OT ML 기초 강의

ToBig's 8기 류호성

Machine Learning

ML 개론

Contents

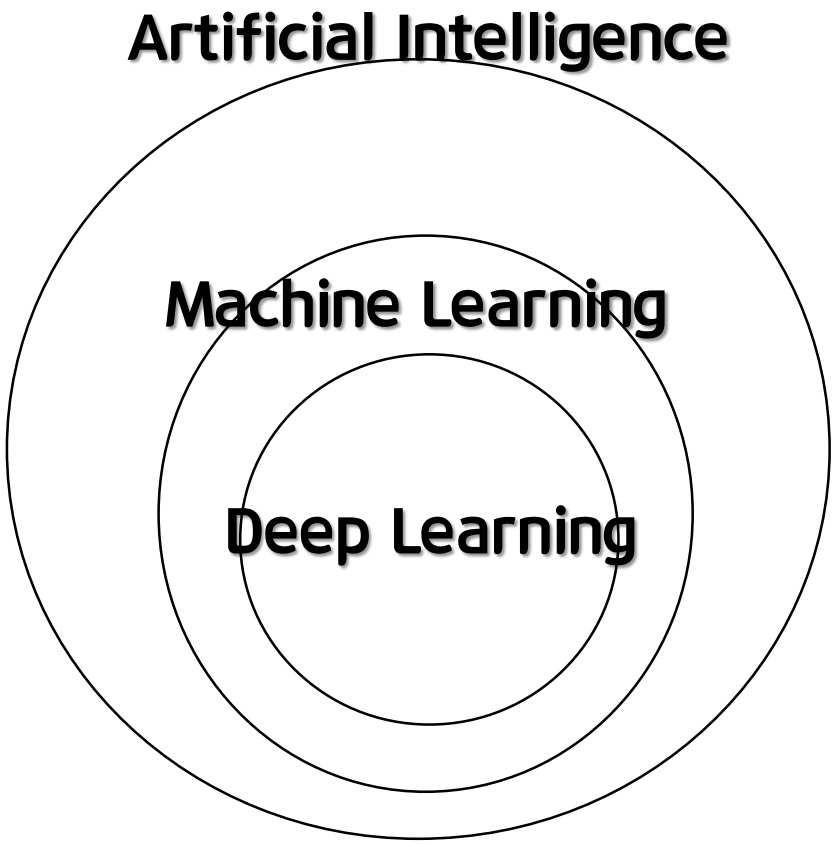
Unit 01 | Machine Learning이란?

Machine Learning

Deep Learning

Artificial Intelligence

Unit 01 | Machine Learning이란?



