[Data Science] 데이터 사이언스 개념 - 5.앙상블 학습 — 나무늘보의 개발 블로그

노트북: 첫 번째 노트북
만든 날짜: 2021-01-18 오후 5:03
URL: https://continuous-development.tistory.com/215?category=833358

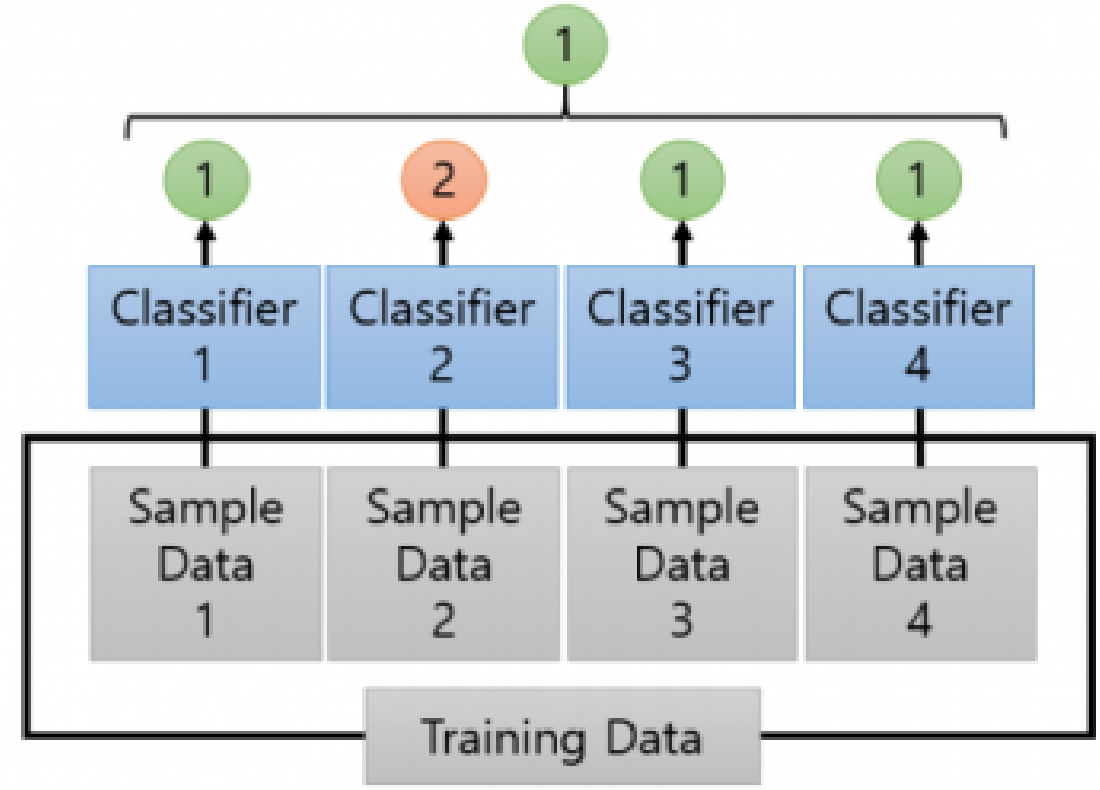
Data Science

[Data Science] 데이터 사이언스 개념 - 5.앙상블 학습

2021. 1. 14. 04:55 수정 삭제 공개

앙상블 학습

1.앙상블 학습이란



앙상블 학습 - 성능이 나오지 않는 모델을 잘 조합함으로써 강력한 성능을 끌어내는 기법
앙상블 학습을 할 때 약한 학습기로서 자주 선택되는 것이 트리라고 불리는 기법이다.
트리는 회귀 문제를 대상으로 할 경우 회귀 트리라고 부르고, 두 값을 다룰 경우는 결정트리라고 부른다.
트리의 경우 과적합하는 경향이 있는데 앙상블을 통해 정밀도를 높일 수 있다.

