6. 결정 트리

정윤서

6-6

지니 불순도 또는 엔트로피?

지니 불순도 또는 엔트로피?

- DecisionTreeClassifier의 criterion 매개변수의 기본값은 "gini“
- criterion 매개변수를 "entropy"로 지정

  → 엔트로피 불순도 사용 가능

엔트로피

- 분자의 무질서함을 측정하는 것 (열역학적 개념)
- 분자가 안정되고 질서 정연하면 엔트로피가 0에 가깝다.

- 메시지의 평균 정보 양을 측정하는 **섀넌의 정보 이론**

  → 모든 메시지가 동일할 때 엔트로피가 0이 된다.

엔트로피

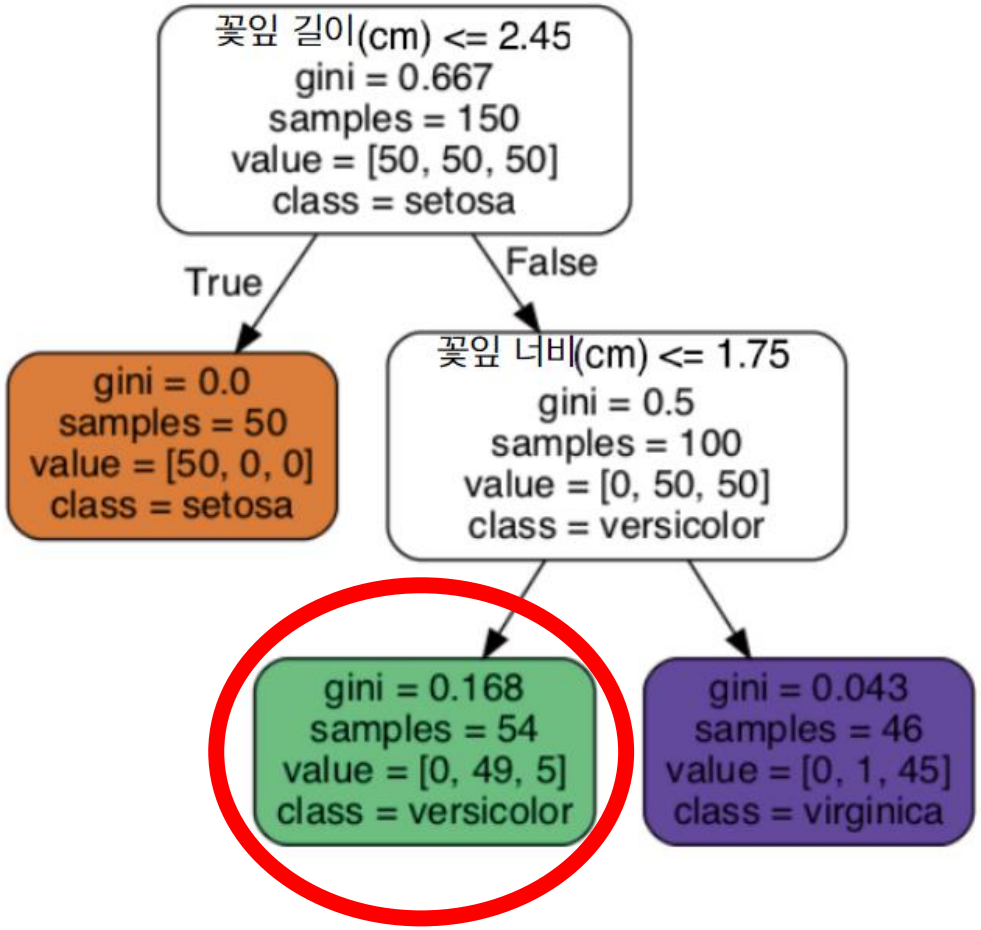
- **머신러닝**에서는 **<u>불순도</u>**의 측정 방법

  → 어떤 세트가 한 클래스의 샘플만 담고 있다면 엔트로피가 0이다.

