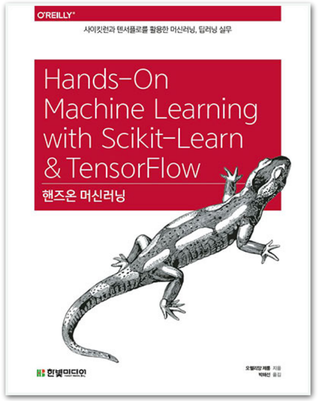


핸즈온 머신러닝

6장. 결정 트리

박해선(옮긴이)
haesun.park@tensorflow.blog
https://tensorflow.blog

결정 트리(Decision Tree)

- 분류, 회귀 문제에 모두 적용 가능합니다.
- 복잡한 데이터셋도 학습할 수 있습니다.
- 데이터 전처리가 필요하지 않습니다.
- 사이킷런은 CART 알고리즘(이진 트리)을 사용합니다.
- 랜덤 포레스트와 그래디언트 부스팅 앙상블 학습의 기본 학습기입니다.
- 화이트 박스(White box) like 선형 모델 vs 블랙 박스 like 랜덤 포레스트, 신경망
- 비모수 모델, 비파라미터 모델(nonparametric model)

Simple Example

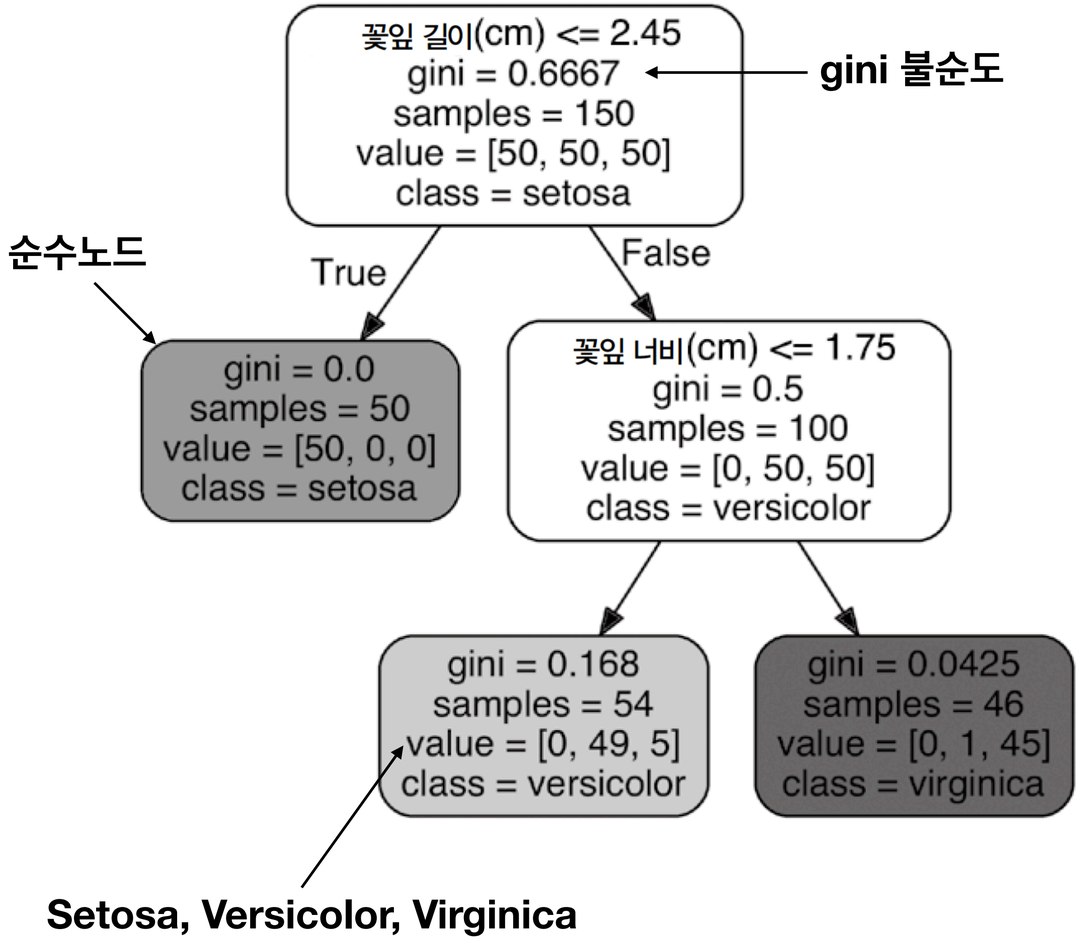
```
from sklearn.datasets import load_iris
from sklearn.tree import DecisionTreeClassifier

iris = load_iris()
X = iris.data[:, 2:]  # 꽃잎의 길이와 너비
y = iris.target

tree_clf = DecisionTreeClassifier(max_depth=2)
tree_clf.fit(X, y)

from sklearn.tree import export_graphviz

export_graphviz(
        tree_clf,
        out_file=image_path("iris_tree.dot"),
        feature_names=iris.feature_names[2:],
        class_names=iris.target_names,
        rounded=True,
        filled=True
    )
```