2.회귀 트리

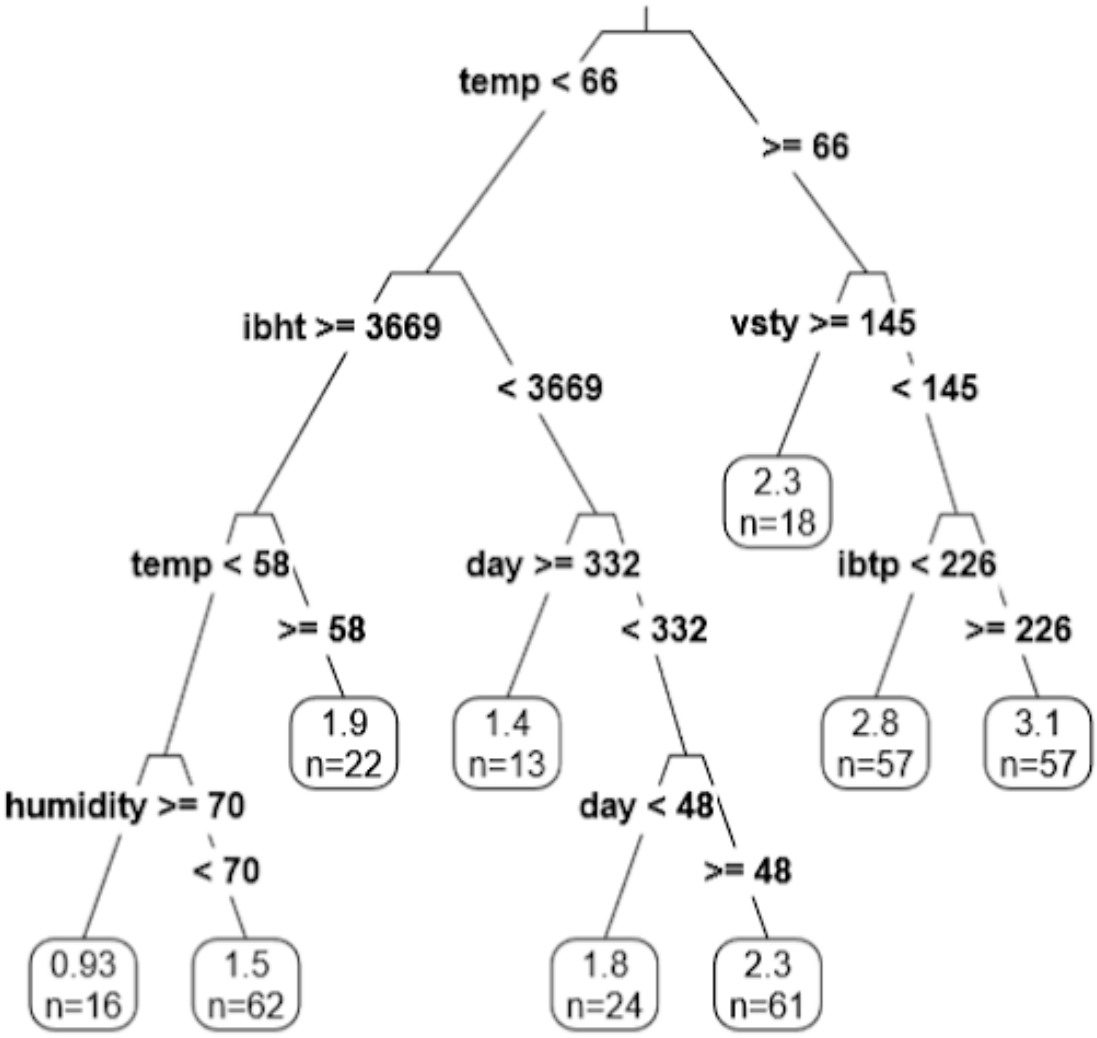


Figure 7.10: Regression tree for the ozone data set.

회귀 트리 - 특징량을 이용해 데이터를 몇개의 그룹으로 나누고 그룹의 편균 값을 예측값으로 하는 방법
회귀트리는 해석성이 높아 선형회귀로는 파악할 수 없는 관계를 추출 할 수도 있다.
회귀트르니는 과적합 되기 쉬우므로 교차검증법 등으로 트리를 가지치기 하는 방법을 사용한다.

