Ch7. 앙상블 학습과 랜덤 포레스트

Kaggle code study
4월 7일(3주차)
임혜민

목차

1. 앙상블 학습

▶ 앙상블 학습이란?

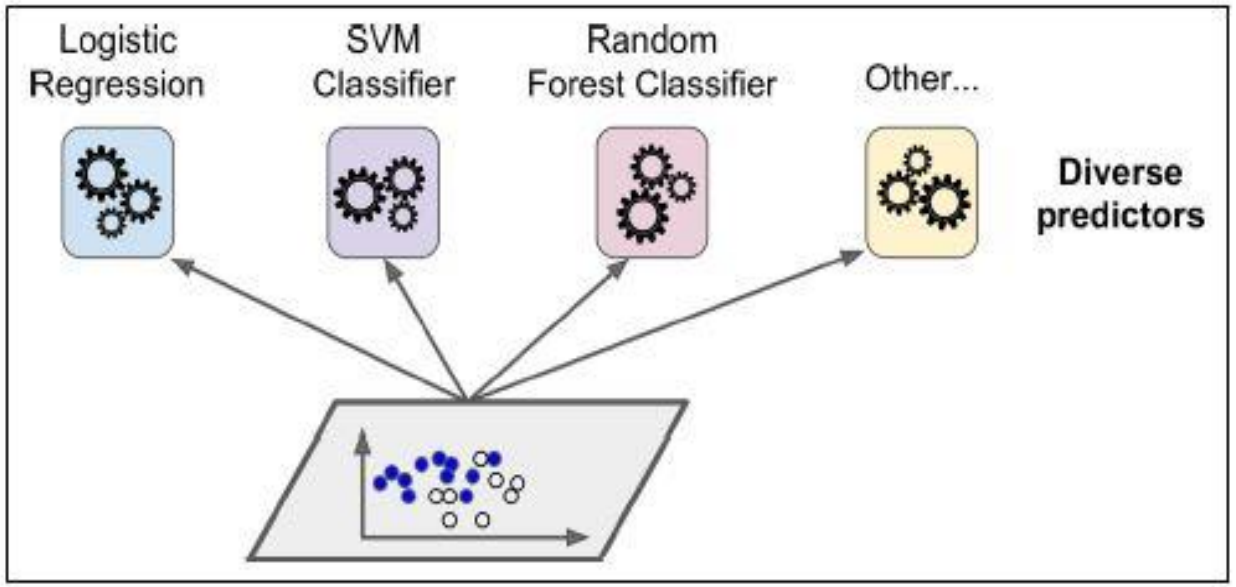
- 일련의 예측기(분류나 회귀 모델)로 부터 예측을 수집하면 가장 좋은 모델 하나보다 더 좋은 예측을 얻을 수 있다.
- 이때, 일련의 예측기를 앙상블이라고 하기때문에 앙상블 학습이라고 한다.
- 앙상블 학습 알고리즘은 앙상블 방법이라 한다.

▶랜덤 포레스트

- 여러 개의 무작위 의사결정 트리(특성들을 랜덤하게 추출)로 이루어진 숲
- 예측을 하기위해 모든 개별 트리의 예측을 구한다.
- 가장 많은 선택을 받은 클래스를 예측으로 삼는다.