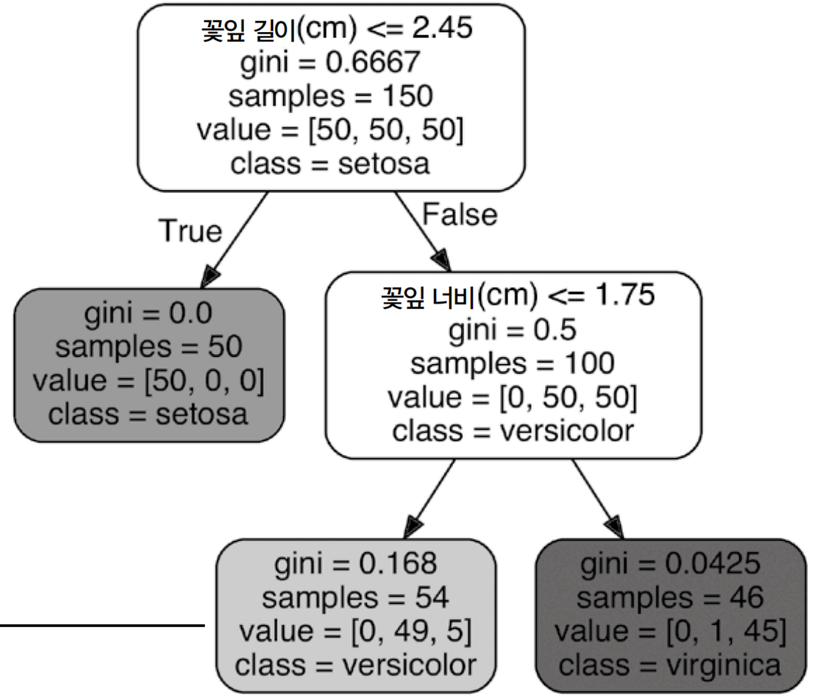
지니 불순도(Gini impurity)

- 지니 불순도는 노드의 샘플 클래스가 얼마나 분산되어 있는지를 측정합니다.
- DecisionTreeClassifier(criterion='gini'), 기본값
- 최악 0.5 ~ 최상 0

