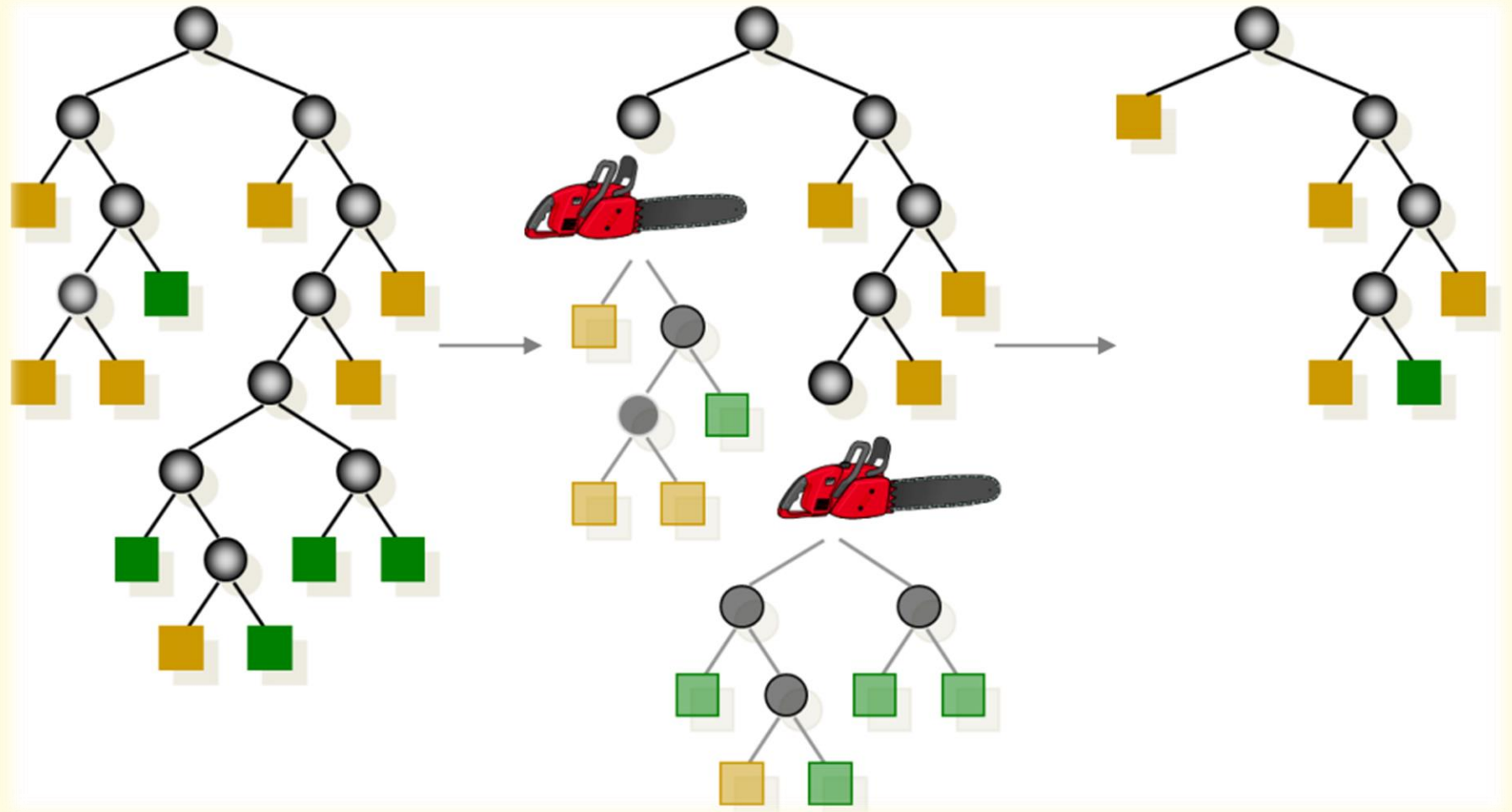
[부적절한 마디를 잘라내 모양을 단순화]

: depth가 깊어질수록 오버피팅의 위험성이 높음

: 불필요한(부적절한) 마디를 제거하는 과정

: 데이터를 버리는 것이 아닌 합치는 과정 (merge 개념)

04 의사결정나무의 형성 과정



04 의사결정나무의 형성 과정

분리규칙

정지규칙

가지치기

예측값
할당

– 예측값 할당

(분류 : class 예측 vs 회귀 : 특정값 예측)

- 타당성 평가 (cross validation 등을 통해 트리 모델 평가)
- 해석 및 예측 (생성한 tree에 새로운 데이터 대입 => 확인)

05 의사결정나무의 장/단점



- 직관적
 - 이상치, 노이즈 큰 영향 x
 - 높은 모델 해석력
 - 연속형 데이터, 범주형 데이터 모두 처리 가능
 - '균일도'에만 초점 가능
- (스케일링, 정규화 등의 과정 불필요)



- 일반화가 어려움 (불안전성)
: 학습데이터에 따른 차이 큼
= 모델 variance가 높음
- 오버피팅의 가능성이 매우 높음



Random Forest의 등장

- 직관적
 - 이산치, 노이즈 큰 영향 x
 - 연속형 데이터, 범주형 데이터 모두 처리 가능
 - '균일도'에만 초점 가능
- (스케일링, 정규화 등의 과정 불필요)
- (트리들에서 만든 예측을 평균 -> 불안정성 극복)

A decorative background featuring stylized trees. On the left, there is a large green circle, a smaller orange circle, and a green triangle. On the right, there is a large yellow circle, a green circle, a smaller green circle, and a green triangle. All trees have brown trunks and are positioned above a horizontal line.

감사합니다

15기 분석 이윤정