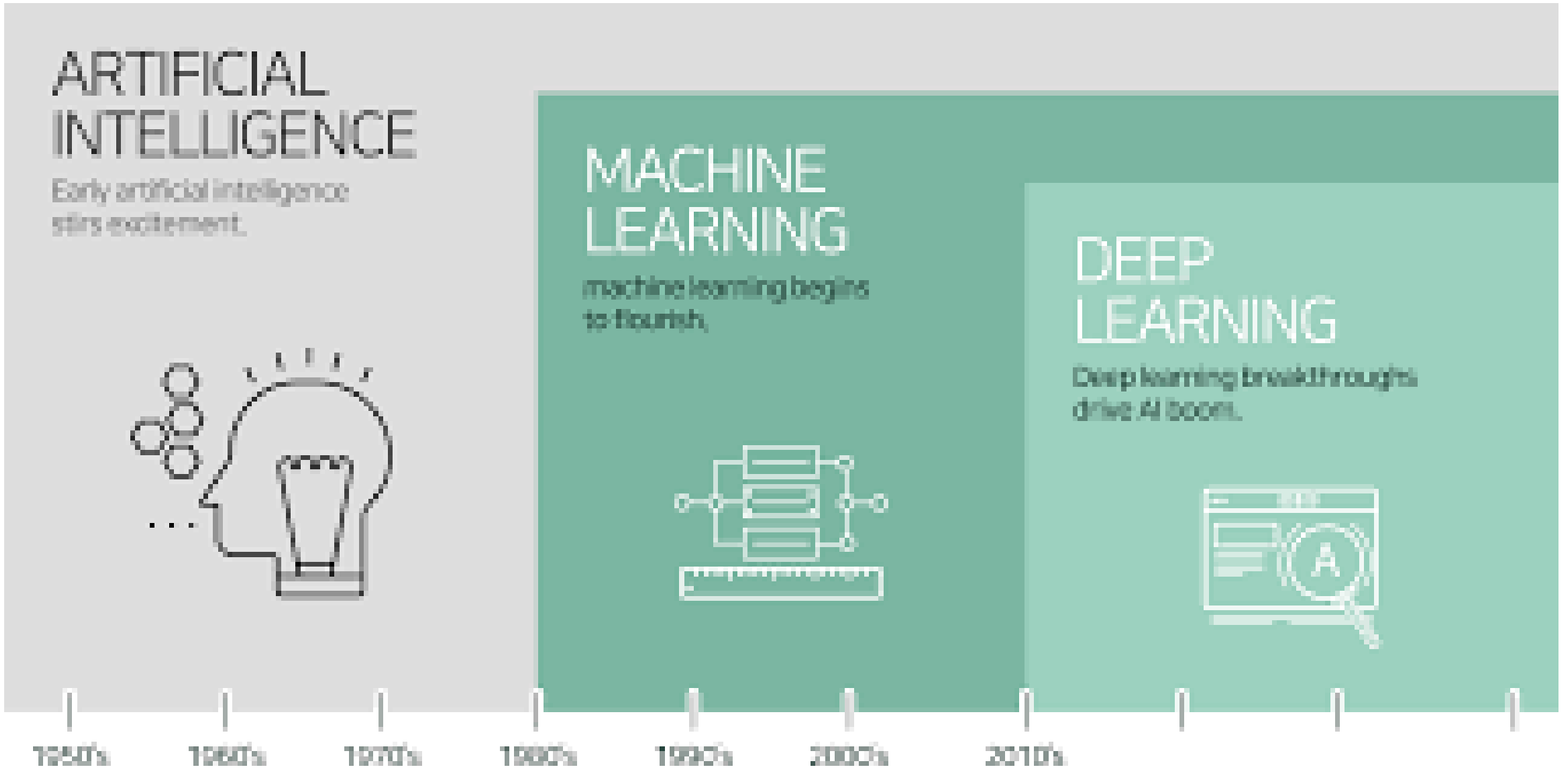
Sports Seminar

머신러닝 활용한 승률 예측

2024.1.21

Table of Contents

머신러닝이란?