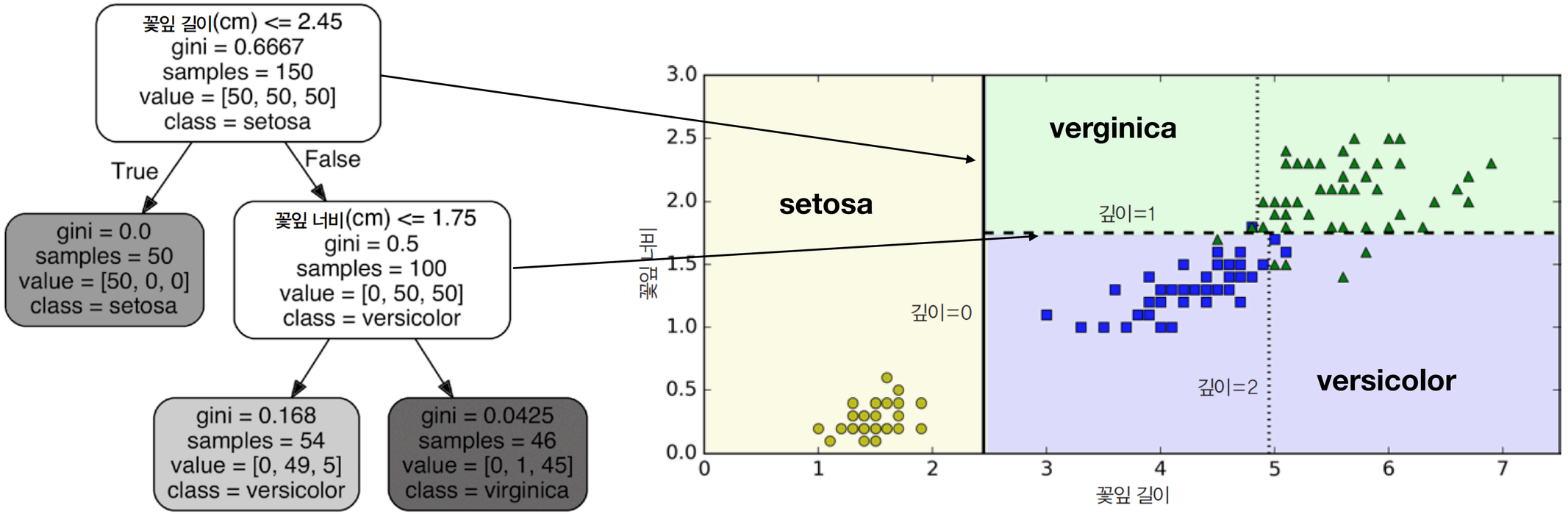
$$G_i = 1 - \sum_{k=1}^{n} p_{i,k}^2$$

- $p_{i,k}$는 i 번째 노드에 있는 훈련 샘플 중 클래스 k에 속한 비율

$$1 - (0/54)^2 - (49/54)^2 - (5/54)^2 \approx 0.168$$



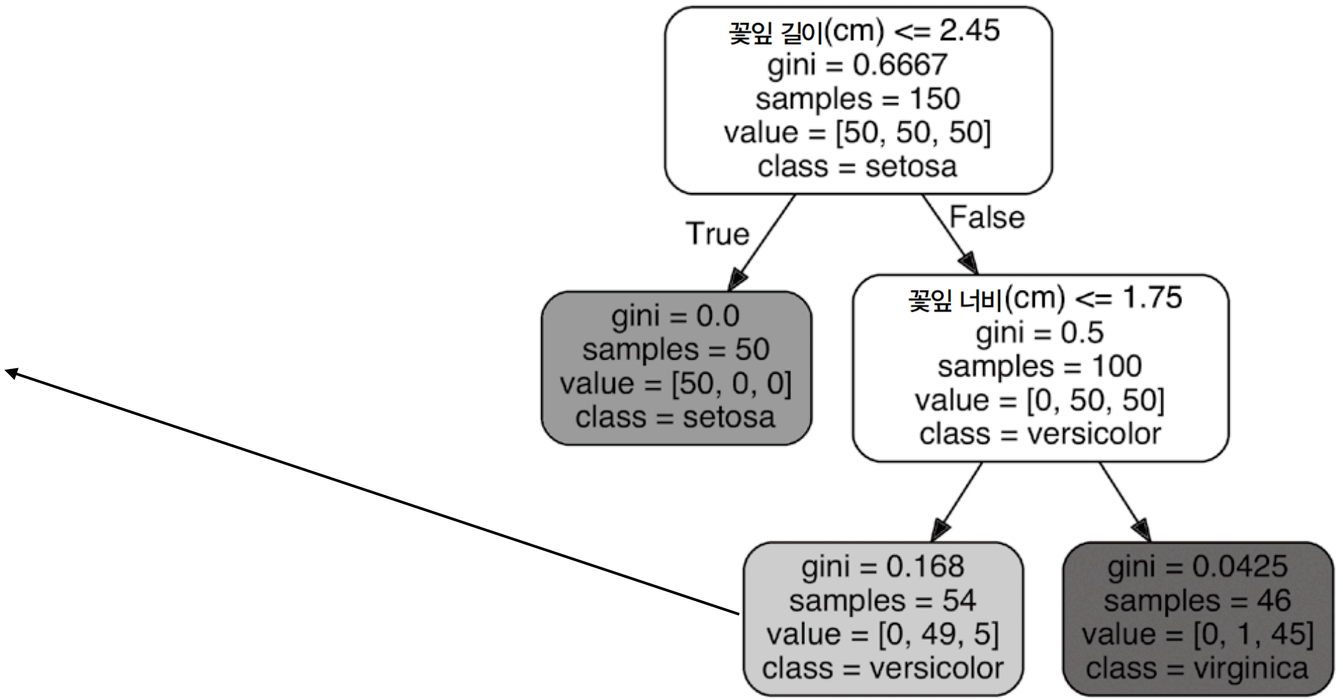
결정 경계(decision boundary)



클래스 확률 추정

- 리프 노드의 훈련 샘플의 클래스 비율
- Ex. 길이 5cm, 너비 1.5cm

```
>>> tree_clf.predict_proba([[5, 1.5]])
array([[ 0. ,  0.90740741,  0.09259259]])
>>> tree_clf.predict([[5, 1.5]])
array([1])
```



CART 비용 함수

- 탐욕적(greedy) 알고리즘
  : 전체 분할을 고려한 최적해가 아니라 현재 노드에서 최적의 분할을 찾습니다(납득할 만한 솔루션).

