Euron 3주차 예습과제_개념정리_한송희

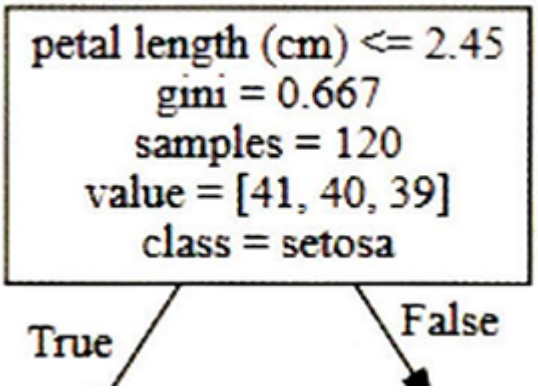
Chapter4 분류

01. 분류(classification)의 개요

classification: 지도학습의 대표적 유형으로 학습 데이터로 주어진 데이터의 피처와 레이블 값을 머신러닝 알고리즘으로 학습해 모델 생성→ 새로운 데이터 값이 주어졌을 때 미지의 레이블 값 예측

종류: Naive Bayes, Logistic Regression, Decision Tree, Support Vector Machine, Nearest neighbor, Neural Network, Ensemble

<시각화 실습>



petal lenght(cm): 자식 노드를 만들기 위한 피처 조건. 없으면 리프노드

gini: value=[]로 주어진 데이터 분포에서의 지니 계수

samples: 현 규칙에 해당하는 데이터 건수

value=[]: 클래스 값 기반 데이터 건수/ 0:Setosa, 1:Versicolor, 2:Virginica

02. Decision Tree

:ML알고리즘 중 직관적으로 이해가 쉬운 알고리즘. 데이터 있는 규칙을 학습을 통해 자동으로 찾아내 Tree 기반의 분류 규칙 생성

일반적으로 if/else 기반으로 나타냄

Decision Node: 규칙 조건

Leaf Node: 결정된 클래스 값

Sub Tree: 새로운 규칙 조건 마다 생성

depth: 트리의 깊이. 깊어질 수록 과적합 위험

⇒ 가능한 한 적은 결정 노드로 높은 예측 정확도를 가지기 위해 트리를 어떻게 split할지가 중요

⇒ 정보 균일도가 높은 데이터 세트를 선택할 수 있도록 규칙 조건을 만들어야함

정보 균일도 측정법: entropy를 이용한 information Gain 지수(1-entropy), 지니 계수(0:평등↔1:불평등)

과정: 1) 데이터 집합의 모든 아이템이 같은 분류에 속하는지 확인 2-1)리프 노드로 만들어서 분류 결정 2-2) 데이터를 분할하는데 가장 좋은 속성과 분할 기준 찾음 3) 해당 속성과 분할 기준으로 데이터 분할하여 Branch 노드 생성 4) Recursive 하게 모든 데이터 집합의 분류가 결정될 때까지 수행