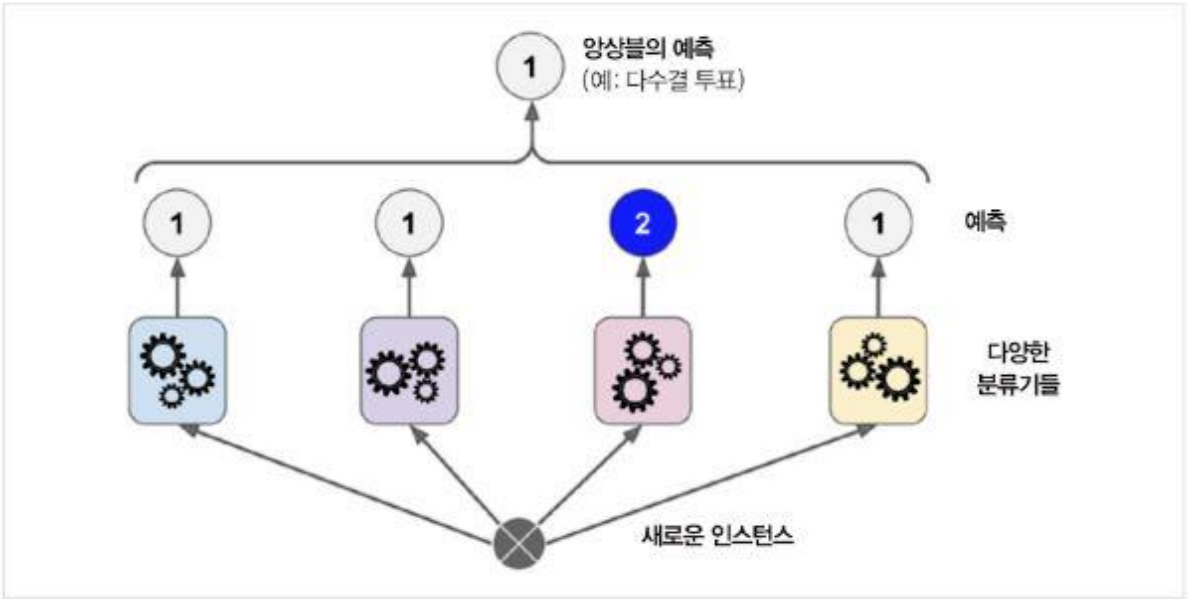
▶ 투표 기반 분류기

- 직접 투표(Hard Voting): 다수결 투표로 정해지는 분류기



여러 분류기 훈련시키기

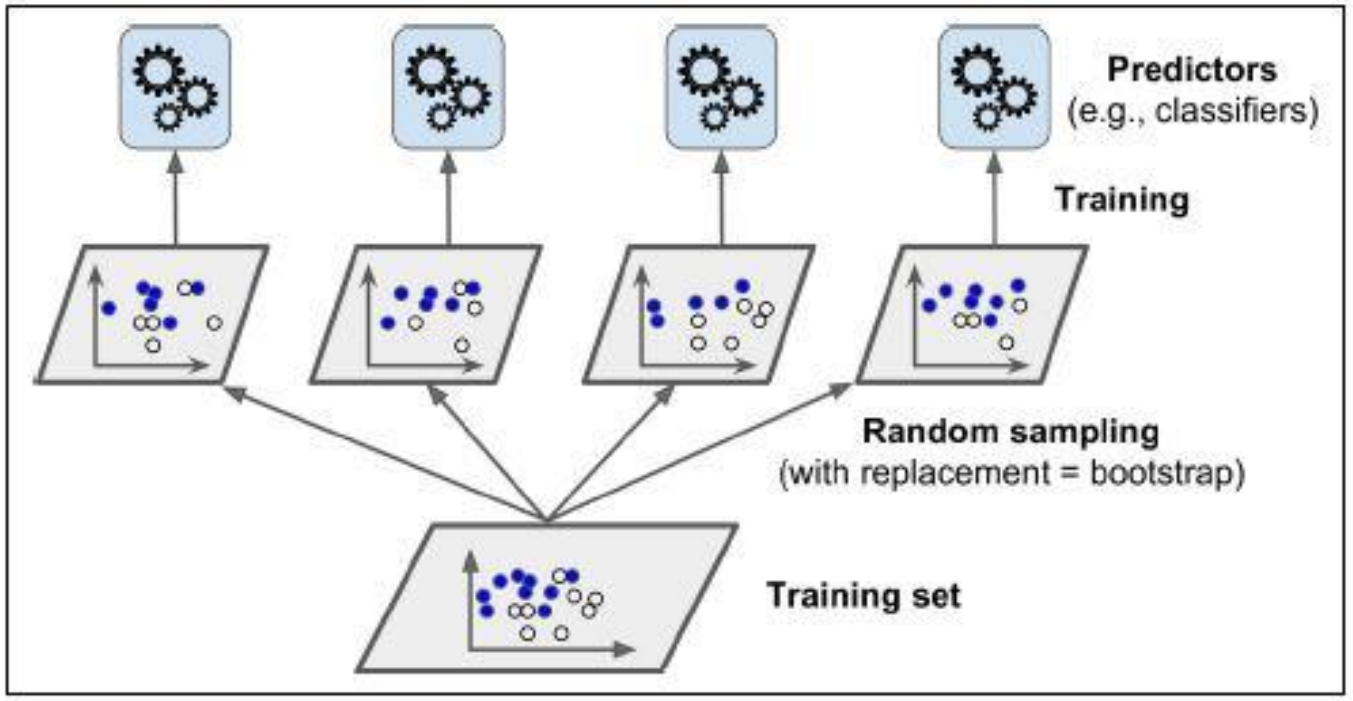
그림 7-2 직접 투표 분류기의 예측



▶ **배깅과 페이스팅**

- **배깅(bagging, bootstrap aggregating):** 훈련 세트에서 중복을 허용하여 샘플링하는 방식
- **페이스팅(pasting):** 훈련 세트에서 중복을 허용하지 않고 샘플링하는 방식