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

i번째 노드의 엔트로피

엔트로피



$$-\frac{49}{54}\log_2\left(\frac{49}{54}\right)-\frac{5}{54}\log_2\left(\frac{5}{54}\right)\approx 0.445$$

지니 불순도 또는 엔트로피?

- 둘 다 비슷한 트리를 만든다.
- 지니 불순도가 조금 더 계산이 빠르기 때문에 기본값으로 좋다.
- 다른 트리가 만들어지는 경우 **지니 불순도**가 가장 빈도 높은 클래스를 **한쪽 가지로 고립**시키는 경향이 있는 반면 **엔트로피**는 **조금 더 균형 잡힌 트리**를 만든다.

6-7

규제 매개변수

규제 매개변수

- 결정 트리는 훈련 데이터에 대한 제약 사항이 거의 없다.
- 반대로 선형 모델은 데이터가 선형일 거라 가정한다.
- 제한을 두지 않으면 트리가 훈련 데이터에 아주 가깝게 맞추려고 해서 대부분 과대적합되기 쉽다.

규제 매개변수

비파라미터 모델

- 결정 트리는 훈련되기 전에 파라미터 수가 결정되지 않는다.
- 모델 구조가 데이터에 맞춰져서 고정되지 않고 자유롭다.

파라미터 모델

- 미리 정의된 모델 파라미터 수를 가지므로 자유도가 제한된다.
- 과대적합될 위험이 줄어든다.
- 하지만 과소적합될 위험은 커진다.

규제 매개변수

- 학습할 때 결정 트리의 자유도를 제한
  - 훈련 데이터에 대한 과대적합을 피하기 위함

  → ***규제***

규제 매개변수

- 규제 매개변수는 보통 적어도 결정 트리의 최대 깊이는 제어할 수 있다.
- 사이킷런에서는 max_depth 매개변수로 이를 조절
  - max_depth를 줄이면 모델을 규제하게 되고 과대적합의 위험이 감소

DecisionTreeClassifier

- min_samples_split(분할되기 위해 노드가 가져야 하는 최소 샘플 수)
- min_samples_leaf(리프 노드가 가지고 있어야 할 최소 샘플 수)
- min_weight_fraction_leaf (min_samples_leaf와 같지만 가중치가 부여된 전체 샘플 수에서의 비율)
- max_leaf_nodes (리프 노드의 최대 수)
- max_features(각 노드에서 분할에 사용할 특성의 최대 수)