장점: 쉽다, 직관적이다, 피처의 스케일링이나 정규화 등의 사전 가공 영향이 크지 않다

단점: 과적합으로 알고리즘이 떨어질 수 있기 때문에 트리의 크기를 사전에 제한하는 튜닝이 필요하다

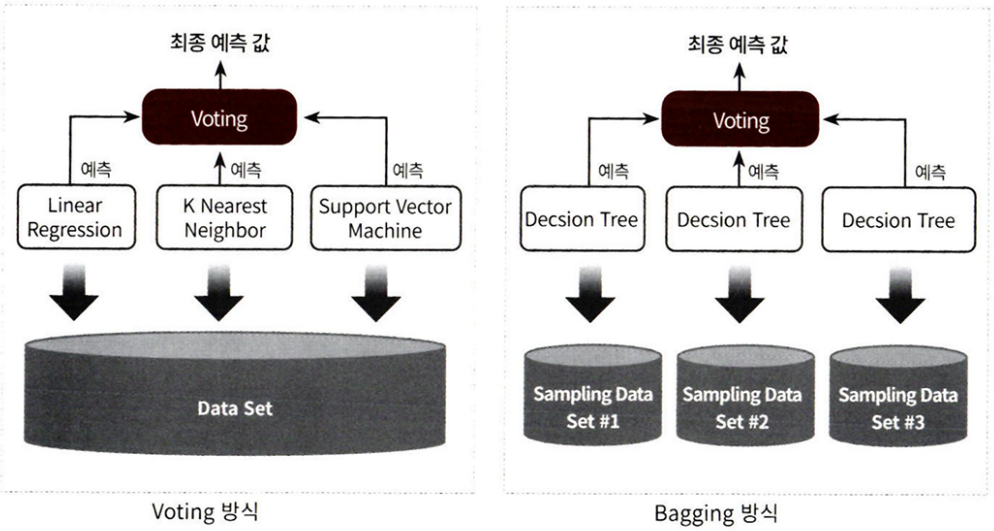
<실습>

코드 파일 참고

03. Ensemble Learning

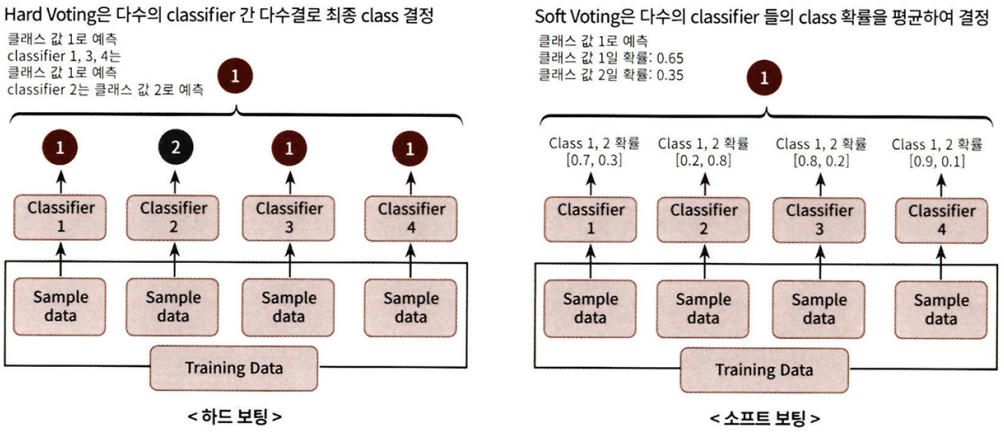
:여러 개의 분류기(classifier)를 생성하고 그 예측을 결합함으로써 보다 정확한 최종 예측을 도출하는 기법

유형:



Voting 방식 Bagging 방식

-Voting: 서로 다른 알고리즘을 가진 분류기를 결합 후 투표를 통해 최종예측 결과 결정



- Hard Voting: 다수의 분류기 간 다수결로 최종 class 결정
- Soft Voting: 다수의 분류기들의 class확률을 평균하여 결정

-Bagging: 같은 유형의 알고리즘이지만 데이터 샘플링 서로 다르게 하여 voting을 수행 (e.g. RandomForest)

-Boosting: 여러 개의 분류기가 순차적으로 학습을 수행하되, 앞에서 학습한 분류기가 예측이 틀린 데이터에 대해서는 올바르게 예측할 수 있도록 다음 분류기에게는 weight(가중치)를 부여하면서 학습과 예측을 진행(e.g. XGBoost, LightGBM)