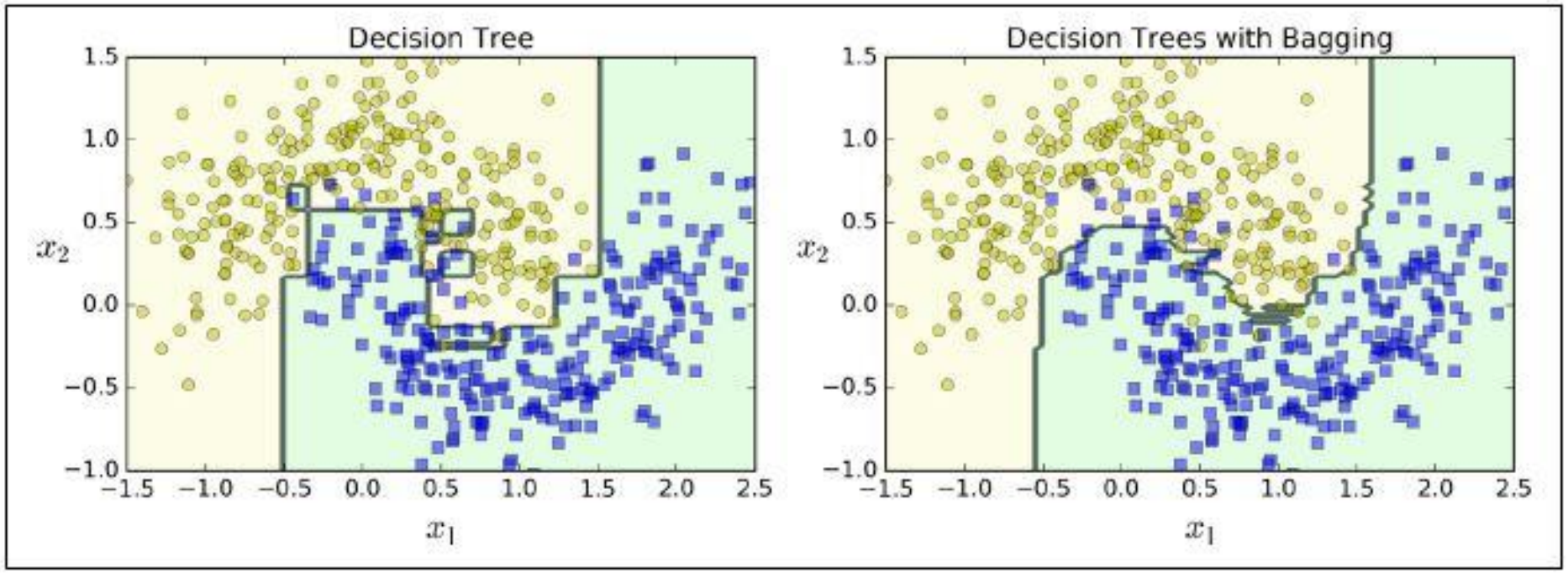
⇒ 배깅과 페이스팅에서는 같은 훈련 샘플을 여러 개의 예측기에 걸쳐 사용 가능



배깅과 페이스팅은 훈련 세트에서 무작위로 샘플링하여 여러 개의 예측기를 훈련

▶ 사이킷런의 배깅과 페이스팅

- 아래의 그림을 보면 앙상블의 예측이 단일 결정 트리 예측보다 일반화가 잘 된 것을 확인 할 수 있다.



단일 결정 트리(왼쪽)와 500개의 트리로 만든 배깅 앙상블(오른쪽) 비교

4. 랜덤 포레스트

▶ 랜덤 포레스트

- 일반적으로 **배깅 방법(또는 페이스팅)을 적용한 결정 트리 앙상블**
- BaggingClassifier에 DecisionTreeClassifier를 넣어 만드는 대신 결정 트리에 최적화되어 사용하기 편리한 RandomForestClassifier를 사용할 수 있다.
- 트리의 노드를 분할할 때, 전체 특성 중에서 최선의 특성을 찾는 대신 무작위로 선택한 특성 후보 중에서 최적의 특성을 찾는 식으로 무작위성을 더 주입한다.
  → 트리를 더 다양하게 만든다.
  → 편향을 손해보는 대신 분산을 낮추어 전체적으로 더 훌륭한 모델을 만들어 낸다.