특성 k, 임계값 t_k

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

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

계산 복잡도

깊이가 ***d*** 인 균형 이진 트리의 리프 노트 개수

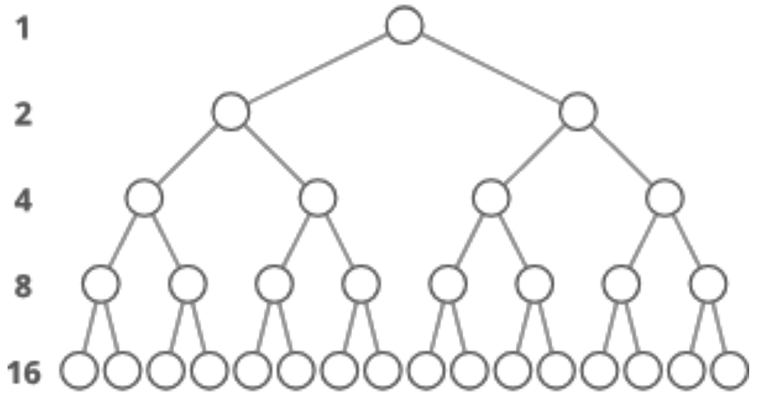
$$n = 2^d$$

리프 노트가 훈련 샘플 개수 ***m*** 개 만큼 있다면

$$m = 2^d$$

$$d = log_2 m$$

예측의 계산 복잡도 $O(log_2 m)$



노드 하나에서 특성 하나를 정렬하는 복잡도

$$mlog(m)$$

노드 하나에서 전체 특성을 정렬하는 복잡도

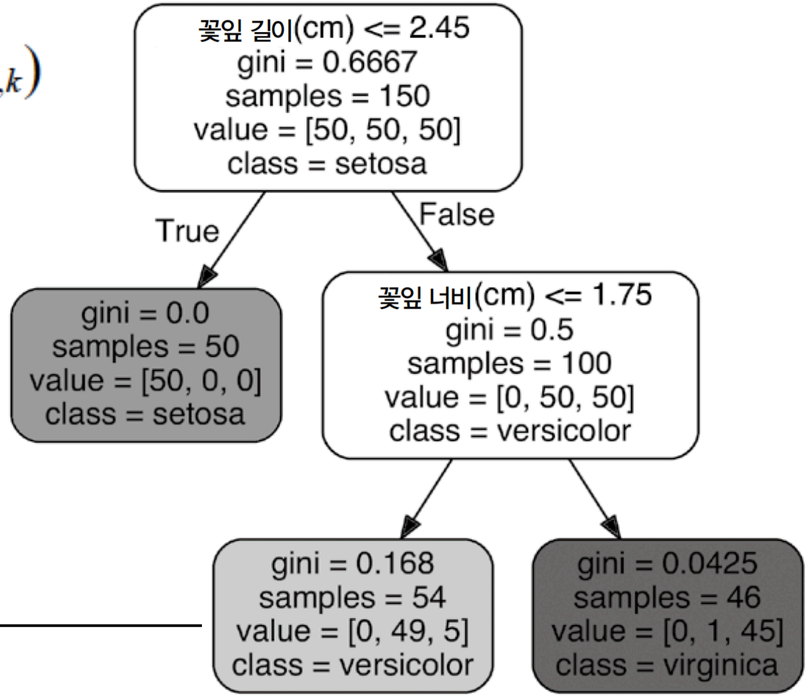
$$nmlog(m)$$

전체 노드에서 전체 특성을 정렬하는 복잡도

$$nm^2log(m)$$

엔트로피(entropy) 불순도

- DecisionTreeClassifier(criterion='entropy')
- 쿨백 라이블러 발산 $D_{KL}(P||Q) = H(P) - H(P,Q)$
- 엔트로피 $H_i = -\sum_{\substack{k=1 \\ p_{i,k} \neq 0}}^{n} p_{i,k} \log_2(p_{i,k})$
- 큰 차이는 없지만 조금 더 균형잡힌 트리를 만듦



$$-\frac{0}{54}log_2(\frac{0}{54}) - \frac{49}{54}log_2(\frac{49}{54}) - \frac{5}{54}log_2(\frac{5}{54}) \approx 0.445$$

불순도에 대한 고찰

정보 이득 $IG(D_p, a) = I(D_p) - \frac{N_{left}}{N} I(D_{left}) - \frac{N_{right}}{N} I(D_{right})$

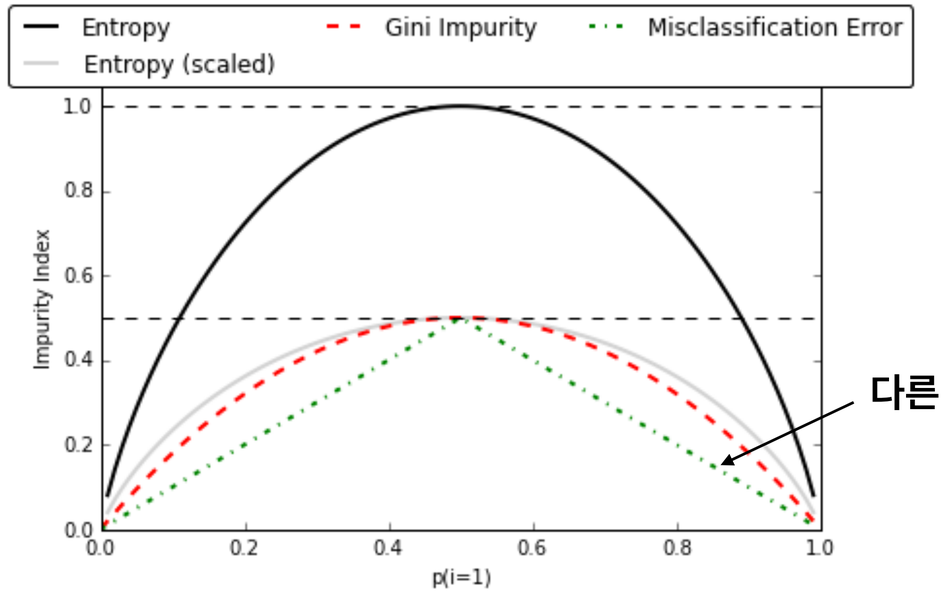
***p(i/t)*: 노드 *t*에서 클래스 *i*에 속한 샘플 비율**

엔트로피 $I_H(t) = -\sum_{i=1}^{c} p(i|t) \, log_2 \, p(i|t)$

클래스가 고루 분포될 때 최대값

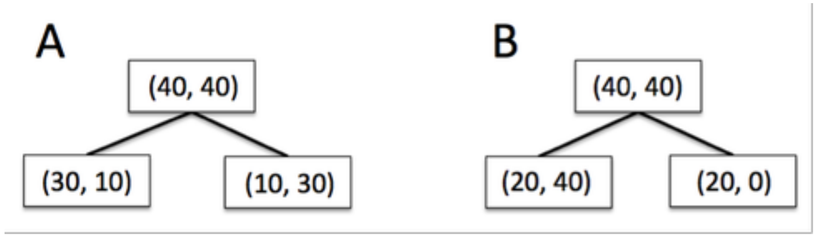
지니 불순도 $I_G(t) = \sum_{i=1}^{c} p(i|t)(1 - p(i|t)) = 1 - \sum_{i=1}^{c} p(i|t)^2$

