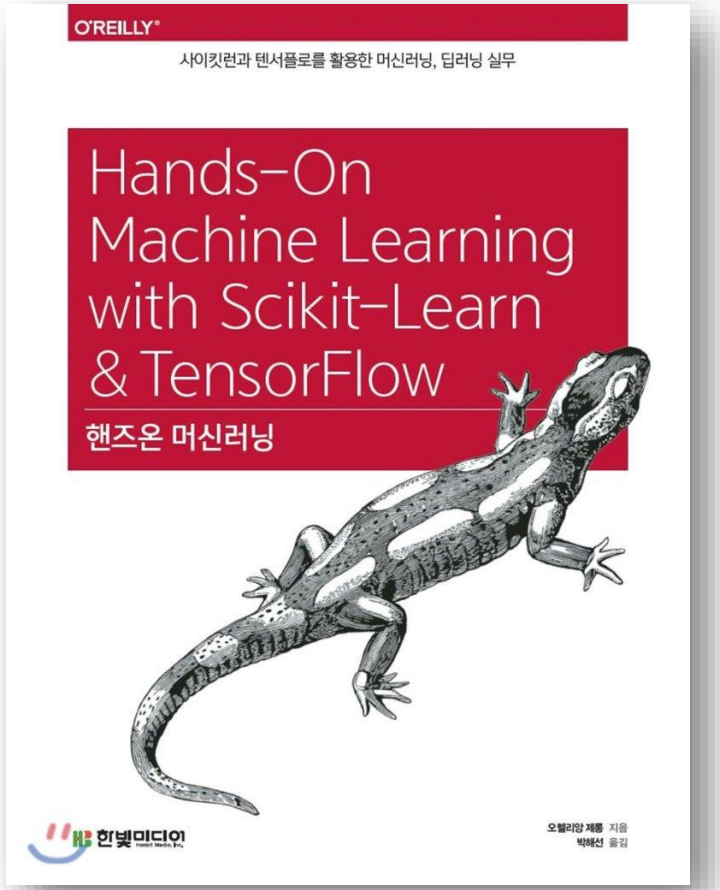
6 결정 트리


O'REILLY
사이킷런과 텐서플로를 활용한 머신러닝, 딥러닝 실무
Hands-On Machine Learning with Scikit-Learn & TensorFlow
핸즈온 머신러닝
한빛미디어
오렐리앙 제롱 지음
박해선 옮김

6장 학습된 결정 트리

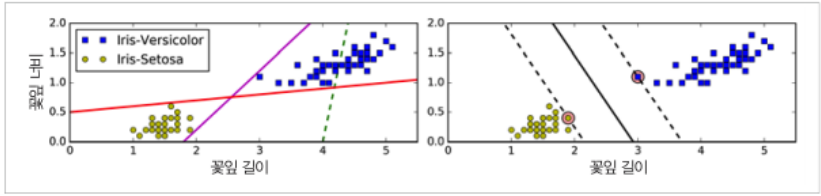
해당 슬라이드 내용은 **핸즈온 머신러닝 6장** 내용 중 일부를 정리한 것입니다. 머신러닝에 관심이 있는 분이라면 해당 책을 직접 읽기를 추천 드립니다.