▶ 엑스트라 트리

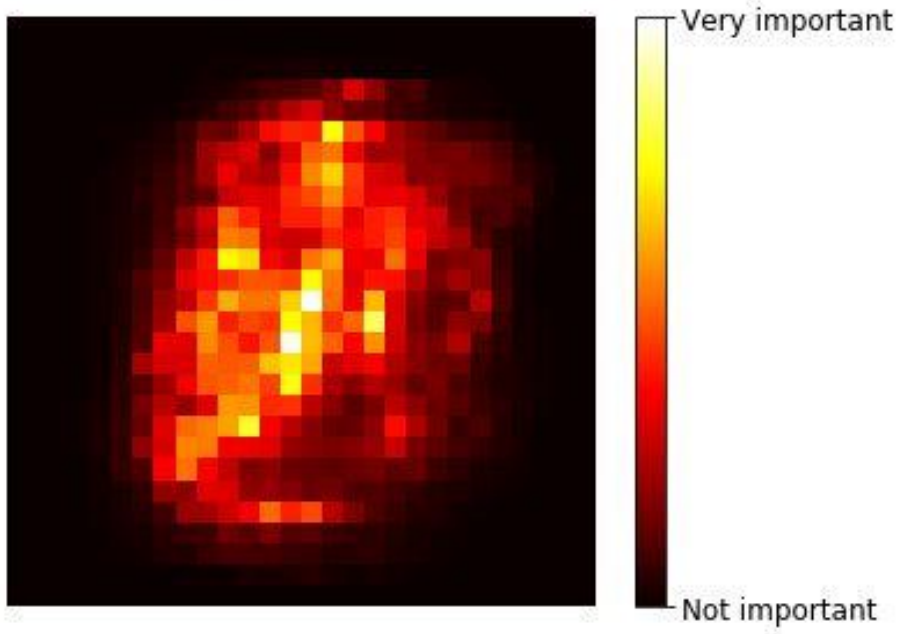
- **익스트림 랜덤 트리 앙상블**(=엑스트라 트리): 극단적으로 무작위한 트리의 랜덤 포레스트
- 엑스트라 트리를 만들기위해 사이킷런의 ExtraTreesClassifier를 사용(RandomForestClassifier과 사용법 동일)

4. 랜덤 포레스트

▶ **특성 중요도**

- 랜덤 포레스트는 **상대적 중요도를 측정**하기 쉽다.
- 사이킷런은 어떤 특성을 사용한 노드가 평균적으로 불순도를 얼마나 감소시키는지 확인해 특성 중요도를 측정한다.
  → 가장치 평균이며, 각 노드의 가중치는 연관된 훈련 샘플수와 같다.

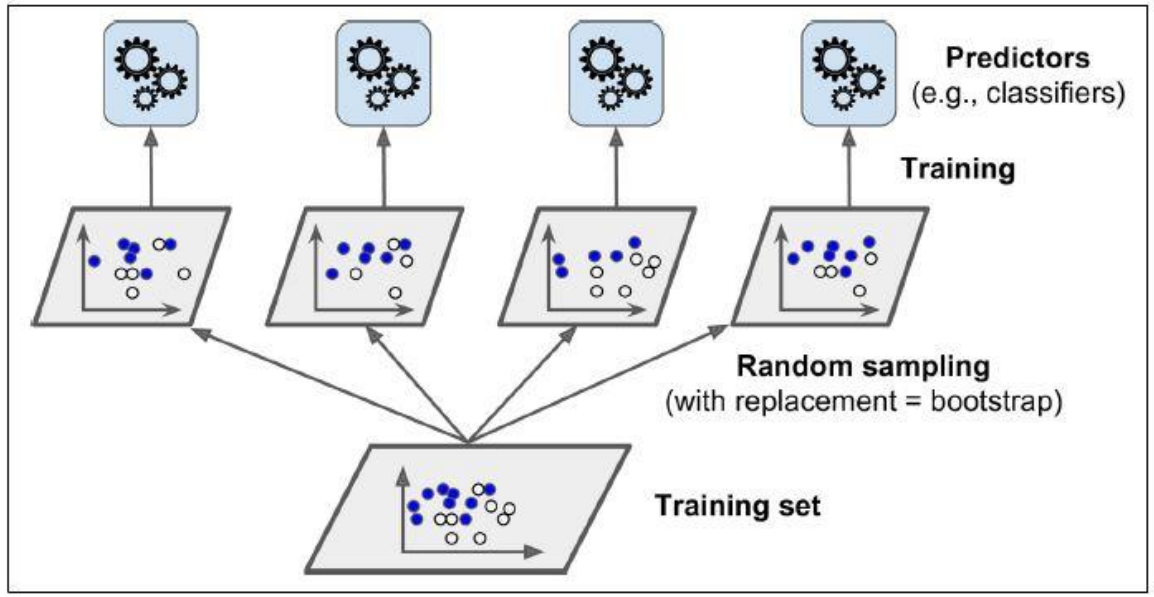
⇒(Conclusion) 랜덤 포레스트는 특성 선택 시, 어떤 특성이 중요한지를 빠르게 확인할 수 있어 매우 편리하다.