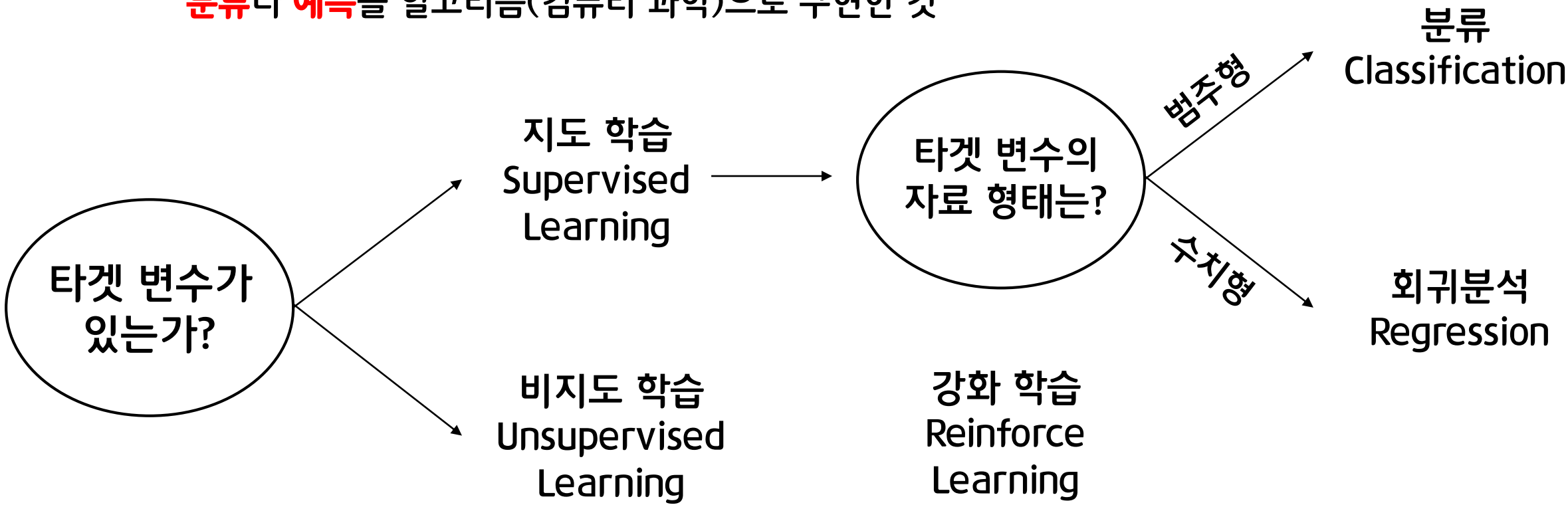
인공지능 : 인간의 지능과 유사한 특성을 가진 복잡한 컴퓨터 /
기계로부터 만들어진 지능

기계학습 : 데이터(빅데이터) 로부터 통계적 분석(통계)을 하고,
이를 바탕으로 컴퓨터를 학습시켜,
분류나 예측을 알고리즘(컴퓨터 과학)으로 구현한 것

딥러닝 : 인공신경망 구조의 hidden layer를 deep하게 쌓은 것으로
비선형 문제를 해결하는데 탁월

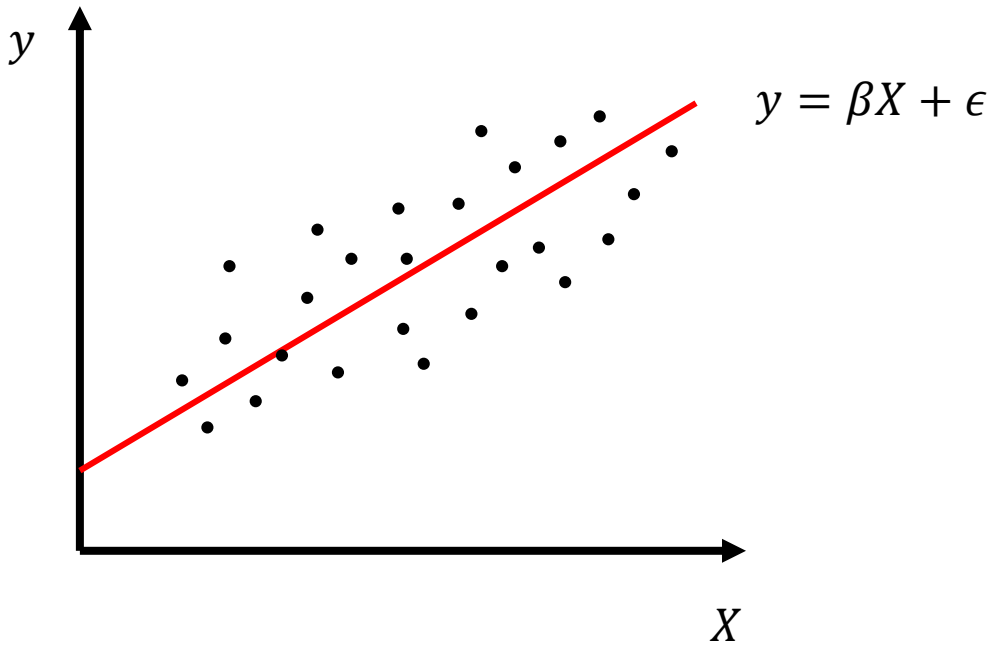
Unit 01 | Machine Learning이란?