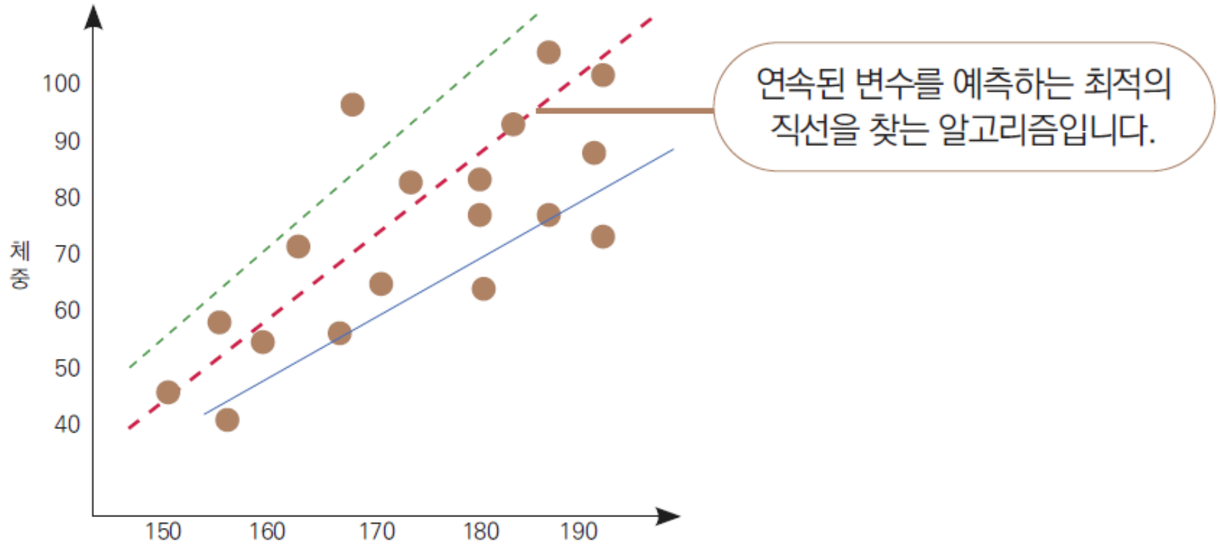
기계 스스로 데이터를 학습하여 서로 다른 변수 간의 관계를 찾아나가는 과정.

1) 예측
2) 분류
3) 군집

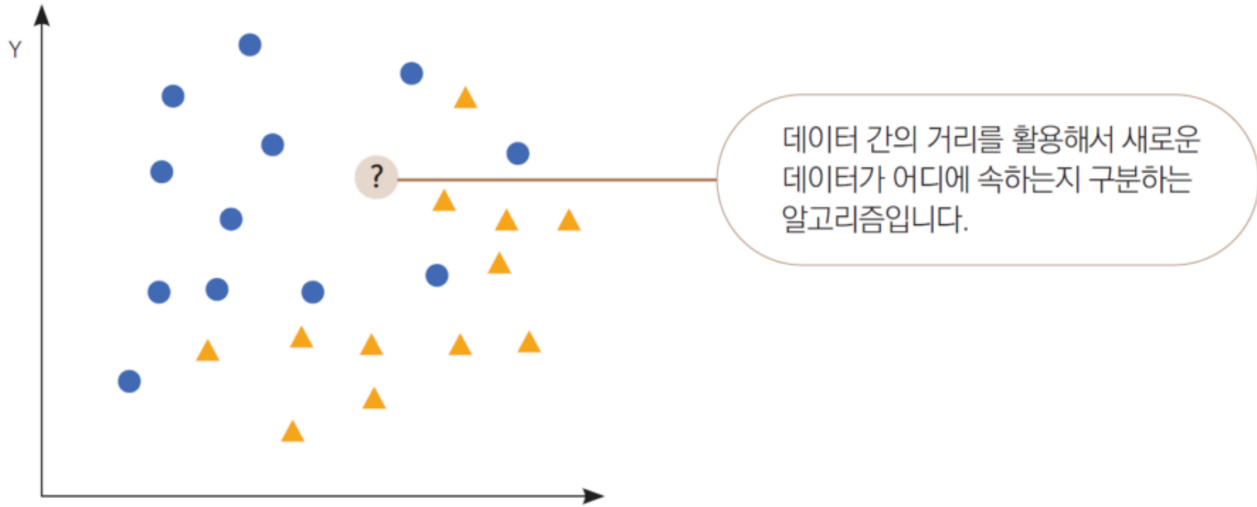
-> 사용하는 알고리즘과 방법론이 무수히 많음.

머신러닝이란?