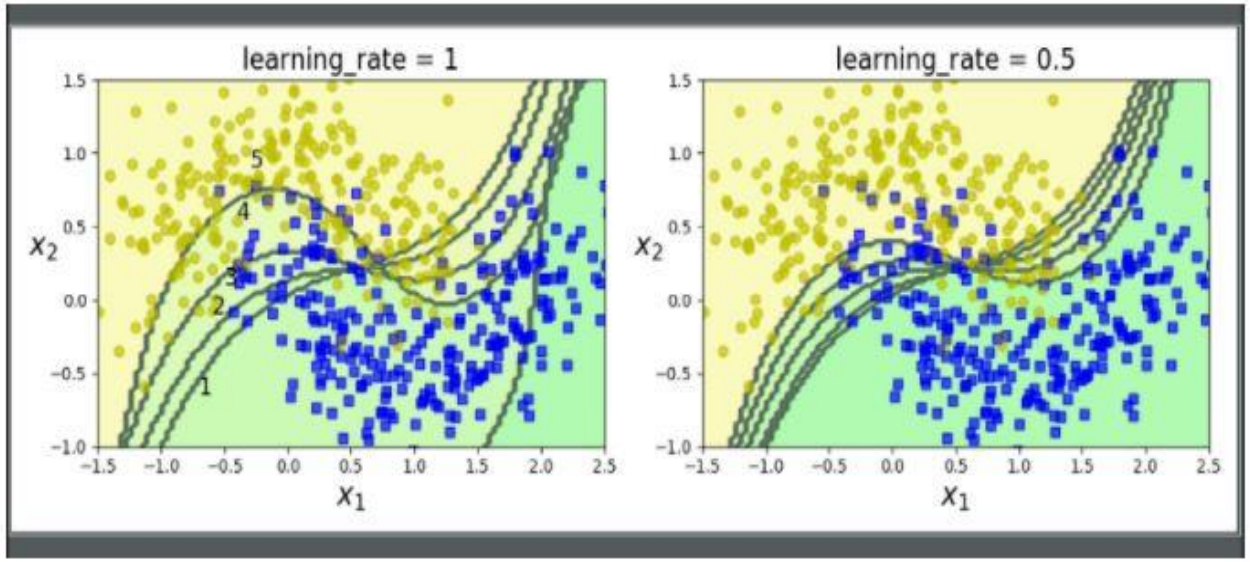
랜덤 포레스트 분류기에서 얻은 MNIST 픽셀 중요도

▶에이다 부스트

- 이전 예측기를 보완하는 새로운 예측기를 만드는 방법
⇒ 이전 모델이 과소적합했던 훈련 샘플의 가중치를 더 높이는 것
- 새로운 예측기는 학습하기 어려운 샘플에 점점 더 맞춰지게 됨