6장 학습된 결정 트리 목차

6.1 결정 트리 학습과 시각화

- 로지스틱 회귀, SVM은 선형모델의 각 특징별 가중치를 학습하는 방식
- 트리를 만들기 위한 특징과 각 특징별 조건을 학습

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$



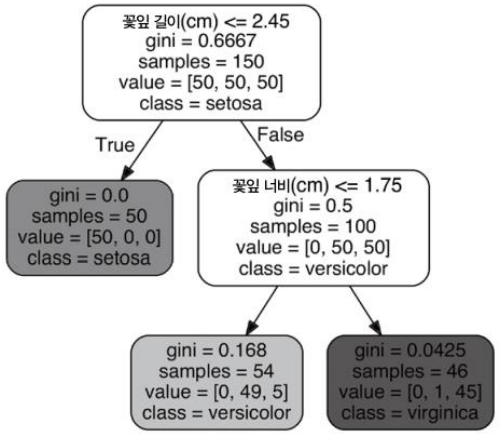
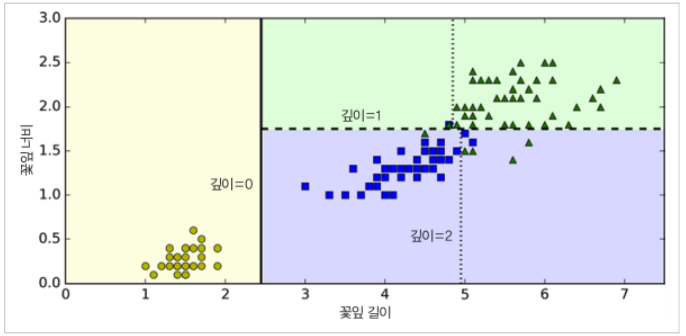


그림 6-2 결정 트리의 결정 경계



6.2 예측하기

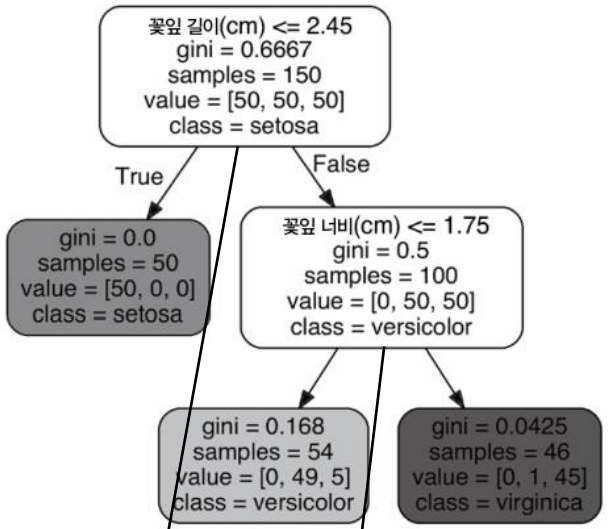
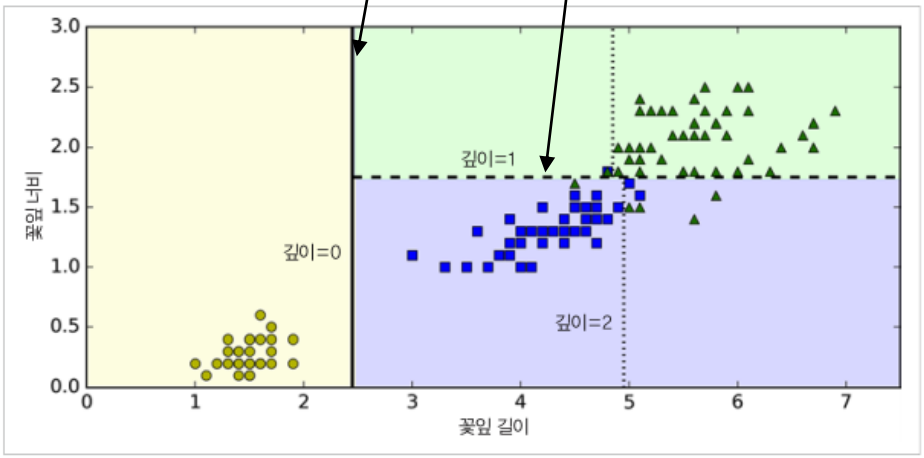