3.부트스트랩과 배깅

부트스트랩 - 가설 검증을 하거나 매트릭을 계산하기전에 random sampling을 적용하는 방법

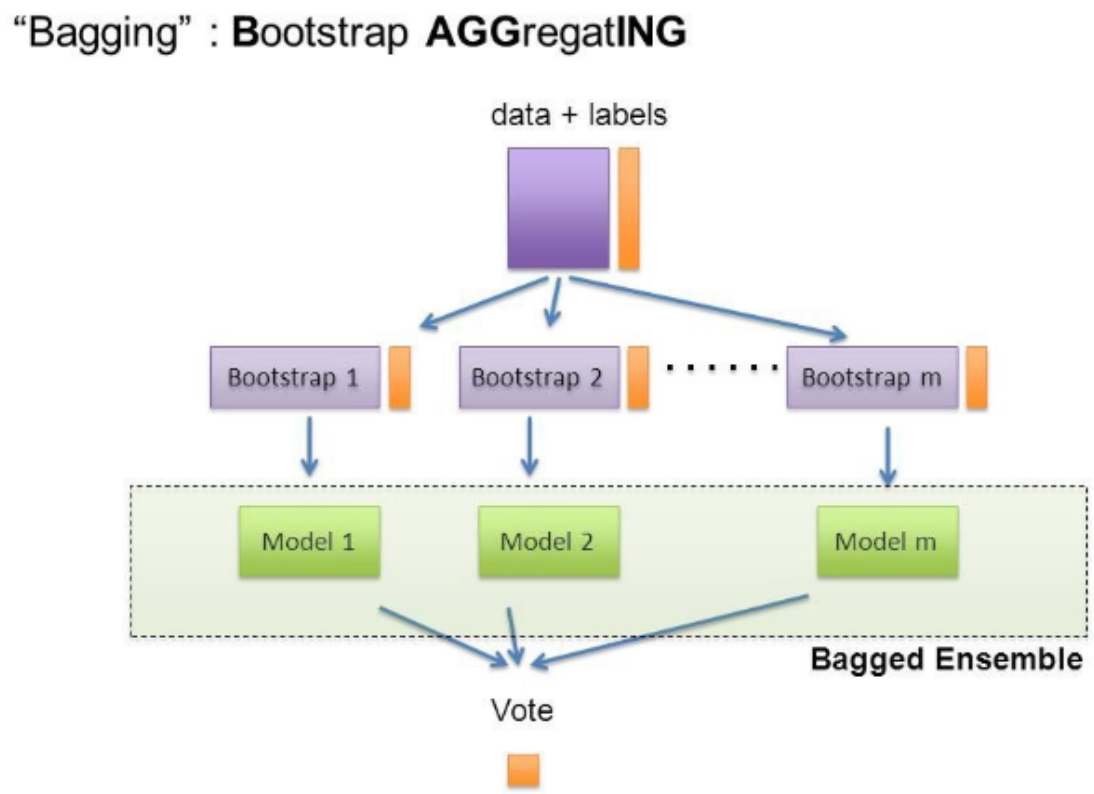
부트스트랩 방법

N개 있는 원본 데이터에서 복원 샘플링으로 N개 데이터를 샘플링

샘플링한 각 데이터로 모델 추정

이 과정을 여러 번 반복해서 추정 결과의 평균값 등을 취한다.