대표적인 머신러닝 알고리즘 소개



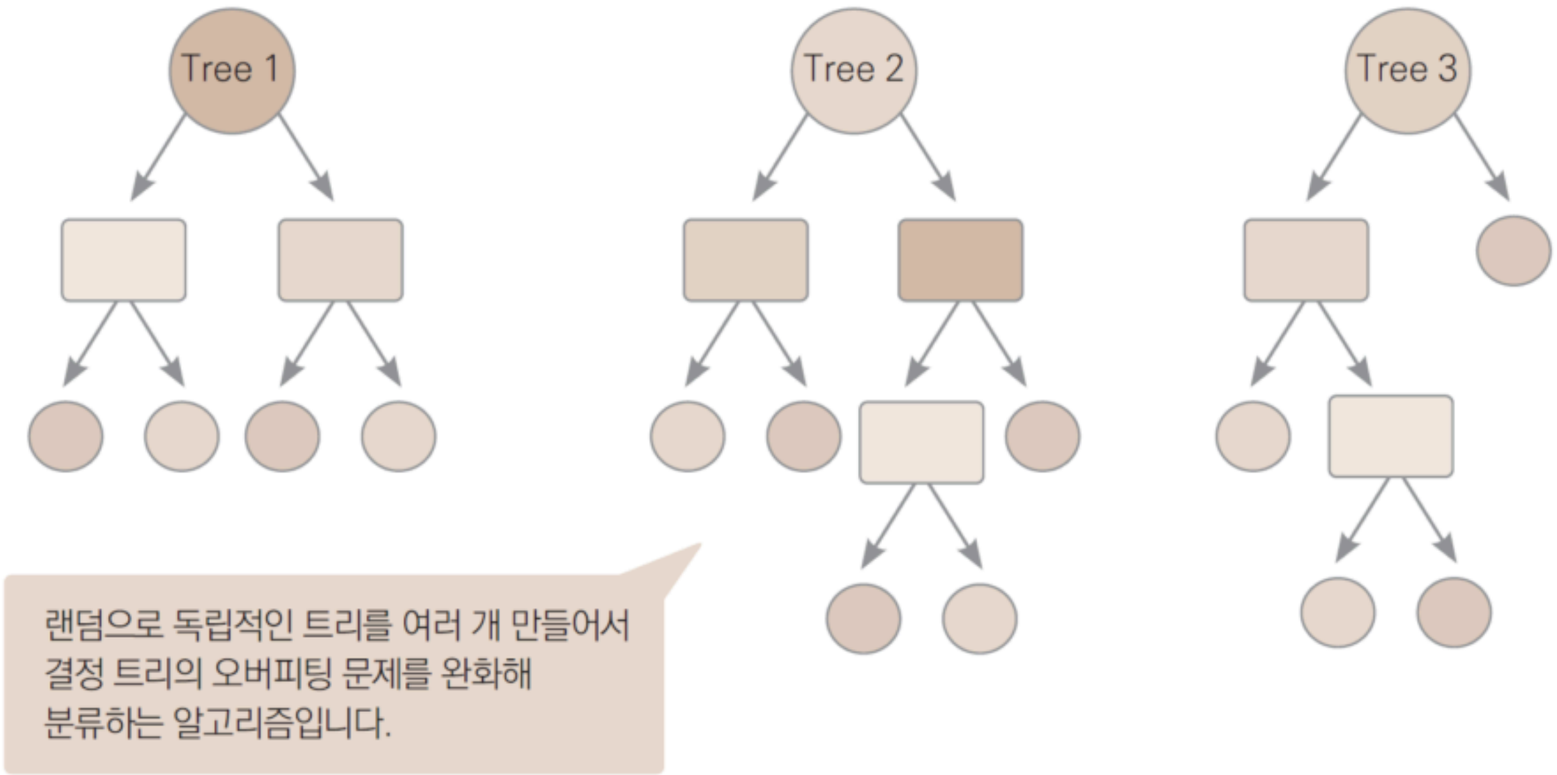
선형 회귀



KNN 알고리즘

머신러닝이란?