그림 6-2 결정 트리의 결정 경계



입력1

꽃잎 길이: 2.2

꽃잎 넓이: 1.0

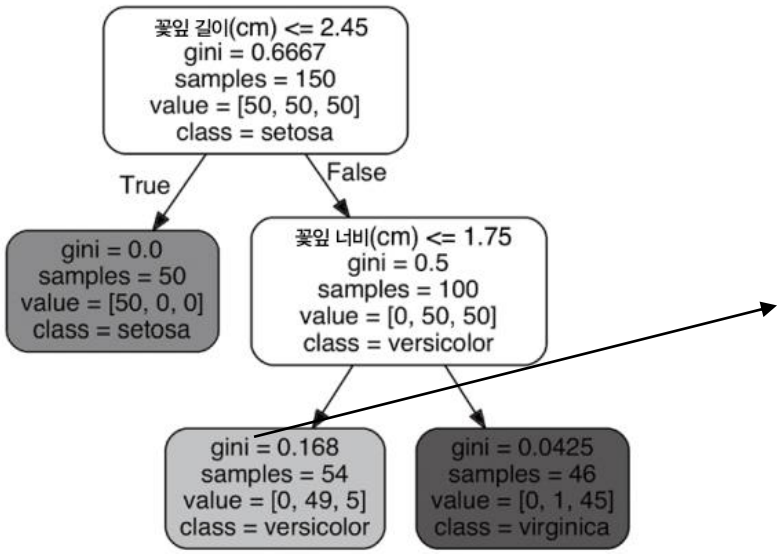
입력2

꽃잎 길이: 2.6

꽃잎 넓이: 1.5

6.2 예측하기

루트 노드

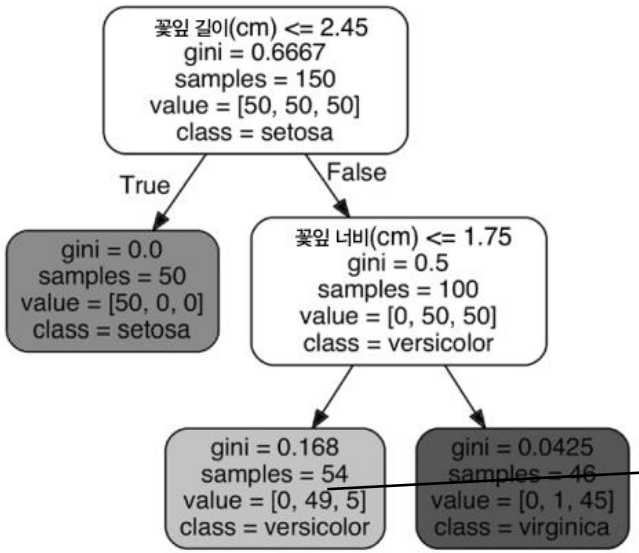


리프노드 노드

식 6-1 지니 불순도

$$G_i = 1 - \sum_{k=1}^{n} p_{i,k}^{2} \qquad 1 - (0/54)^2 - (49/54)^2 - (5/54)^2 \approx 0.168$$

6.3 클래스 확률 추정



꽃잎 길이: 5
꽃잎 넓이: 1.5

0/54 = 0.0
49/54 = 0.907
5/54 = 0.092

화이트박스와 블랙박스

모델 해석: 화이트박스와 블랙박스

여기에서 볼 수 있듯이 결정 트리는 매우 직관적이고 결정 방식을 이해하기 쉽습니다. 이런 모델을 **화이트박스**white box 모델이라고 합니다. 반대로 앞으로 보게 될 랜덤 포레스트나 신경망은 **블랙박스**black box 모델입니다. 이 알고리즘들은 성능이 뛰어나고 예측을 만드는 연산 과정을 쉽게 확인할 수 있습니다. 그렇지만 왜 그런 예측을 만드는지는 쉽게 설명하기 어렵습니다. 예를 들어 신경망이 어떤 사람이 사진에 있다고 판단했을 때 무엇이 이런 예측을 낳게 했는지 파악하기 매우 어렵습니다. 모델이 그 사람의 눈을 인식한 걸까요? 아니면 입 또는 코 또는 신발일까요? 아니면 그 사람이 앉아 있는 소파 때문일까요? 반면에 결정 트리는 필요하다면 (예를 들면 붓꽃 분류를 위해) 수동으로 직접 따라 해볼 수도 있는 간단하고 명확한 분류 방법을 사용합니다.