규제 매개변수

기본 매개변수(규제 없음)　　min_samples_leaf=4

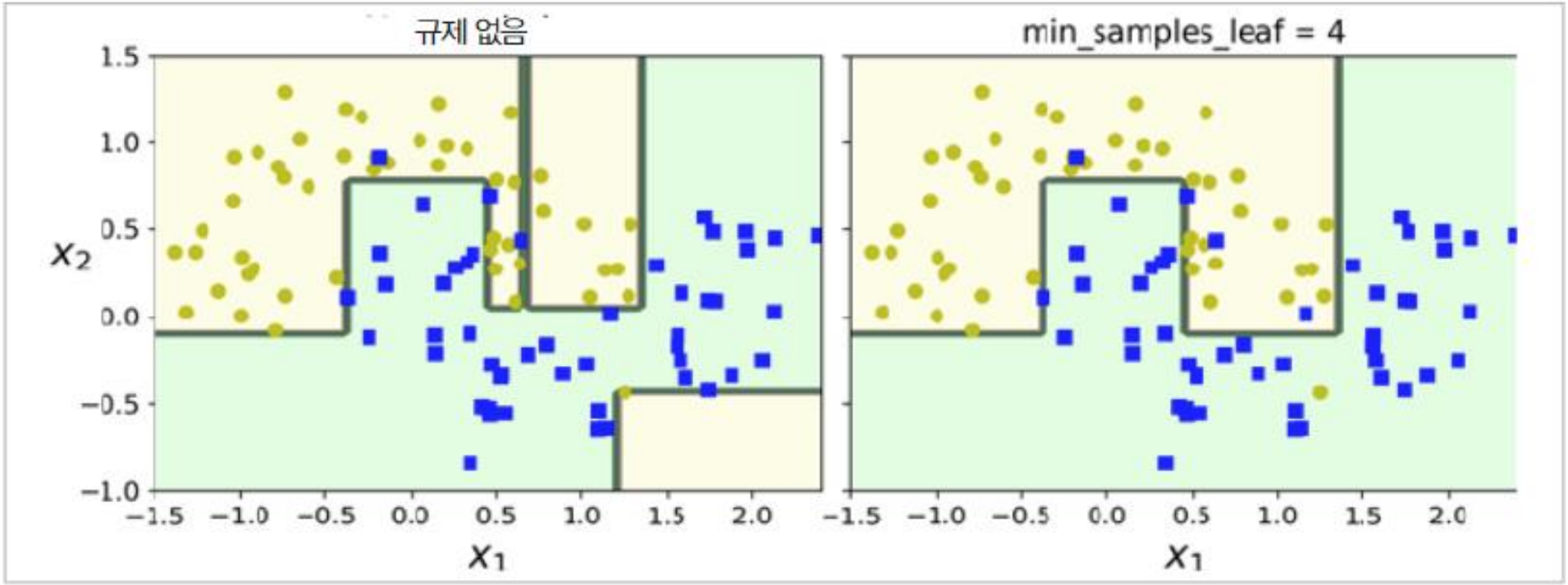


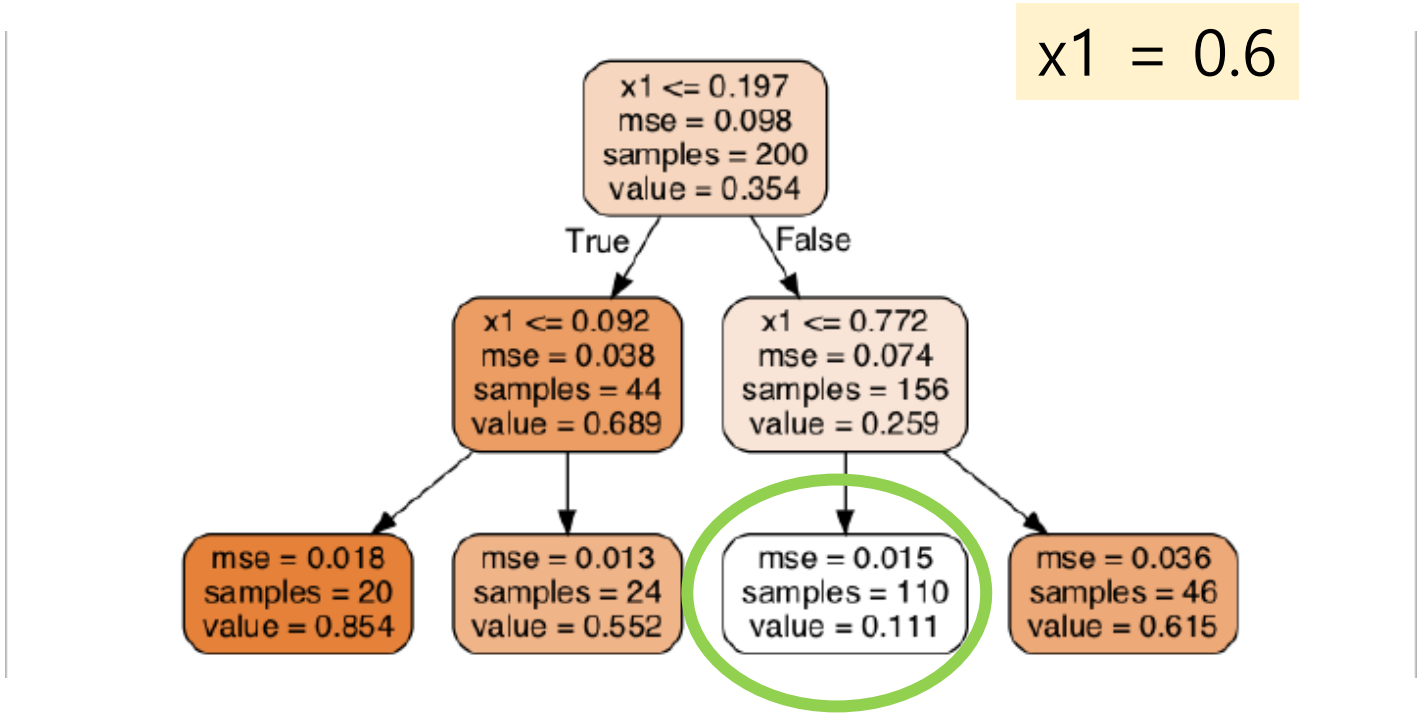
그림 6-3 min_samples_leaf 매개변수를 사용한 규제

6-8

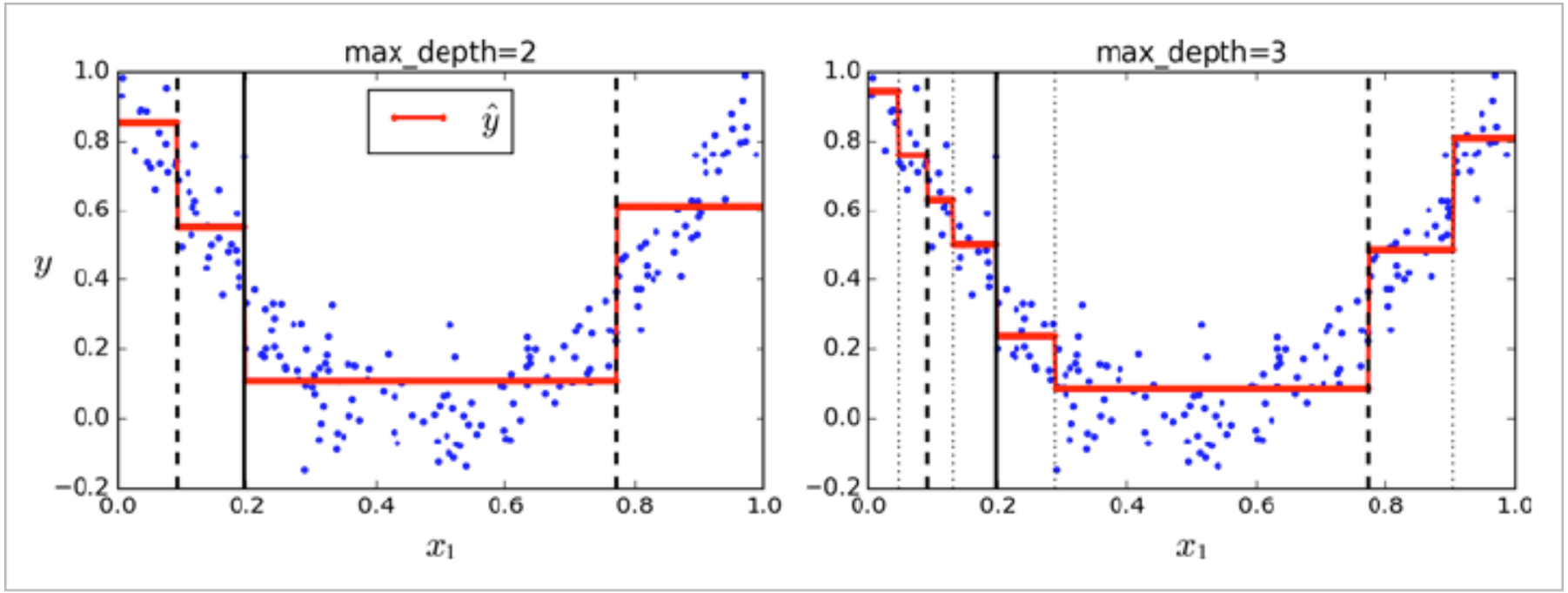
회귀

회귀