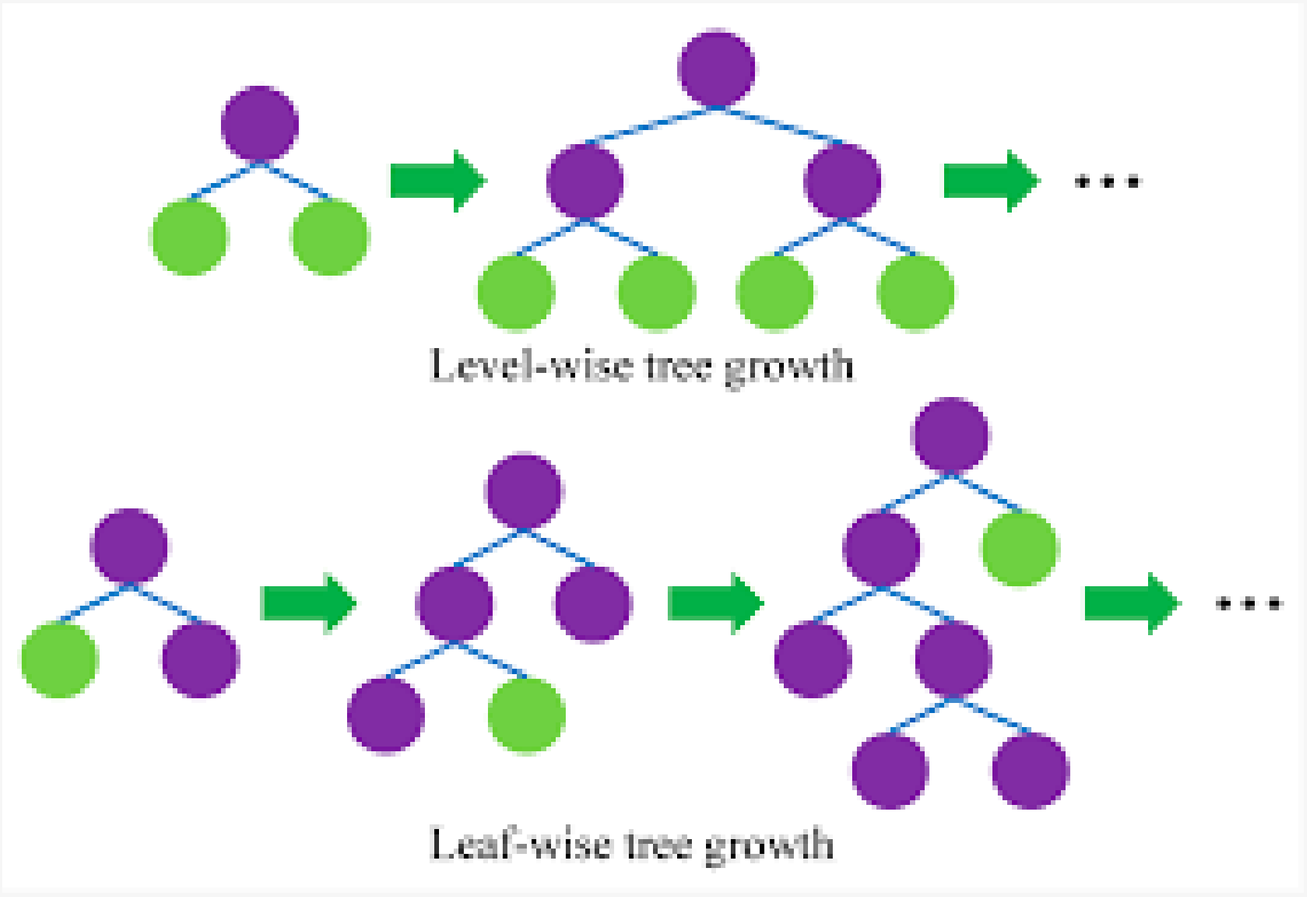
대표적인 머신러닝 알고리즘 소개



Random Forest

머신러닝이란?

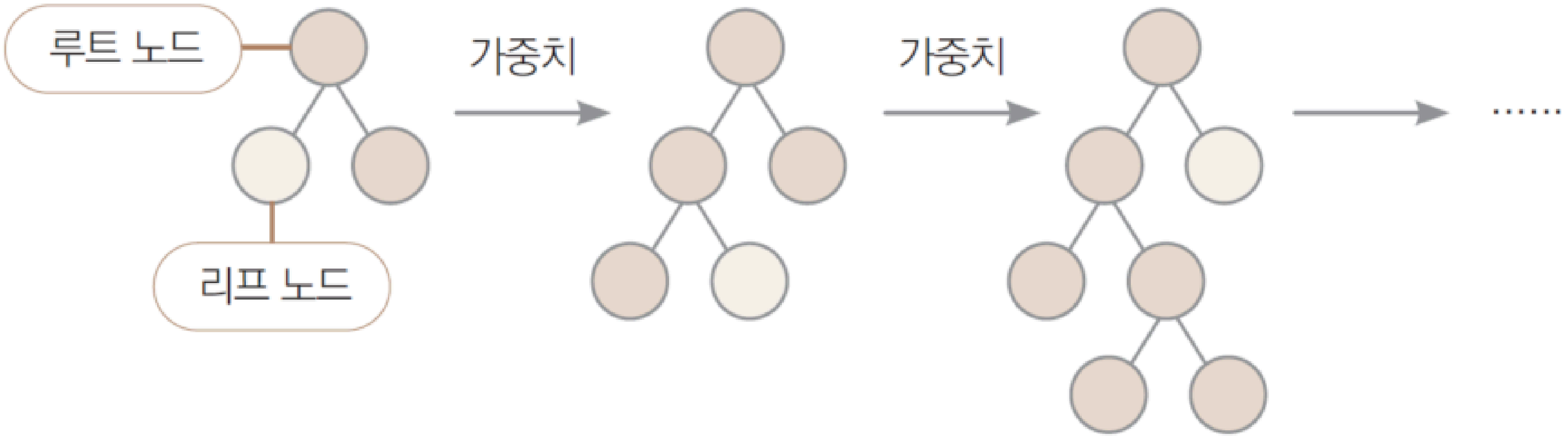
대표적인 머신러닝 알고리즘 소개



XGboost

머신러닝이란?