분류 오차 $I_E = 1 - \max\{p(i|t)\}$



다른 지표보다 불순도 변화가 느림

불순도 계산 예제



$$I_E(D_p) = 1 - \max(0.5 + 0.5) = 0.5$$

$$I_E(D_{left}) = 1 - \max(0.75, 0.25) = 0.25$$

$$I_E(D_{right}) = 1 - \max(0.25, 0.75) = 0.25$$

$$IG_E = 0.5 - \frac{40}{80} \times 0.25 - \frac{40}{80} \times 0.25 = 0.25$$

$$I_E(D_{left}) = 1 - \max(\frac{20}{60}, \frac{40}{60}) = \frac{1}{3}$$

$$I_E(D_{right}) = 1 - \max(1, 0) = 0$$

$$IG_E = 0.5 - \frac{60}{80} \times \frac{1}{3} - \frac{20}{80} \times 0 = 0.25$$

$$I_G(D_p) = 1 - (0.5^2 + 0.5^2) = 0.5$$

$$I_G(D_{left}) = 1 - \left(\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right) = \frac{3}{8} = 0.375$$

$$I_G(D_{right}) = 1 - \left(\left(\frac{1}{4}\right)^2 + \left(\frac{3}{4}\right)^2\right) = \frac{3}{8} = 0.375$$

$$IG_G = 0.5 - \frac{40}{80} \times 0.375 - \frac{40}{80} \times 0.375 = 0.125$$

$$I_G(D_{left}) = 1 - \left(\left(\frac{2}{6}\right)^2 + \left(\frac{4}{6}\right)^2\right) = \frac{4}{9}$$

$$I_G(D_{right}) = 1 - \left(\left(\frac{2}{2}\right)^2 + \left(\frac{0}{2}\right)^2\right) = 1 - 1 = 0$$

$$IG_G = 0.5 - \frac{60}{80} \times \frac{4}{9} - \frac{40}{80} \times 0 = 0.1666$$

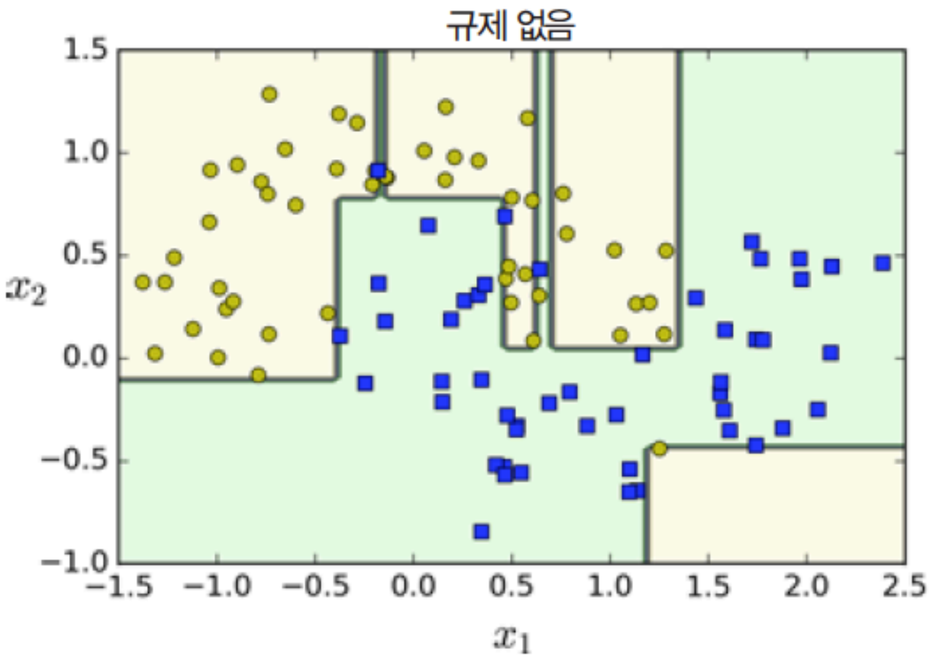
https://tensorflow.blog/2018/03/25/결정-트리와-불순도에-대한-궁금증/

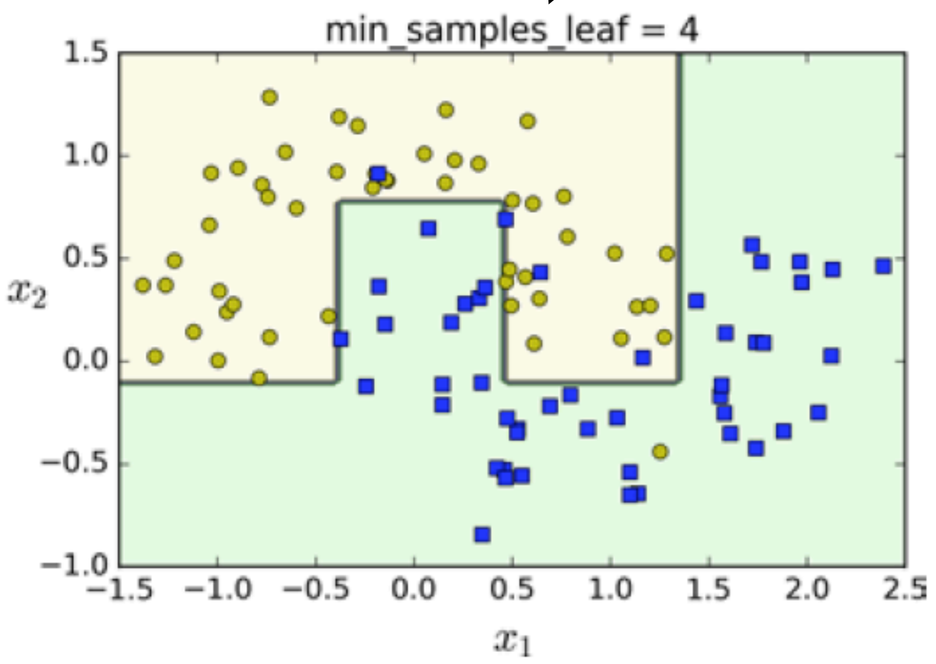
규제 매개변수

- max_depth: 결정 트리 최대 깊이, 기본값 None
- min_samples_split: 분할되기 위해 노드가 가져야 하는 최소 샘플 수, 2
- min_samples_leaf: 리프 노드가 가지고 있어야 할 최소 샘플 수, 1
- min_weight_fraction_leaf: 리프 노드가 가질 가중치(sample_weight) 부여된 전체 샘플 수에서의 비율, 0
- max_leaf_nodes: 리프 노드의 최대 수, None
- max_features: 각 노드에서 분할에 사용할 특성의 최대 수, None
- min_impurity_decrease: 분할로 얻어질 최소한의 불순도, 0
- min_impurity_split: 분할을 위해 필요한 최소 불순도, 0.21에서 삭제 예정
- 사이킷런은 사전 가지치기(pre-pruning)만 지원합니다.