기계학습 : **데이터**(빅데이터) 로부터 **통계적 분석**(통계)을 하고,
이를 바탕으로 컴퓨터를 **학습**시켜,
분류나 **예측**을 알고리즘(컴퓨터 과학)으로 구현한 것



1. 예측(Prediction) - 회귀분석

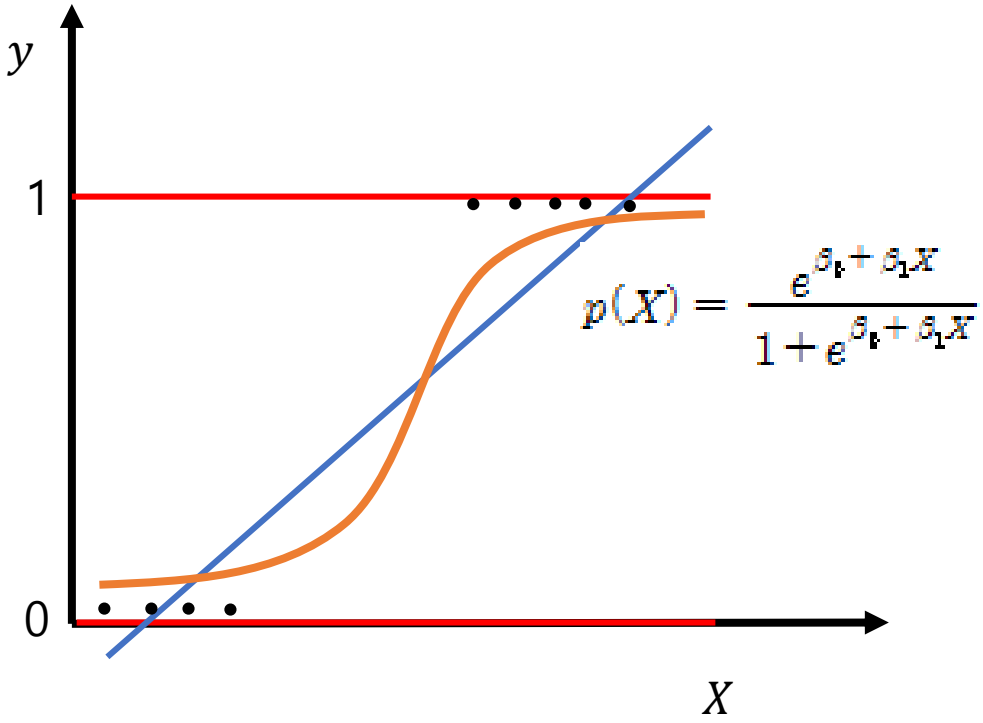
: 연속형 반응변수를 가진 데이터에서 종속변수와 설명변수의 관계를 선형 회귀식으로 나타내고,
새로운 데이터가 들어왔을 때 회귀식으로부터 새로운 반응변수를 예측하는 방법



- 종류 : 단순 회귀, 다중 회귀, 비선형 회귀
- 회귀식을 찾는다.
- 최적의 모수 (β)를 추정한다. → 최소 제곱법
- 변수 선택 : 변수선택기준(AIC , BIC 등)에 부합하는 변수 선택
- 가정 : 선형성, 비상관성, 정상성, 등분산성, 독립성
- 고려사항 : outlier / high leverage / 다중공선성

2. 분류(Classification) – 로지스틱 회귀

: Binary 범주형 변수를 반응변수로 가진 데이터에 대해 설명 변수와 종속 변수 간의 회귀관계를 파악하고, 새로운 데이터가 들어왔을 때, 확률값을 통해 반응변수(binary)를 예측하는 방법



- 회귀분석을 위한 binary 변수의 연속화 :

  Odds → logit

- 최적의 모수 (β)를 추정한다.