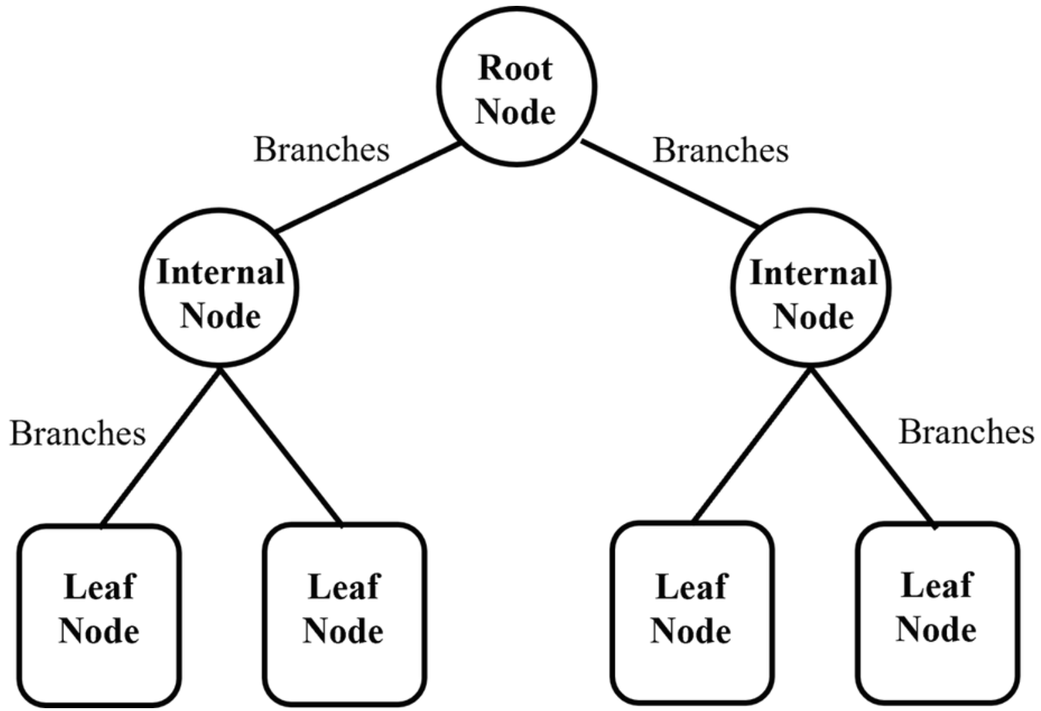
대표적인 머신러닝 알고리즘 소개



LightGBM

사용할 분석 기법

RandomForest



분류와 회귀에 사용되는 지도학습 알고리즘
-> 여러 개의 의사결정나무를 조합한 모델.
(중간마디, 끝마디)

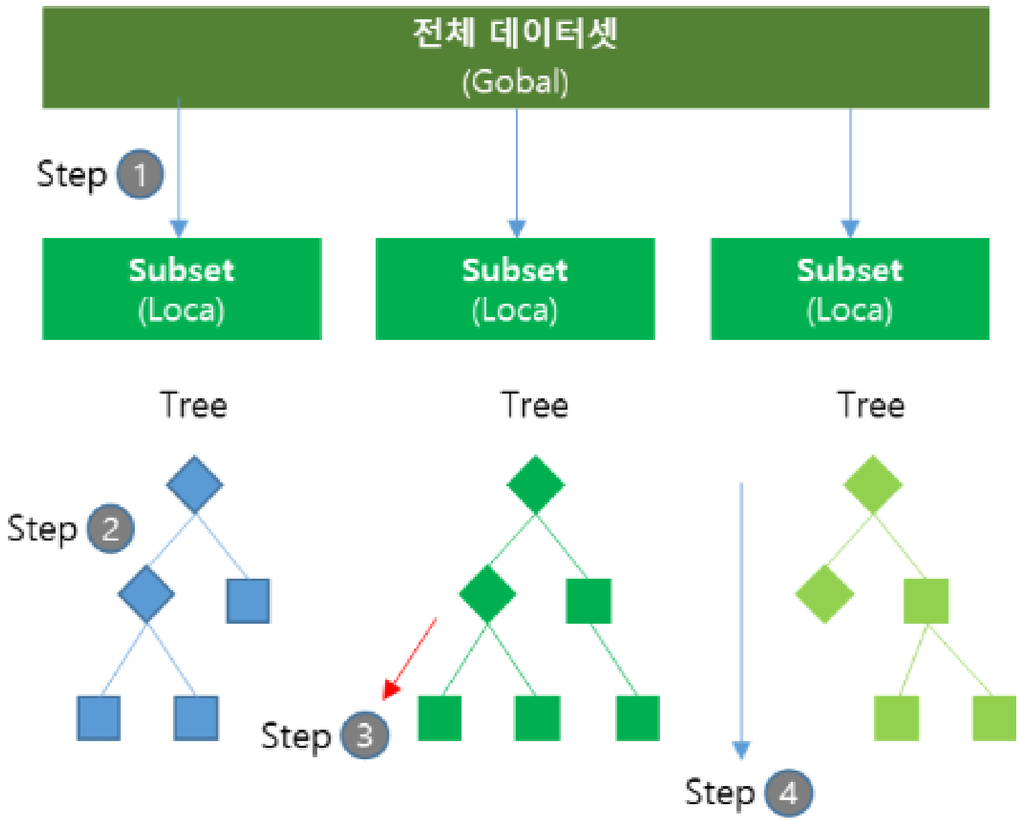
주로 분류 or 회귀에 사용! (오늘 내용은 회귀모델을 사용)

의사결정나무에 앙상블 학습을 사용(배깅)

장점 : 과적합을 해소하고 분산을 감소시켜 정확도가 높음.
단점 : 계산 비용이 높고 규칙이 많아 추론 로직을 설명하기 어려움.

사용할 분석 기법

RandomForest



기존 Gradient Tree Boosting 알고리즘에 과적합 방지를 위한 기법이 추가된 지도 학습 알고리즘.

-동작 원리-

1. 정렬된 전체 데이터셋이 가진 영역을 몇 개의 영역으로 분할
2. 분할된 데이터셋들에 대해 별도 Split Point를 찾기.
3. 각 Bucket별 기울기를 계산, Best Split Point를 찾기
4. 각 Bucker별로 병렬처리가 가능.