규제 사례

리프 노드 최소 샘플 수



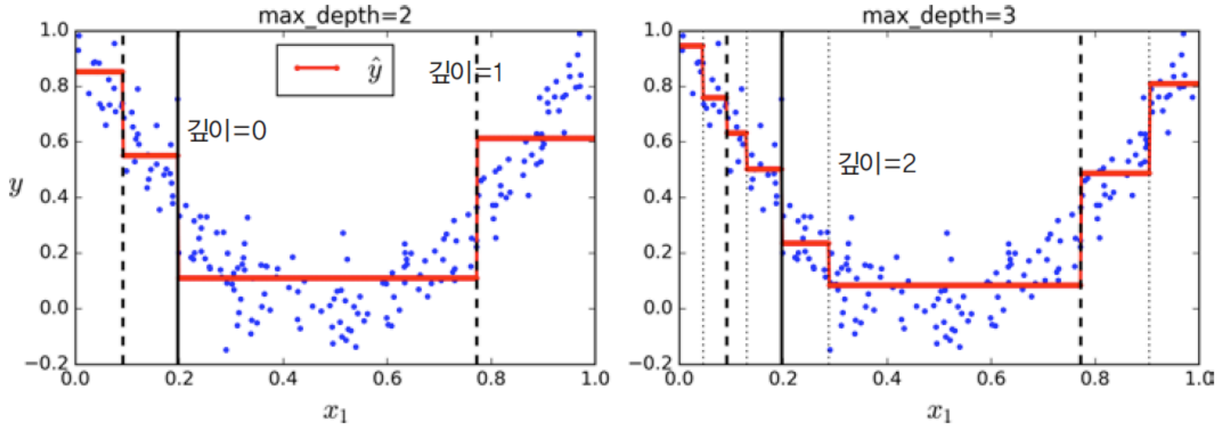


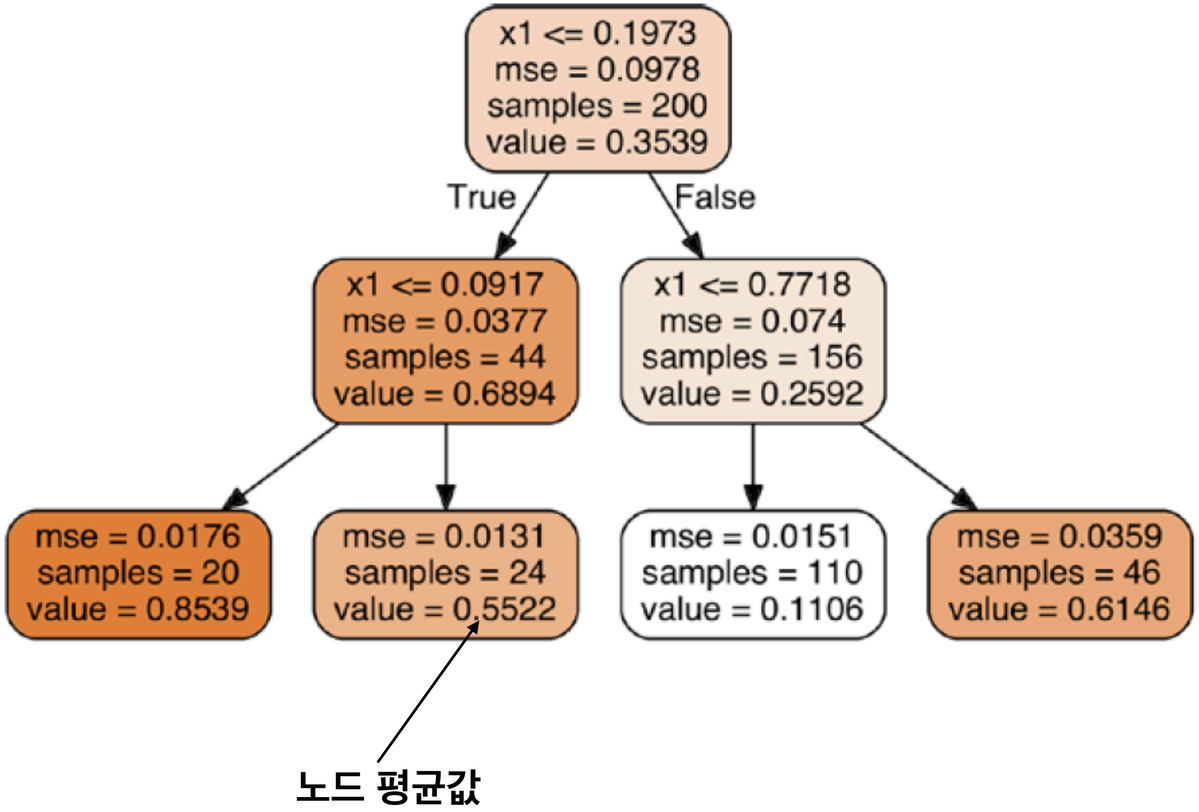
DecisionTreeRegressor

- DecisionTreeClassifier(criterion='mse'), 또는 'mae'

```
from sklearn.tree import DecisionTreeRegressor

tree_reg = DecisionTreeRegressor(max_depth=2)
tree_reg.fit(X, y)
```