사이킷런의 DecisionTreeRegressor를 사용하여 max_depth=2 설정으로 만든 회귀 트리



회귀



- 각 영역의 예측값: 그 영역 타깃값의 평균
- 알고리즘은 예측값과 많은 샘플이 가까이 있도록 영역 분할

CART 알고리즘

- 평균제곱오차(MSE)를 최소화하도록 분할
- 비용 함수

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

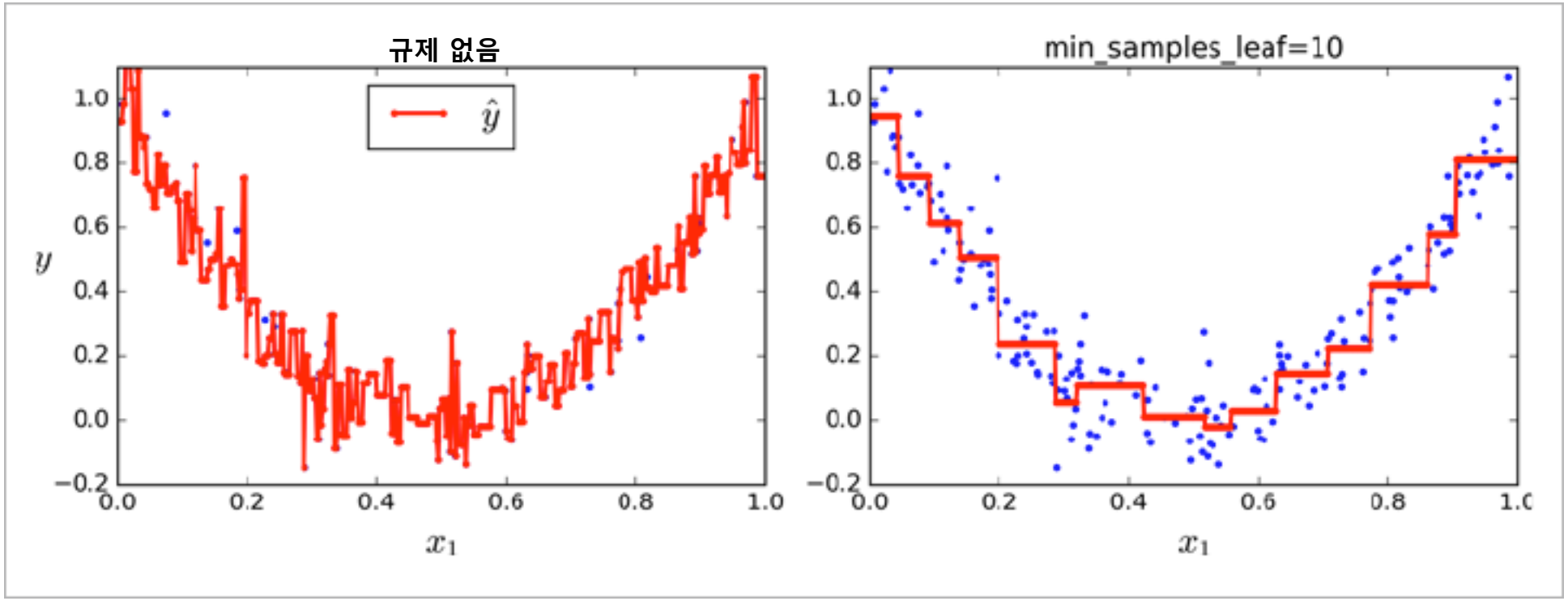
여기서

$$\text{MSE}_{\text{node}} = \sum_{i \in \text{node}} (\hat{y}_{\text{node}} - y^{(i)})^2$$

$$\hat{y}_{\text{node}} = \frac{1}{m_{\text{node}}} \sum_{i \in \text{node}} y^{(i)}$$

CART 알고리즘

- 회귀에서도 결정 트리가 과대적합되기 쉽다.