장점 : 과적합 방지가 잘 되며, 예측 성능이 좋음.
단점 : 작은 데이터에 과적합 가능성이 큼, 해석이 어려움.

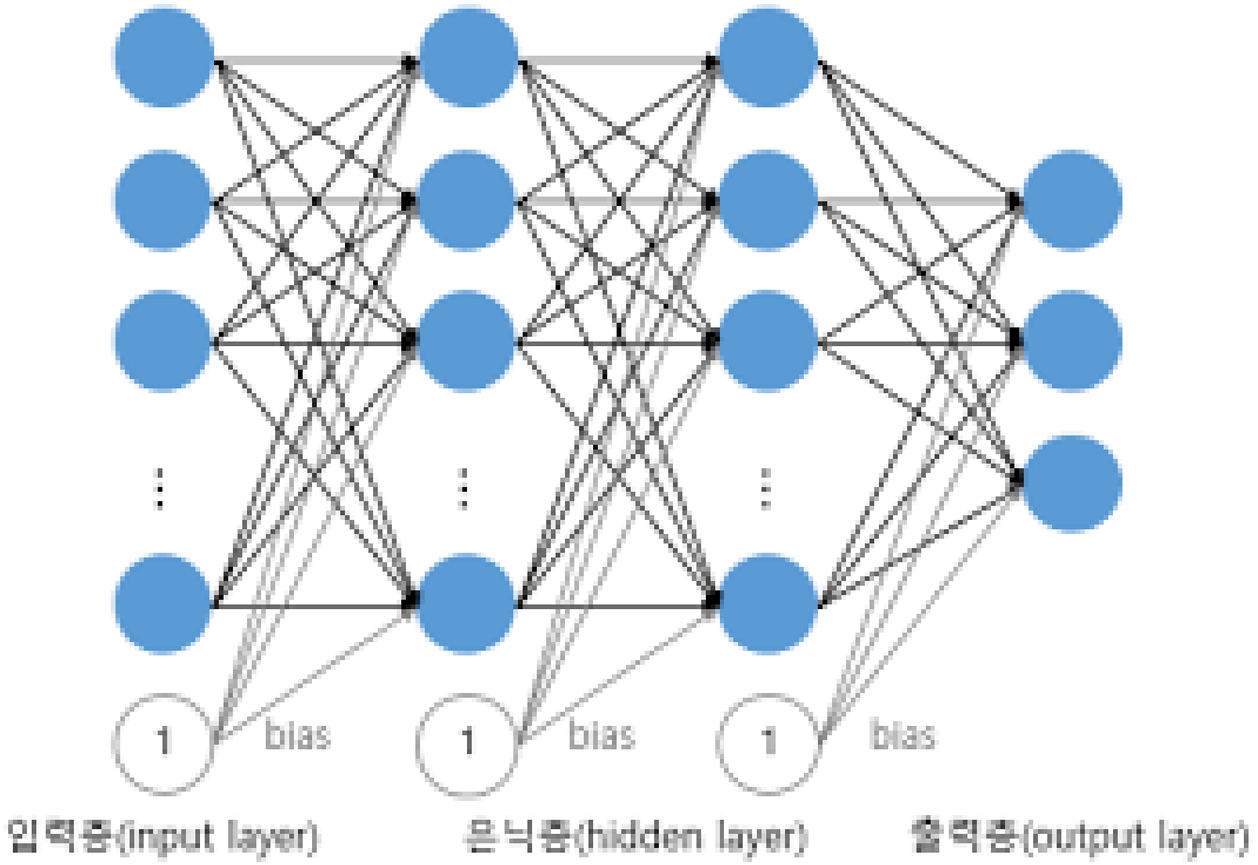
변수 조절을 위한 추가 기법

변수선택법

변수가 여러개 있을 때, 최적의 변수 조합을 찾아내는 기법.
(조합을 다 해봄으로 가장 좋은 조합을 찾기)

If, 변수가 늘어날수록 조합의 수가 기하급수적으로 늘어난다면?
-> 전체 조합을 확인하는 시간이 길어지며, 추론 속도도 늘어날 것.

해결하기 위해 합리적으로 가장 좋은 조합을 찾아내도록 함.