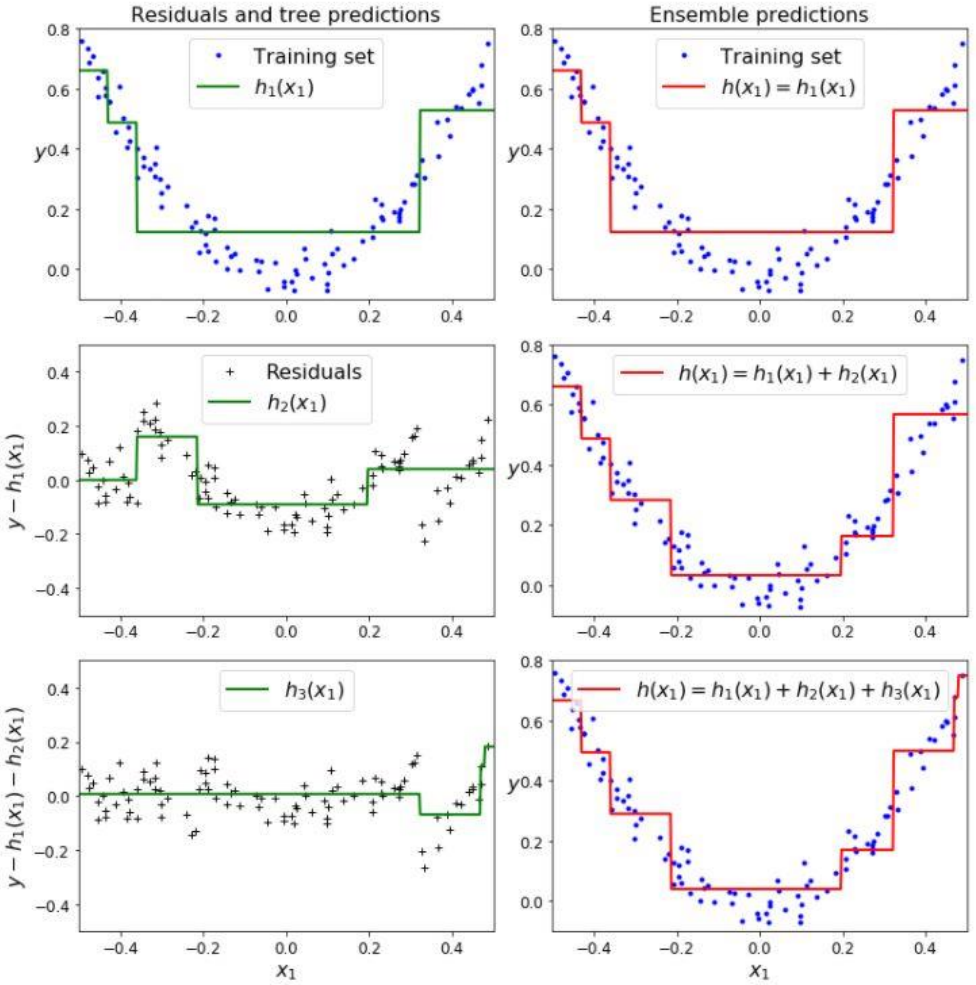


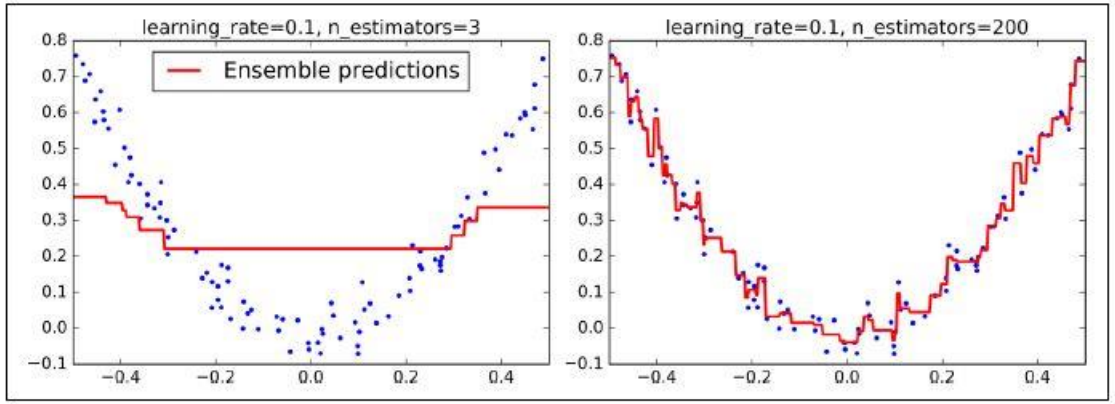
샘플의 가중치를 업데이트하면서 순차적으로 학습하는 에이다부스트



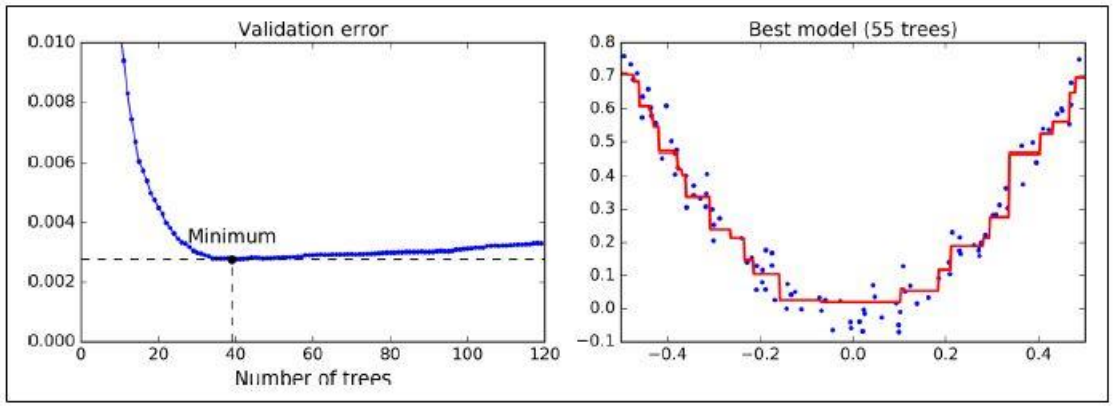
훈련 수가 늘어갈 수록 정확한 그래프가 만들어지는 것을 볼 수 있다.