회귀의 비용 함수

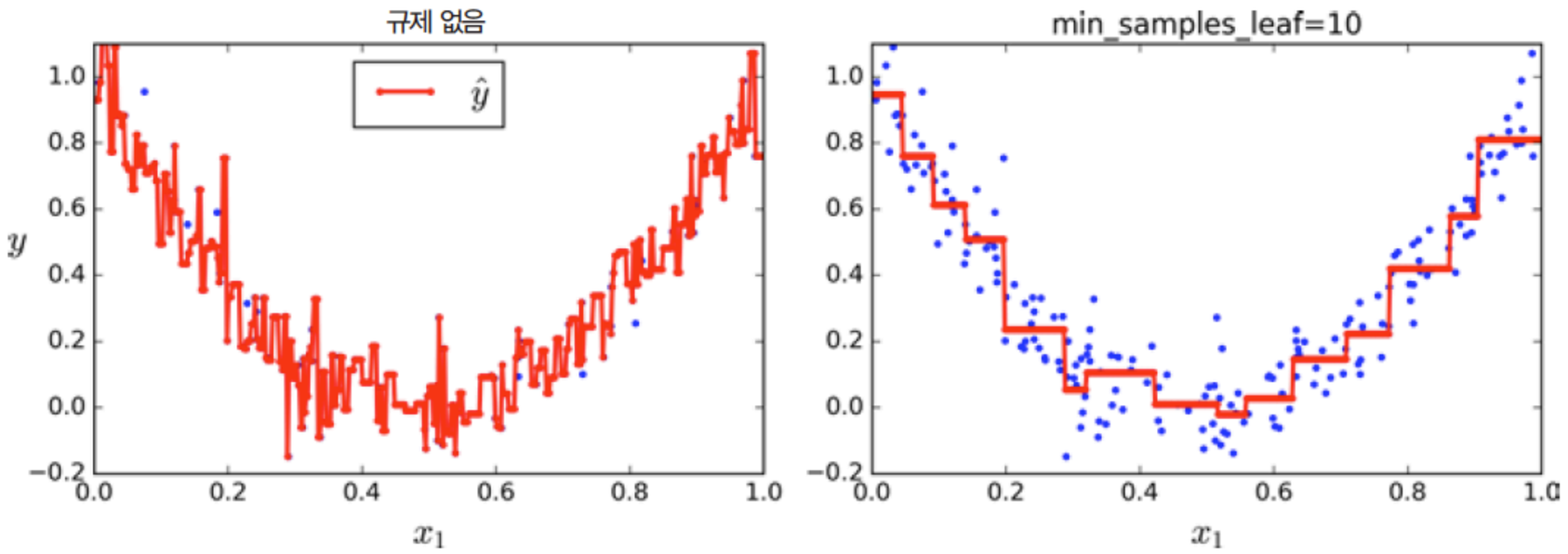
$$J(k, t_k) = \frac{m_{\text{left}}}{m} \text{MSE}_{\text{left}} + \frac{m_{\text{right}}}{m} \text{MSE}_{\text{right}}$$

여기서

$$\begin{cases} \text{MSE}_{\text{node}} = \sum_{i \in \text{node}} \left(\hat{y}_{\text{node}} - y^{(i)} \right)^2 \\ \hat{y}_{\text{node}} = \frac{1}{m_{\text{node}}} \sum_{i \in \text{node}} y^{(i)} \end{cases}$$