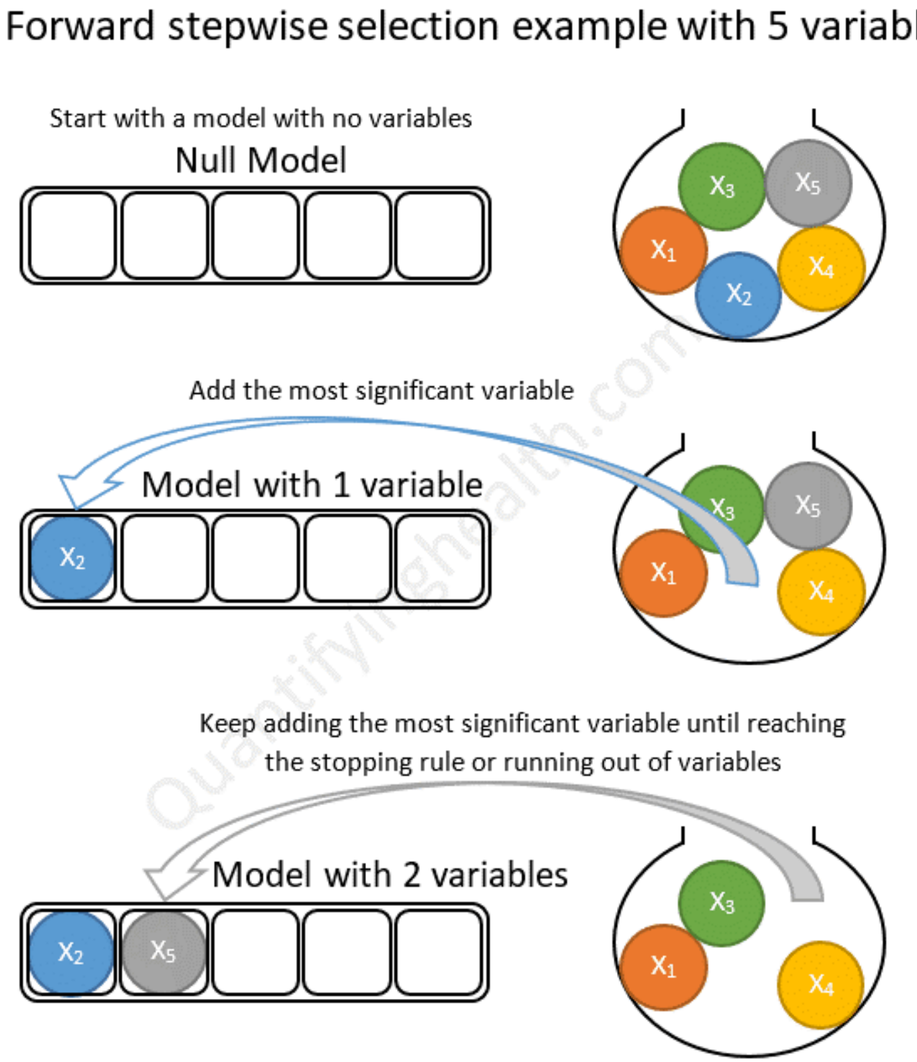


변수 조절을 위한 추가 기법

전진선택법

모델에 투입할 변수를 하나씩 추가하는 과정
-> 변수를 하나씩 순차적으로 추가!

성능지표를 비교하며, 각 변수의 유의수준을 고려해 모델에 들어갈 수 있는 변수들을 선정.