6.4 CART

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

여기서 $G_{\text{left/right}}$는 왼쪽/오른쪽 서브셋의 불순도
$m_{\text{left/right}}$는 왼쪽/오른쪽 서브셋의 샘플 수

CAUTION_ 여기에서 볼 수 있듯이 CART 알고리즘은 **탐욕적 알고리즘**greedy algorithm입니다. 맨 위 루트 노드에서 최적의 분할을 찾으며 각 단계에서 이 과정을 반복합니다. 현재 단계의 분할이 몇 단계를 거쳐 가장 낮은 불순도로 이어질 수 있을지 없을지는 고려하지 않습니다. 탐욕적 알고리즘은 종종 납득할만한 훌륭한 솔루션을 만들어냅니다. 하지만 최적의 솔루션을 보장하지는 않습니다.

불행하게도 최적의 트리를 찾는 것은 **NP-완전**NP-Complete 문제[3]로 알려져 있습니다. 이 문제는 $O(\exp(m))$ 시간이 필요하고 매우 작은 훈련 세트에도 적용하기 어렵습니다. 그러므로 '납득할 만한 좋은 솔루션'으로만 만족해야 합니다.

6.6 지니 불순도 또는 엔트로피

식 6-3 엔트로피

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

지니 불순도와 엔트로피 중 어떤 것을 사용해야 할까요? 실제로는 큰 차이가 없습니다. 즉, 둘 다 비슷한 트리를 만들어냅니다. 지니 불순도가 조금 더 계산이 빠르기 때문에 기본값으로 좋습니다. 그러나 다른 트리가 만들어지는 경우 지니 불순도가 가장 빈도 높은 클래스를 한쪽 가지 branch로 고립시키는 경향이 있는 반면 엔트로피는 조금 더 균형 잡힌 트리를 만듭니다.[8]

칸아카데미 정보측정
https://ko.khanacademy.org/computing/computer-science/informationtheory/moderninfotheory/v/how-do-we-measure-information-language-of-coins-10-12

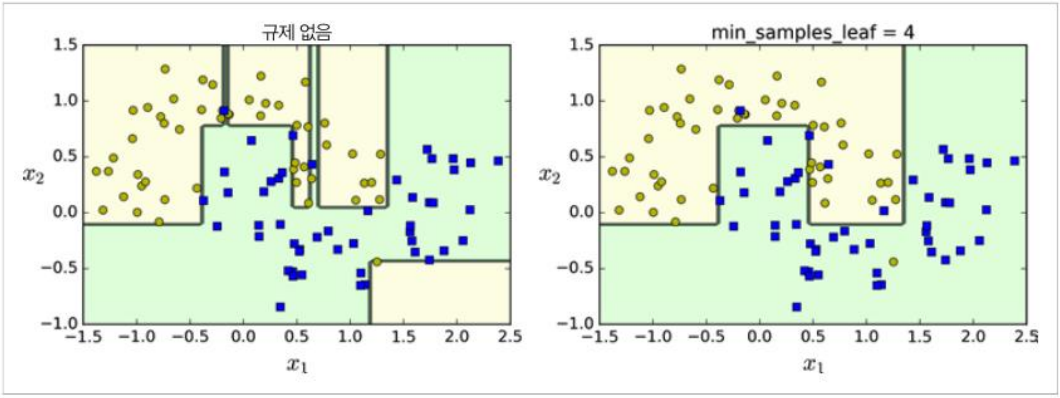
칸아케데미 엔트로피 영상
https://ko.khanacademy.org/computing/computer-science/informationtheory/moderninfotheory/v/information-entropy

6.7 규제 매개변수

훈련 데이터에 대한 과대적합을 피하기 위해 학습할 때 결정 트리의 자유도를 제한할 필요가 있습니다. 이미 알고 있듯이 이를 규제라고 합니다. 규제 매개변수는 사용하는 알고리즘에 따라 다르지만, 보통 적어도 결정 트리의 최대 깊이는 제어할 수 있습니다. 사이킷런에서는 `max_depth` 매개변수로 이를 조절합니다(기본값은 제한이 없는 것을 의미하는 `None`입니다). `max_depth`를 줄이면 모델을 규제하게 되고 과대적합의 위험이 감소합니다.