  → MLE(Maximum likelihood Estimation)

- 예측 : 추정된 회귀식에 새로운 설명변수 대입 후 0.5를 기준으로 Group 분류

2. 분류(Classification) – 나이브 베이즈

: 특성들 사이의 독립을 가정하는 베이즈 정리를 적용한 확률 분류기

복습 **베이즈 정리**

사전확률과 Likelihood를 이용해 사후확률을 구하는 방법

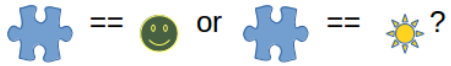
$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{P(B)} = \frac{P(B|A_k)P(A_k)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}$$

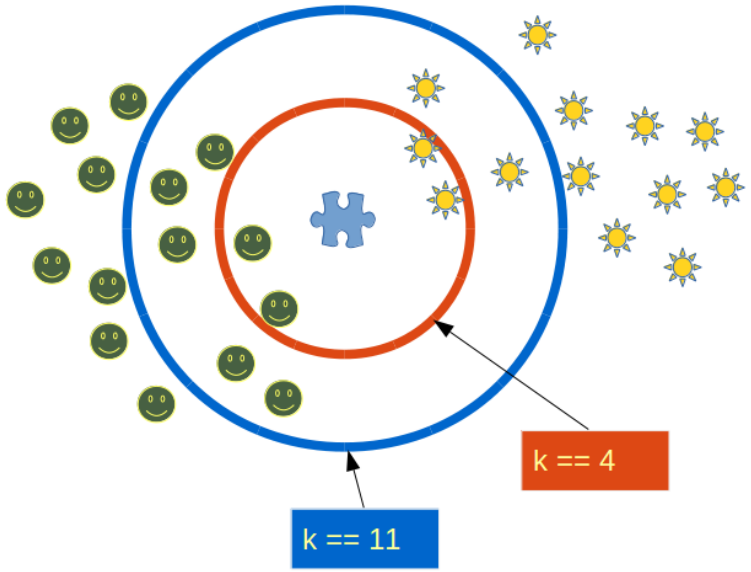
$$Posterior = \frac{Prior * Likelihood}{Evidence} \qquad 사후확률 = \frac{사전확률 * 가능도}{관찰값}$$

- 가정 : feature의 조건부 독립
  (feature의 개수가 증가할수록 likelihood 계산량을 줄이기 위해서 독립 가정)
- 문제점&해결 : train set 에 존재하지 않는 featur가 등장하면 '사후확률=0'이므로 분류 불가능 → Laplace Smoothing

2. 분류(Classification) – KNN(K-Nearest Neighbor)

: 주어진 train set를 좌표공간에 배치한 후,
새로운 데이터데 대해 이 데이터와 가장 가까운 k개의 train 데이터의 성질을 이용해 group을 분류