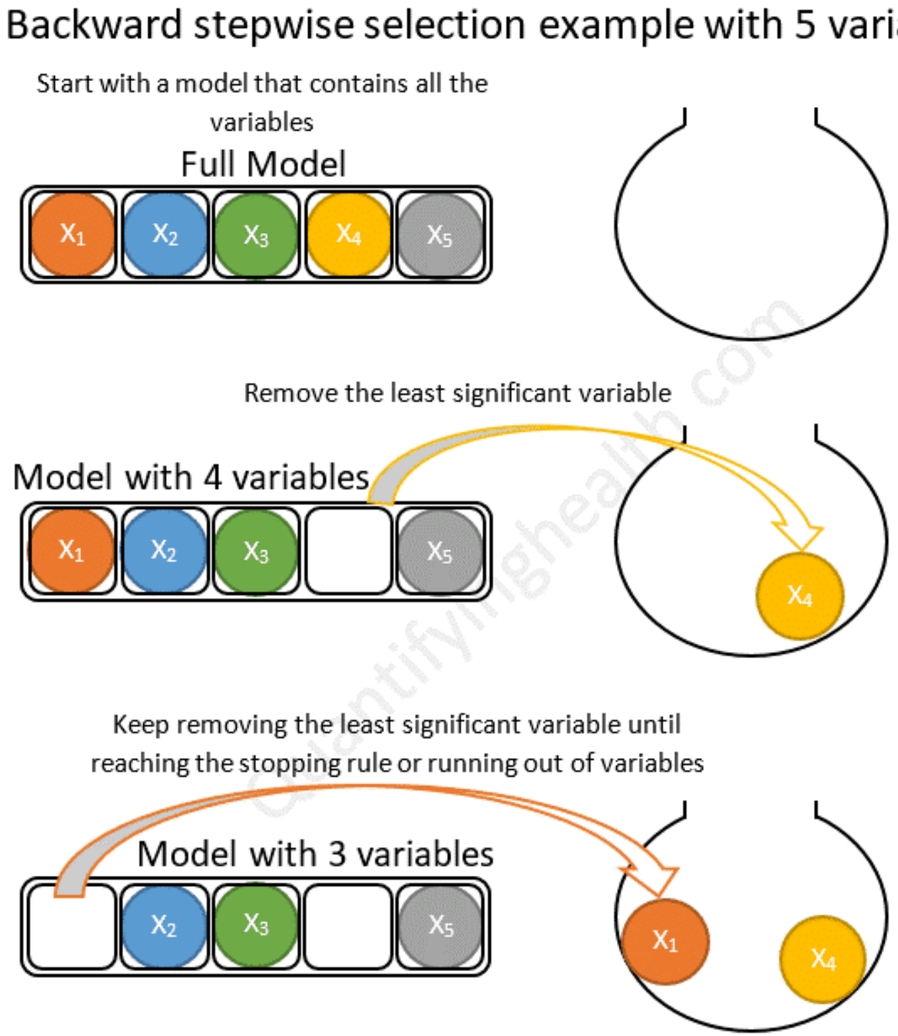


변수 조절을 위한 추가 기법

후진선택법

모델에 있는 변수를 모두 고려하여, 순차적으로 하나씩 제거하도록 진행.
-> 변수를 하나씩 제거

성능지표를 비교하며, 각 변수의 유의수준을 고려해 모델에 들어가면 안되는 변수들을 선정.

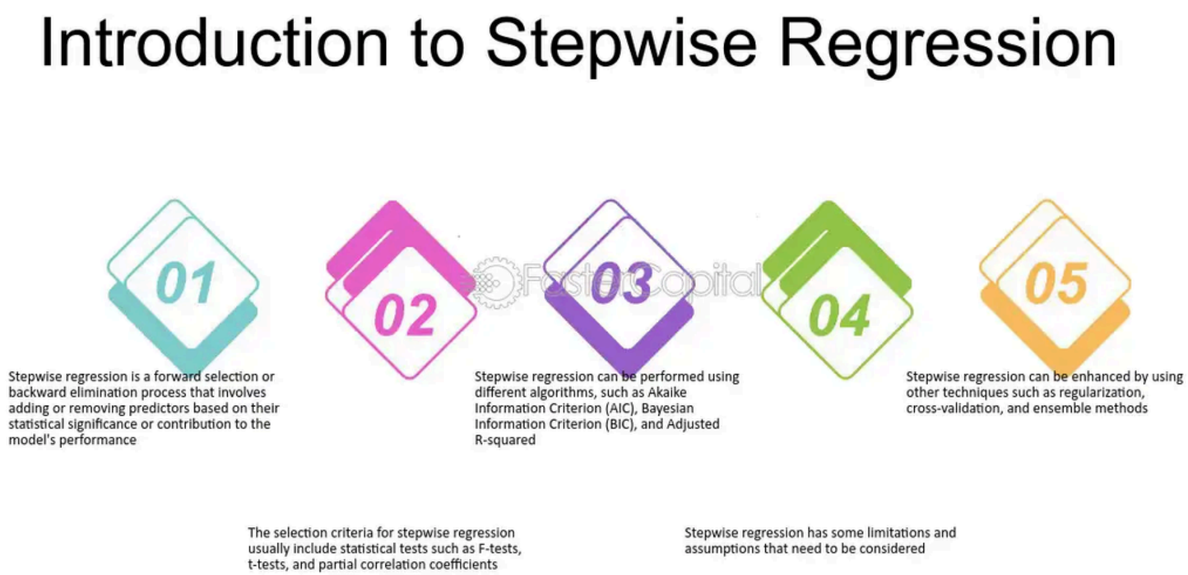


변수 조절을 위한 추가 기법

단계적 선택법

각 단계에서 변수의 중요성을 체크, 중요하지 않은 변수는 제거
+ 다시 다른 변수를 추가하는 식으로 나아가는 과정

1. 변수 입력/제거를 위해 p-value 임계치를 설정
2. 전진 선택법을 통해 변수를 설정
3. 선택된 변수 중 중요하지 않은 변수는 제거
4. 추가하거나 제거할 변수가 없을 때 종료



사용할 데이터

Korean Baseball Pitching Data (1982 - 2021)

KBO Korean Baseball Pitching Data (1982 - 2021)

Data Card | Code (4) | Discussion (0) | Suggestions (0)

About Dataset

The dataset contains team pitching data from every season of KBO Baseball.

The data was collected from Sports Reference then cleaned for data analysis.

Usability
9.71

License
CC0: Public Domain

1982~2021년까지 각 팀의 투수 기록을 담은 데이터(1차 스탯 위주로 구성)
-> 해당 데이터를 사용하여 머신러닝 모델을 구현

오늘의 목표 : 높은 정확도로 승률을 예측하는 모델을 구현해보자!

실습

사용할 데이터로 구성된 코드

Korean Baseball Pitching Data (1982 - 2021)

KBO Korean Baseball Pitching Data (1982 - 2021)

Data Card | Code (4) | Discussion (0) | Suggestions (0)

About Dataset

The dataset contains team pitching data from every season of KBO Baseball.

The data was collected from Sports Reference then cleaned for data analysis.

Usability ⓘ
9.71

License
CC0: Public Domain

1. 데이터 EDA(전처리, 시각화 코드)
2. 데이터 분석 기법 사용(변수선택법)
3. 머신러닝 코드 사용하여 모델 구현
4. 성능지표를 통한 모델 평가

감사합니다.

정보통계학과 이현섭