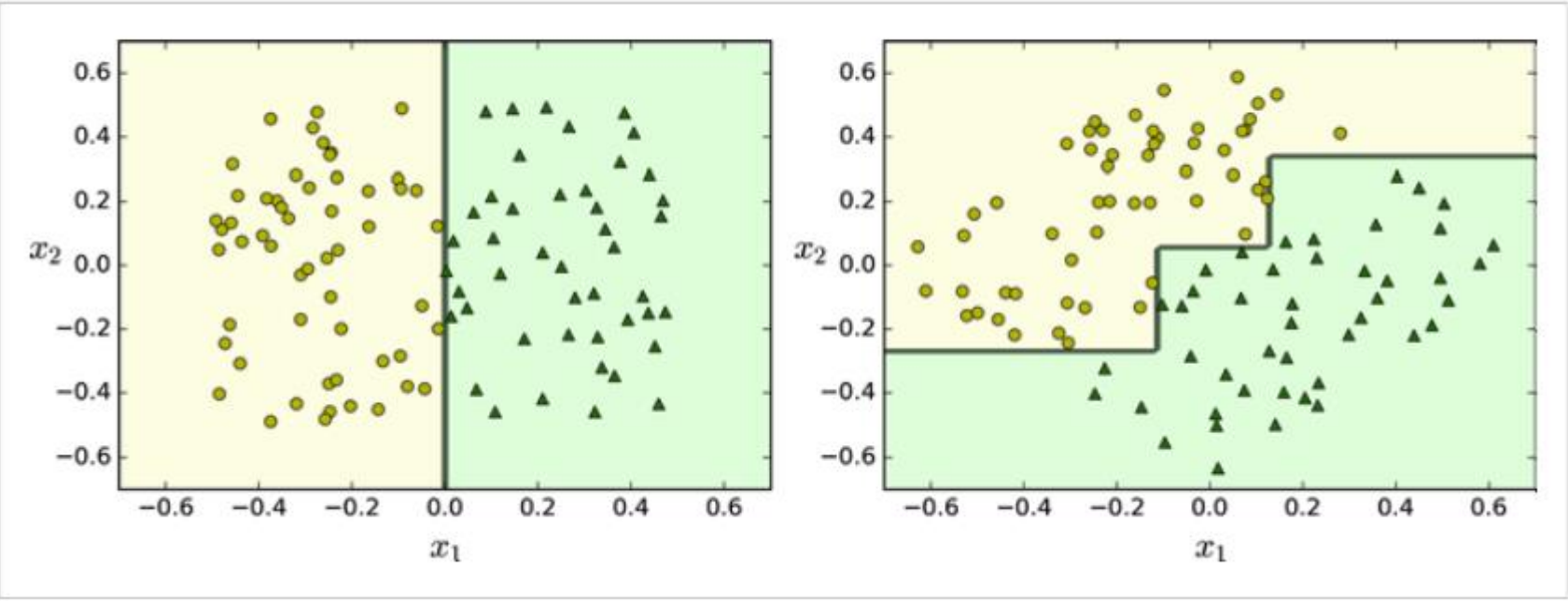
`DecisionTreeClassifier`에는 비슷하게 결정 트리의 형태를 제한하는 다른 매개변수가 몇 개 있습니다. `min_samples_split`(분할되기 위해 노드가 가져야 하는 최소 샘플 수), `min_samples_leaf`(리프 노드가 가지고 있어야 할 최소 샘플 수), `min_weight_fraction_leaf`(`min_samples_leaf`와 같지만 가중치가 부여된 전체 샘플 수에서의 비율), `max_leaf_nodes`(리프 노드의 최대 수), `max_features`(각 노드에서 분할에 사용할 특성의 최대 수)가 있습니다. `min_`으로 시작하는 매개변수를 증가시키거나 `max_`로 시작하는 매개변수를 감소시키면 모델에 규제가 커집니다.[9]

그림 6-3 min_samples_leaf 매개변수를 사용한 규제



6.8 규제 매개변수

그림 6-7 훈련 세트의 회전에 민감한 결정 트리



감사합니다

실습링크
https://colab.research.google.com/github/rickiepark/handson-ml2/blob/master/06_decision_trees.ipynb