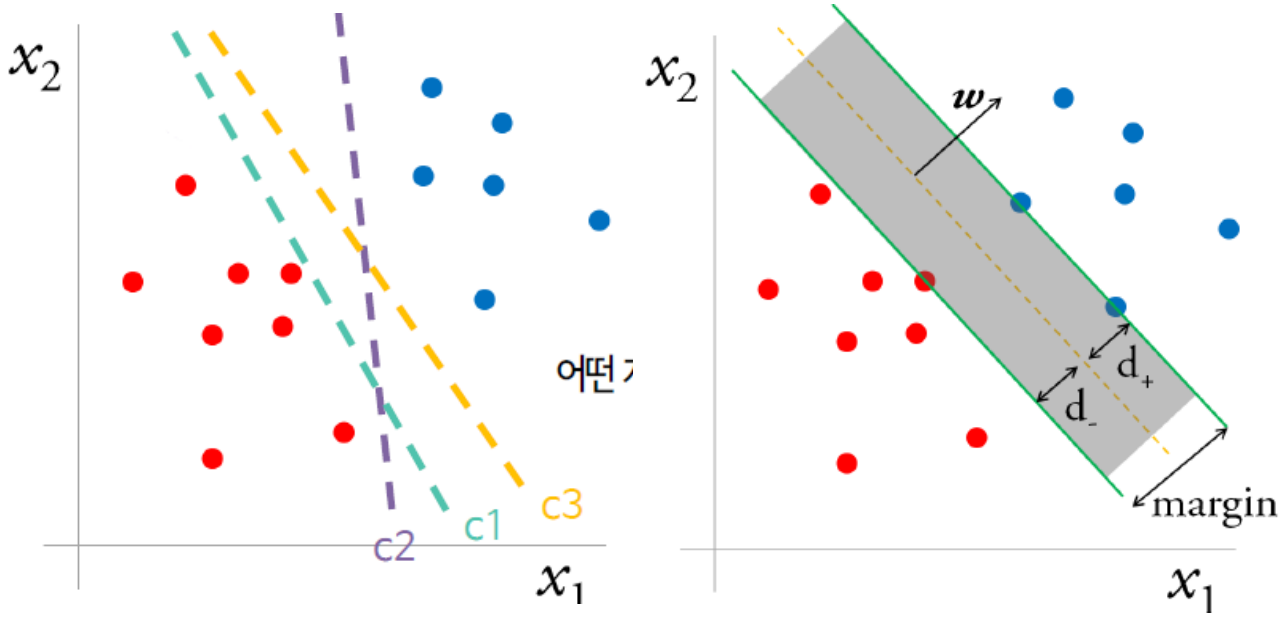




- 거리 측정 방법 : 유클리디안, 코사인 유사도
- K 결정 : 여러가지 k를 적합해본 후 적절한 k
- Group 결정 : k개의 가까운 관측치에 대해 가장 많이 나온 group 혹은 가중치가 큰 group으로 분류

2. 분류(Classification) – SVM(Support Vector Machine)

: Margin(여백)을 최대화하는 결정경계를 찾아 이를 기준으로 group을 분류하는 방법



- SVM 종류 :

  선형 VS 비선형 SVM(결정경계 형태 기준)

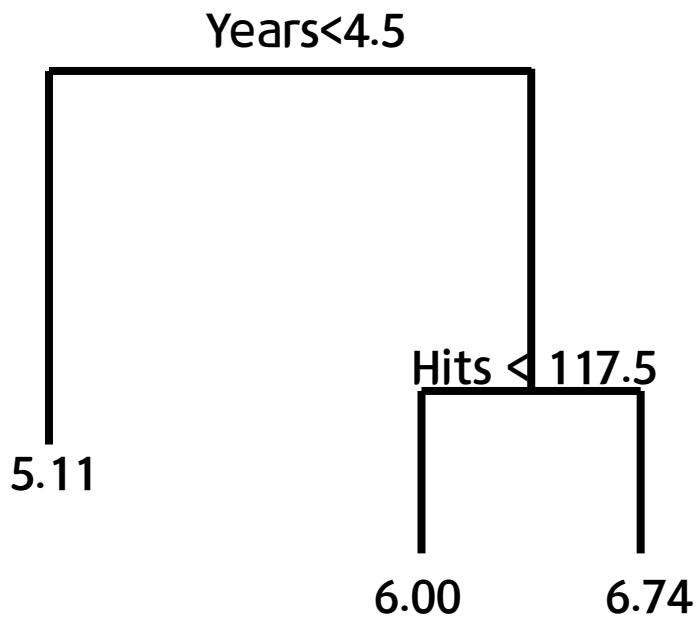
  Hard & Soft margin (margin 안의 데이터 허용)

- 분류 : 결정경계가 정해지면 새로운 데이터에

  그 결정경계를 기준으로 group 분류

3. 예측 & 분류 – 의사결정 나무 (Decision Tree): 회귀나무

: 의사결정 나무 규칙을 통해 나무구조로 도표화하여 데이터를 예측하는 방법



- Split Rule: feature 결정 기준

  모든 가능한 변수들의 절단점을 고려해 SSE가 가장 작게 되는 공간으로 나눈다.