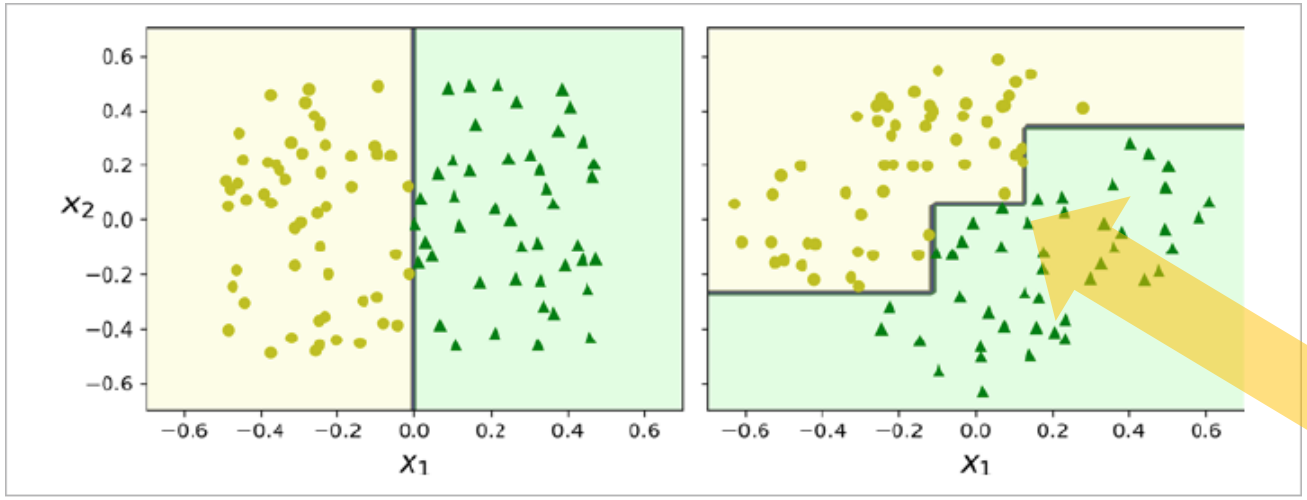


6-9

불안정성

결정 트리의 제한 사항

- 회전(Rotation) 문제
  - 계단 모양의 결정 경계를 만든다. (모든 분할이 축에 수직)
  - 훈련 세트의 회전에 민감하다.
  - 데이터셋의 축이 변경되면 결정 경계도 변경되어 성능이 저하될 수 있다.



결정 트리의 제한 사항

- 회전(Rotation) 문제
  - PCA 기법 사용 (훈련 데이터를 더 좋은 방향으로 회전시킨다.)

    (8장 참조)

결정 트리의 제한 사항

- 규제(Regularization) 문제
  - 과대적합 문제 발생
  - ➢ 규제 적용

    - max_depth, min_samples_leaf, max_leaf_nodes 등과 같은 하이퍼파라미터를 사용하여 결정 트리의 규제를 조절

결정 트리의 제한 사항

- 불안정성 문제
  - 결정 트리는 데이터셋의 작은 변화에도 민감하게 반응
  - ➢ 앙상블 학습(ensemble learning) 기법과 함께 사용
    - 여러 개의 결정 트리를 사용하여 예측을 수행하고 그 결과를 종합하여 더 강력한 예측 모델을 만드는 기법

    ex) 랜덤 포레스트(Random Forest), 부스팅(Boosting) 등

결정 트리의 제한 사항

- 다중 클래스 분류(Multi-class classification) 문제
  - 이진 분류(Binary classification) 문제에 더 특화되어 있다.
  - ➢ 다중 클래스 분류 문제를 다룰 때는 일반적으로 일대다(one-vs.-rest) 방법을 사용
    - 각 클래스를 다른 모든 클래스와 구분하여 이진 분류 문제로 변환한다.
    - 이러한 변환은 데이터셋이 불균형한 경우에 특히 중요하다.