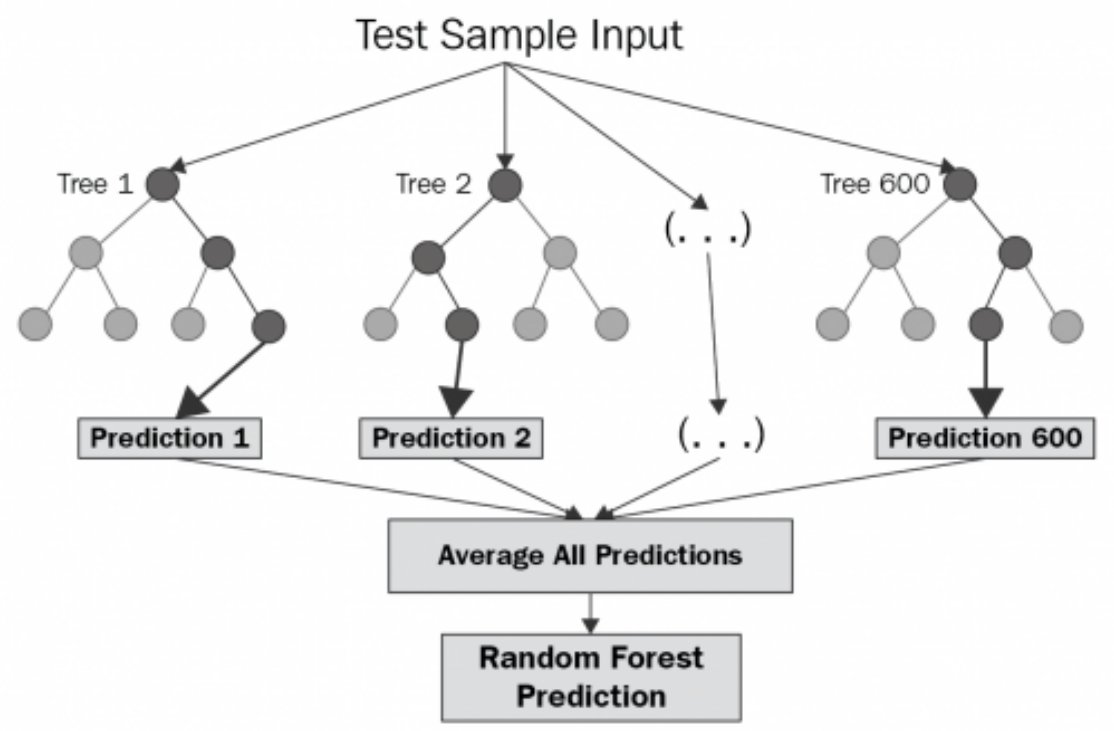
배깅 - 부트스트랩과 같은 아이디어를 이용해 의사적으로 작성한 데이터 셋에 회귀 트리를 적용해 각 모델의 에측 평균값을 모델의 예측으로 하는 것



보통 회귀트리보다 예측 정밀도가 향상 된다.
배깅의 문제점은 표본마다 추정 결과가 서로 비슷해지는 결과가 나온다.
이처럼 예측 정밀도를 높이는 것이 앙상블의 핵심이다.