▶그레디언트





예측기가 부족한 경우와 많은 경우의 GBRT 앙상블

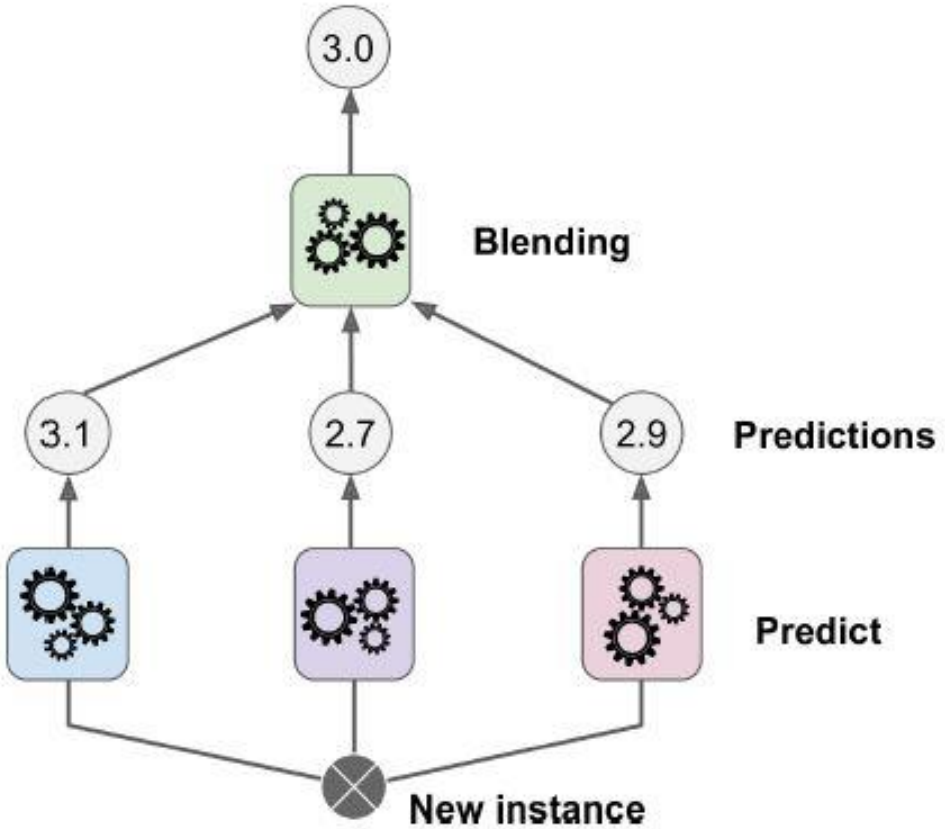


조기 종료를 사용하여 트리 수 튜닝

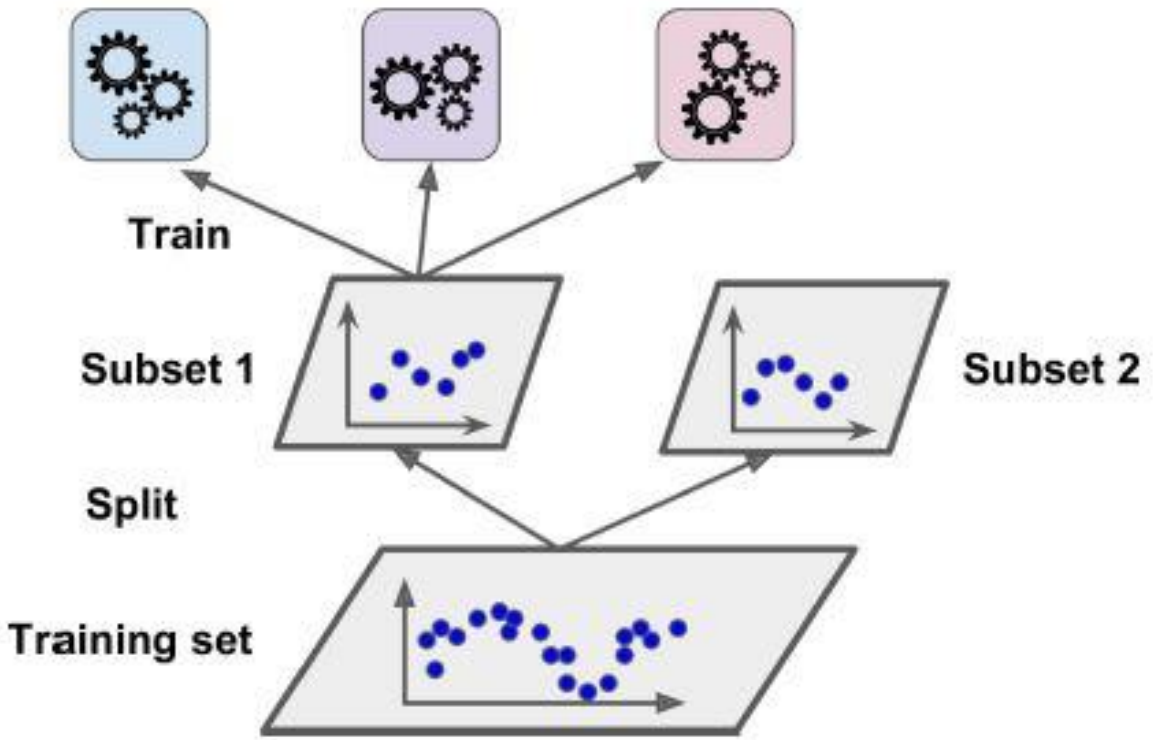
6. 스태킹

▶스태킹

: 앙상블에 속한 모든 예측기의 예측을 취합하는 간단한 함수를 