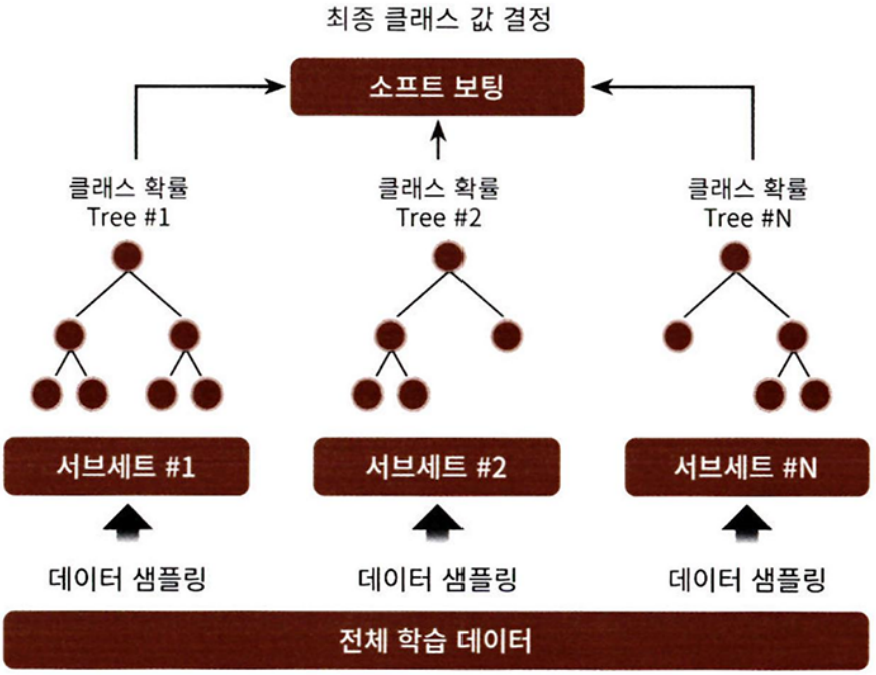
-Stacking: 여러 가지 다른 모델의 예측 결과값을 다시 학습 데이터로 만들어서 다른 모델로 재학습시켜 결과를 예측

04. Random Forest

: 같은 알고리즘으로 여러 개의 분류기를 만들어 보팅으로 최종 결정하는 bagging의 대표적 예시

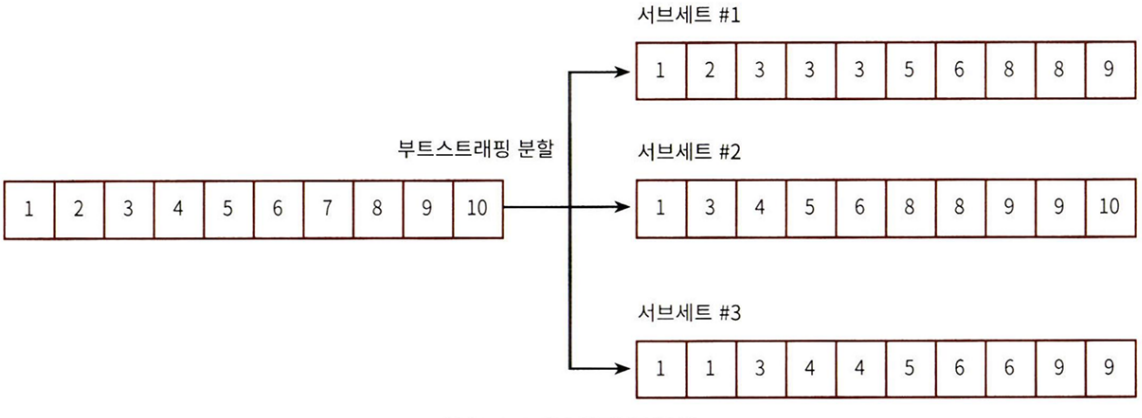
빠르고 다양한 영역에서 높은 예측 성능을 보임

기반알고리즘: decision tree (직관적인 장점)



〈 랜덤 포레스트 〉

bootstrapping: 여러 개의 데이터 세트를 중첩되게 분리



〈 부트스트래핑 샘플링 방식 〉

n_estimators: 결정 트리 개수

max_features: 결정 트리에 사용된 max_features 파라미터와 같음

max_depth, min_samples_leaf 같은 결정 트리 과적합 방지 파라미터가 랜덤 포레스트에도 적용될 수 있음