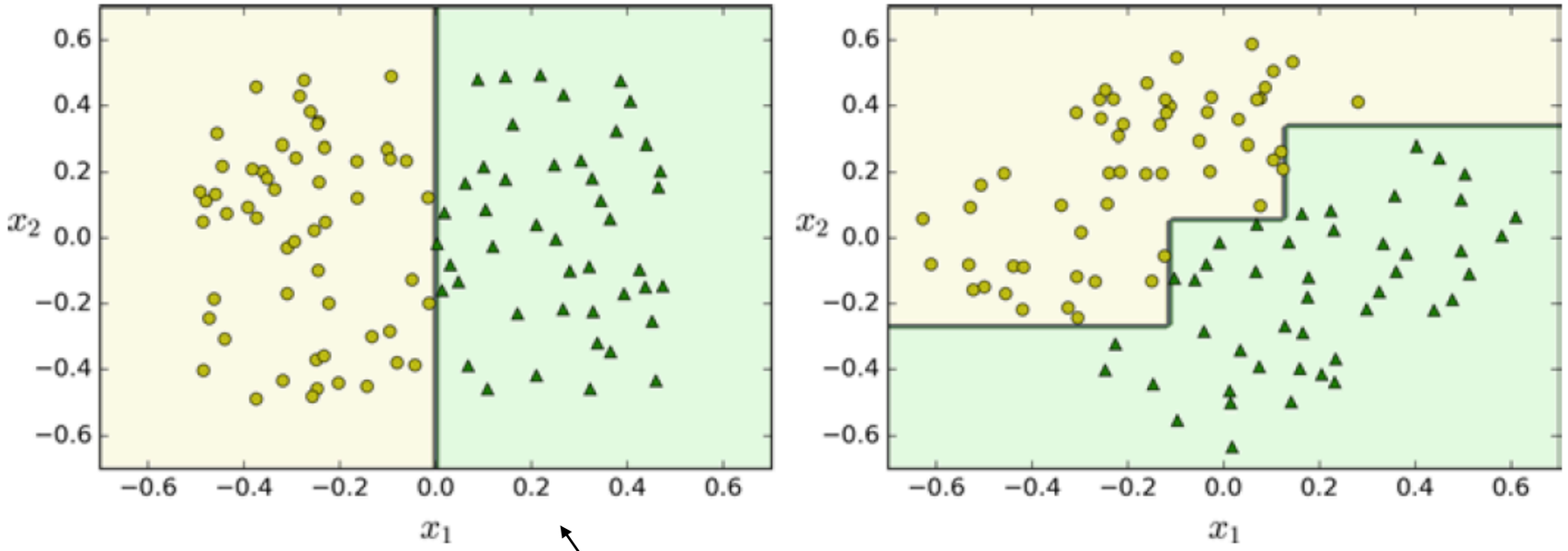
회귀 모델의 규제

리프 노드 최소 샘플 수



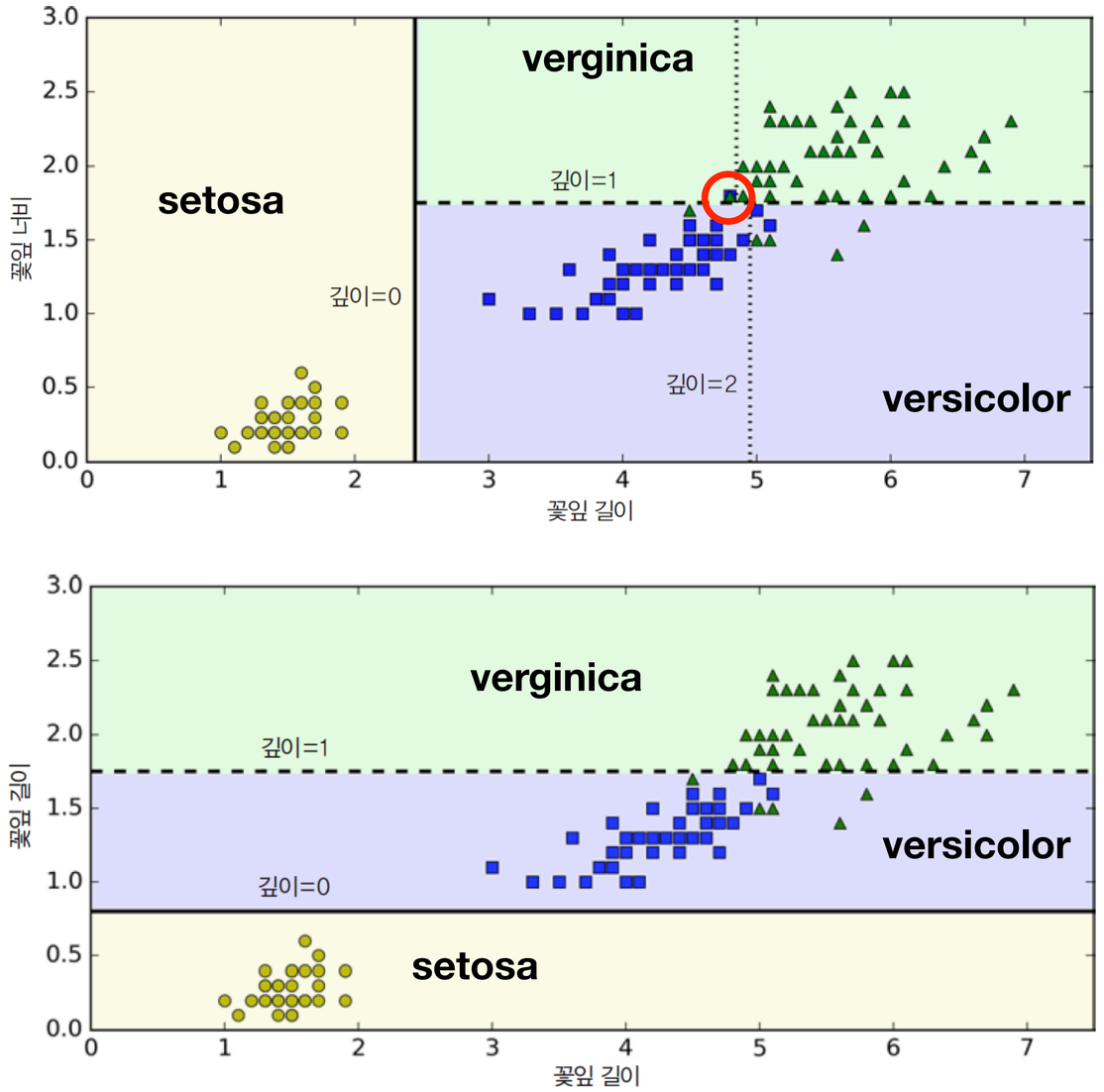
회전 불안정성

- PCA를 사용하여 직교하는 주성분으로 표현하는 것이 좋음



일반화에 더 나음

훈련 세트 분할에 민감


verginica
setosa
versicolor
깊이=1
깊이=0
깊이=2
꽃잎 너비
꽃잎 길이
verginica
versicolor
setosa
깊이=1
깊이=0
꽃잎 길이
꽃잎 길이

감사합니다