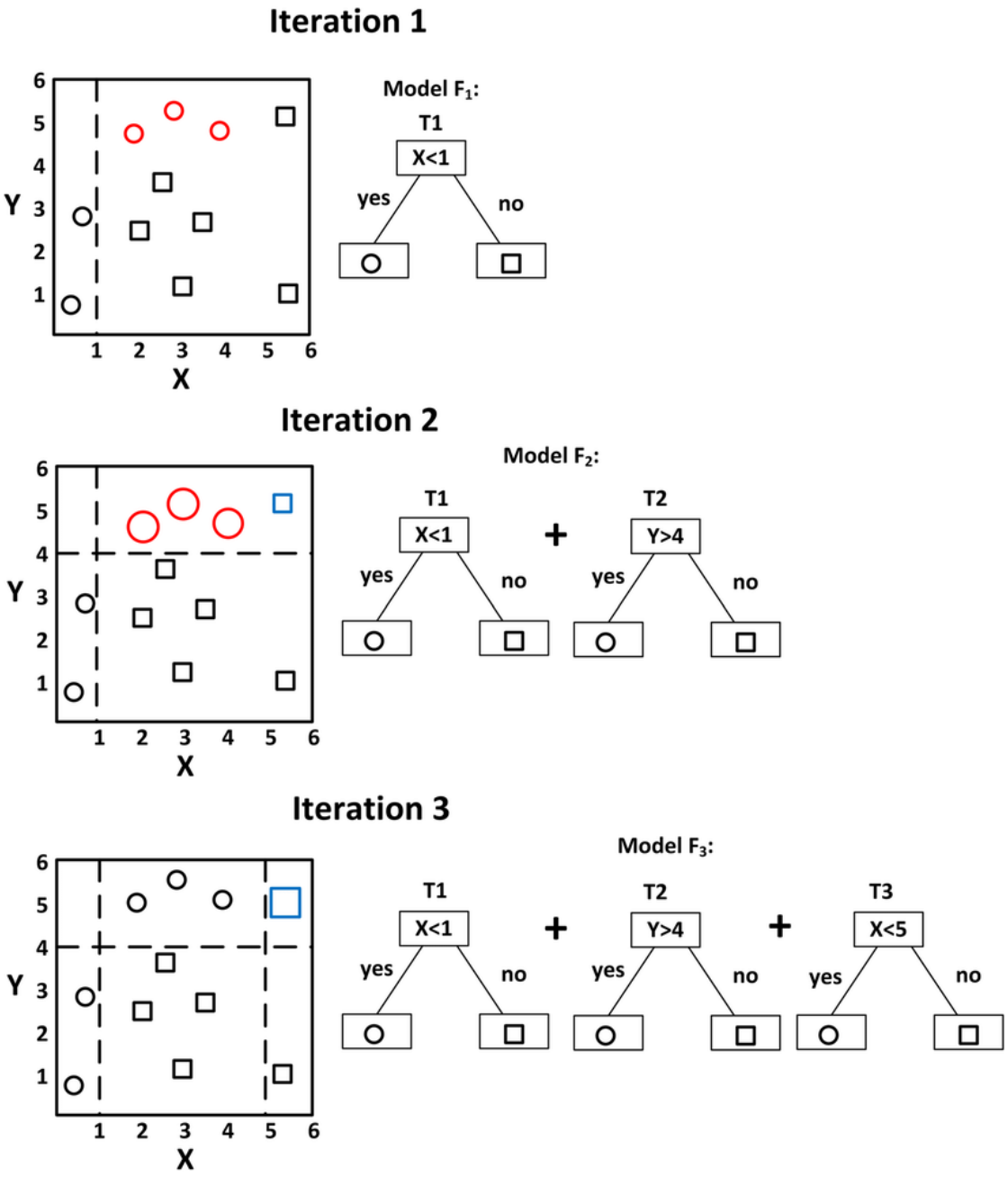
4.랜덤포레스트

랜덤 포레스트 - 표본마다 회귀트리를 추정할 때 변수도 랜덤하게 선택하는 것



보통 데이터셋에 특징량의 수가 p개 있을 때, 일반적으로 루트 p 개의 변수를 선택하면 좋은 결과를 얻을 수 있다고 한다.
보통 추정 결과마다 중요성을 계산하고 그 평균값을 변수의 중요성으로 판정한다.
부분의존성 - 모델 핵성을 위해 변수별로 부분 반응 함수를 그리는 경우

5.그래디언트 부스팅



그래디언트 부스팅 - 추정한 회귀 트리의 결과를 조합하는 방법

그래디언트 부스팅 의 핵심은 비교적 깊이가 얕은 회귀 트리를 반복적으로 학습해 조금씩 추정 정밀도를 높여가는 것이다.

위의 그림처럼 결과를 조합해 더 나은 결과를 만드는 방법이다.

지금까지의 배깅과 랜덤 포레스트는 부트스트랩에 의해 의사적으로 재구성한 표본에 대해 회귀트리를 추정했다.

주의 할 점은 트리의 최대수가 너무 크면 과적합하는 경향이 있다.

'Data Science' 카테고리의 다른 글

[Data Science] 데이터 사이언스 개념 - 7.비지도 학습

[Data Science] 데이터 사이언스 개념 - 6.분류문제

[Data Science] 데이터 사이언스 개념 - 5.앙상블 학습

[Data Science] 데이터 사이언스 개념 - 4.회귀 모델

[Data Science] 데이터 사이언스 개념 - 3.과적합과 모델 선택

[Data Science] 데이터 사이언스 개념 - 2.머신러닝의 기본

그래디언트 부스팅 배깅 부스팅 랜덤포레스트 앙상블 앙상블 학습 회귀트리

나아무늘보

혼자 끄적끄적하는 블로그 입니다.