사용하는 대신, 취합하는 모델을 훈련시키고자 하는 방법

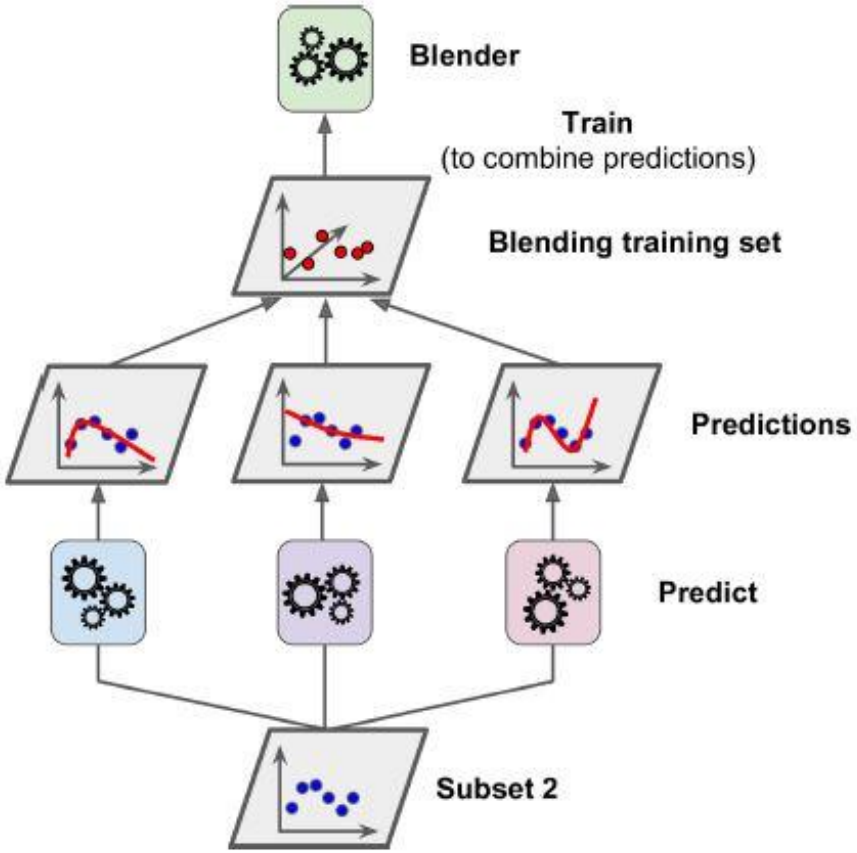


블렌딩 예측기를 사용한 예측 취합

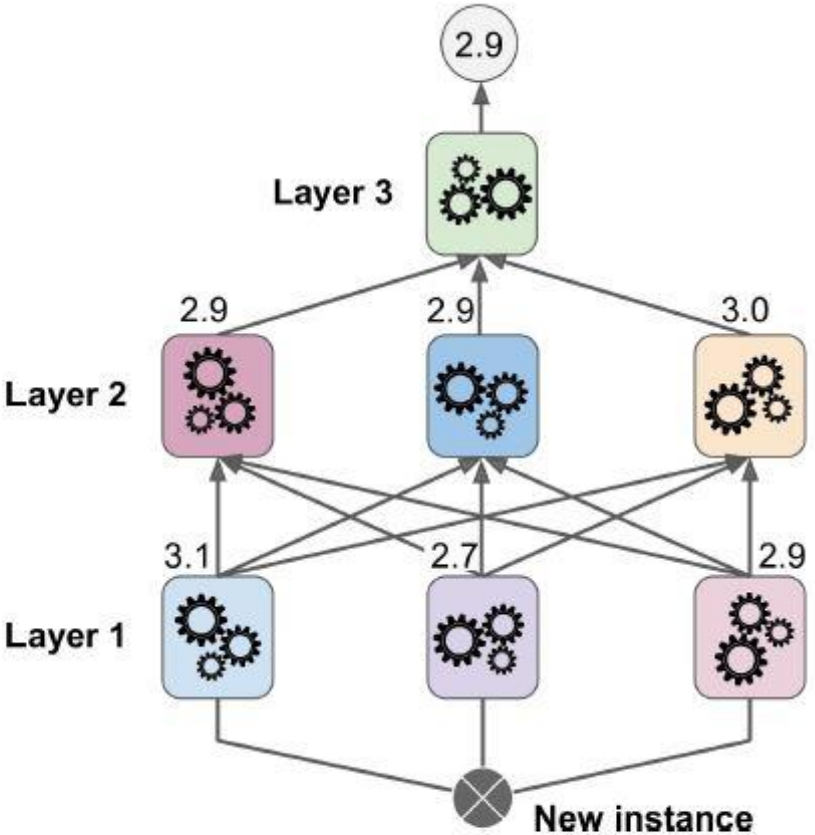


첫 번째 레이어 훈련하기

▶**블렌더 훈련**



블렌더 훈련



멀티 레이어 스태킹 앙상블의 예측