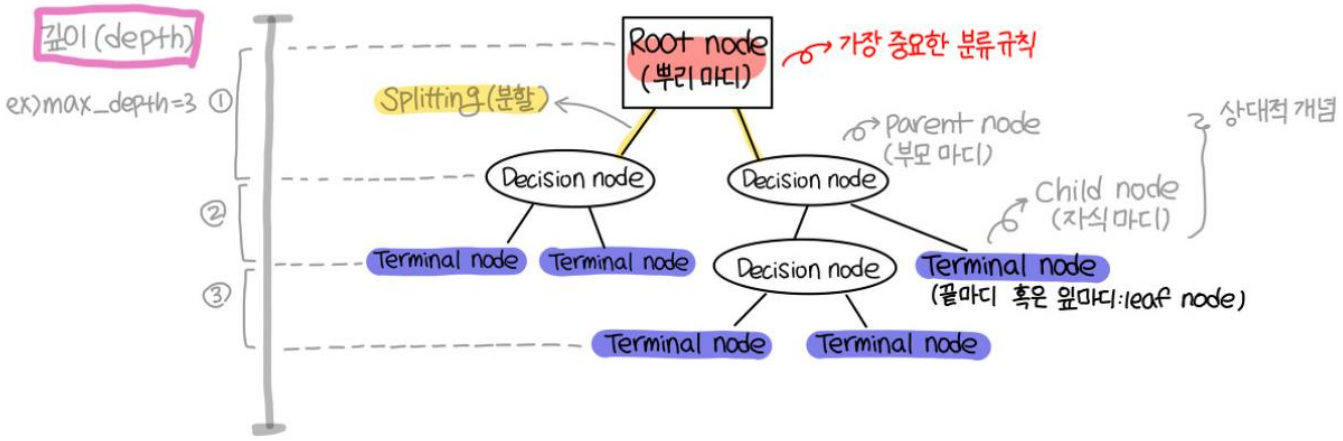
[핸즈온 머신러닝_Chapter 6. 결정트리]

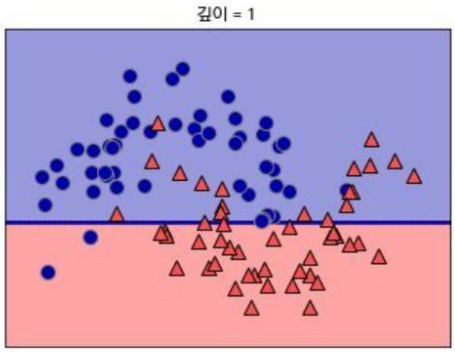


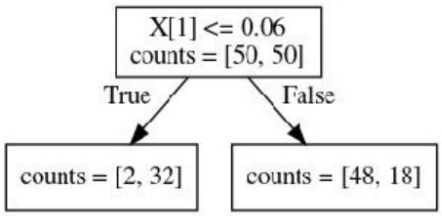
4. 결정 트리가 훈련 세트에 과소적합되었다면 입력 특성의 스케일을 조정하는 것이 좋을까요?

결정 트리는 훈련 데이터의 스케일이나 원점에 맞추어져 있는지 상관하지 않는다. 이것이 결정트리의 장점. 그러므로 결정 트리가 훈련세트에 과소적합되었다고 입력 특성의 스케일을 조정하는 것은 시간 낭비

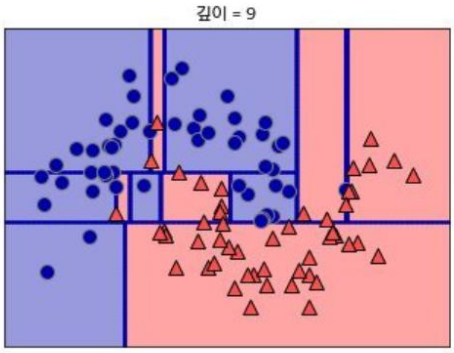
의사결정 나무의 깊이 조정 이해

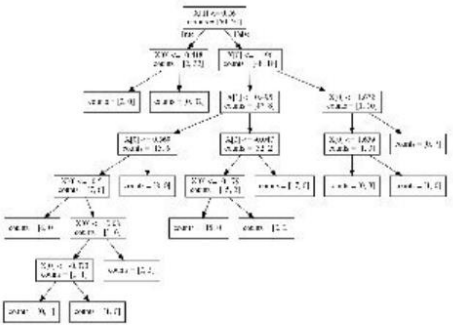
트리의 깊이는 특성을 몇번이나 분류할지를 정하는 것





깊이를 너무 조금 주면 거의 분류가 되지 않는 과소적합 상태가 됨





깊이를 너무 많이 주면 모델의 복잡도가 커져 시각화 하였을 때 확인도 어렵고, 훈련데이터에 과대적합 됨

매개변수

max_depht : 트리의 최대 깊이 설정

max_leaf_nodes : 리프 노드의 최대 개수 지정

max_samples_leaf : 최소 리프 노드 샘플 개수 지정

* 트리의 모델 복잡도를 조절하는 매개변수는 사전 가지치기 매개변수 중 하나만 지정해도 과대적합을 막기에 충분

5. 백만 개의 샘플을 가진 훈련 세트에 결정 트리를 훈련시키는 데 한 시간이 걸렸다면, 천만 개의 샘플을 가진 훈련 세트에 결정 트리를 훈련시키는 데는 대략 얼마나 걸릴까요?

예측을 하려면 결정 트리를 루트 노드에서 리프 노트까지 탐색해야 한다. 결정 트리는 거의 균형을 이루고 있으므로 결정 트리를 탐색하기 위해서는 약 $O(log_2(m))$개의 노드를 거쳐야 한다. 각 노드는 하나의 특성 값만 확인하기 때문에 예측에 필요한 전체 복잡도는 특성 수와 무관하다.

*균형 이진 트리에서 깊이 d에서 리프 노드의 개수 2^d, 리프 노드가 훈련데이터수(m)만큰 있다면,

- 트리의 깊이: $log_2(m) = \log(m) / \log(2)$

- 결정 트리 훈련의 계산 복잡도: $O(n \times m \log(m))$

훈련세트의 크기에 10을 곱하면 훈련시간은 K배 늘어난다.

$K = \frac{n \times 10m \times \log(10m)}{n \times m \times \log(m)} = 10 \times \frac{long(10m)}{\log(m)}$

만약 $m = 10^6$이면 $K \approx 11.7$이므로 훈련에 대략 11.7시간이 걸릴 것으로 예상

6. 십만 개의 샘플을 가진 훈련 세트가 있다면 presort=True로 지정하는 것이 훈련 속도를 높일까요?

데이터셋의 샘플 수가 수천 개 미만일 때 훈련 세트를 사전에 정렬하여 훈련 속도를 높일 수 있다.

100,000 개의 샘플을 포함하고 있을 때 presort=True로 지정하면 훈련 속도가 매우 느려질 것이다.