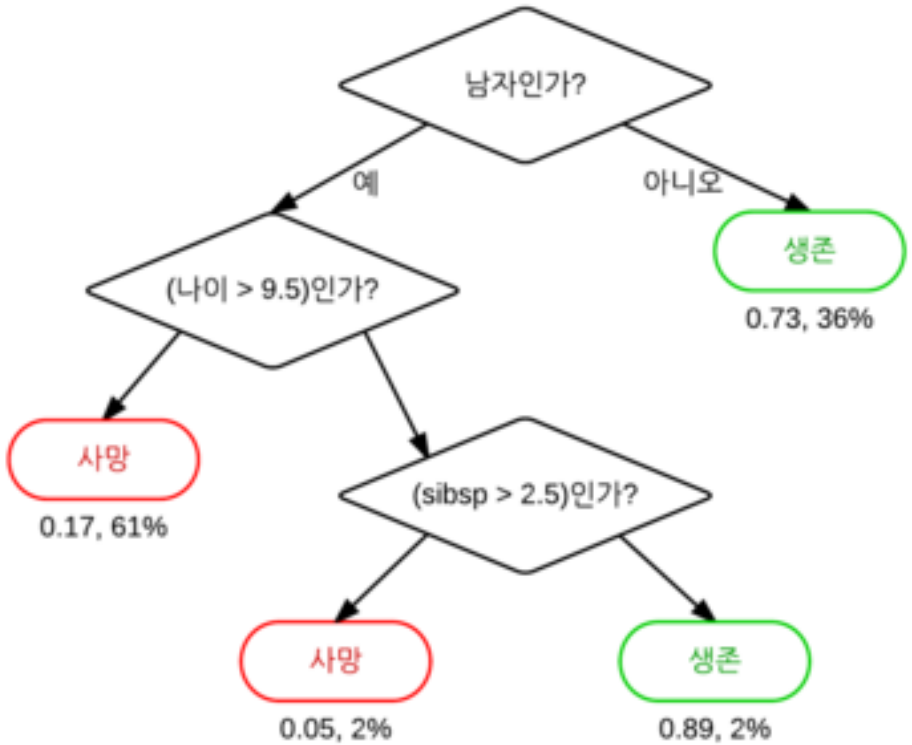
- Stopping Rule: 노드 정지 기준

  split시 sse 감소 여부 / 끝 마디의 sample 개수 / 목표 깊이 도달 여부 등

- 예측 : 새로운 데이터를 넣었을 때 최종 도착하는 끝마디 group 의 평균을 예측 값으로 사용

3. 예측 & 분류 – 의사결정 나무 (Decision Tree): classification 나무

: 의사결정 나무 규칙을 통해 나무구조로 도표화하여 데이터를 분류하는 방법



- Split Rule: feature 결정 기준

  지니계수 / 엔트로피 계수 / 불순도

- Stopping Rule: 노드 정지 기준

  split시 불순도 감소 여부 / 끝 마디의 sample 개수 /

  목표 깊이 도달 여부 등

3. 예측 & 분류 – Ensemble method

: 학습 알고리즘들을 따로 쓰는 경우에 비해 더 좋은 예측 성능을 얻기 위해
다수의 학습 알고리즘을 사용하는 방법

1. 배깅(bagging; bootstrap aggregating)

: 데이터에서 여러 개의 복원추출 샘플을 뽑아 각 샘플에 대한 모델들을 학습

2. 부스팅(boosting)

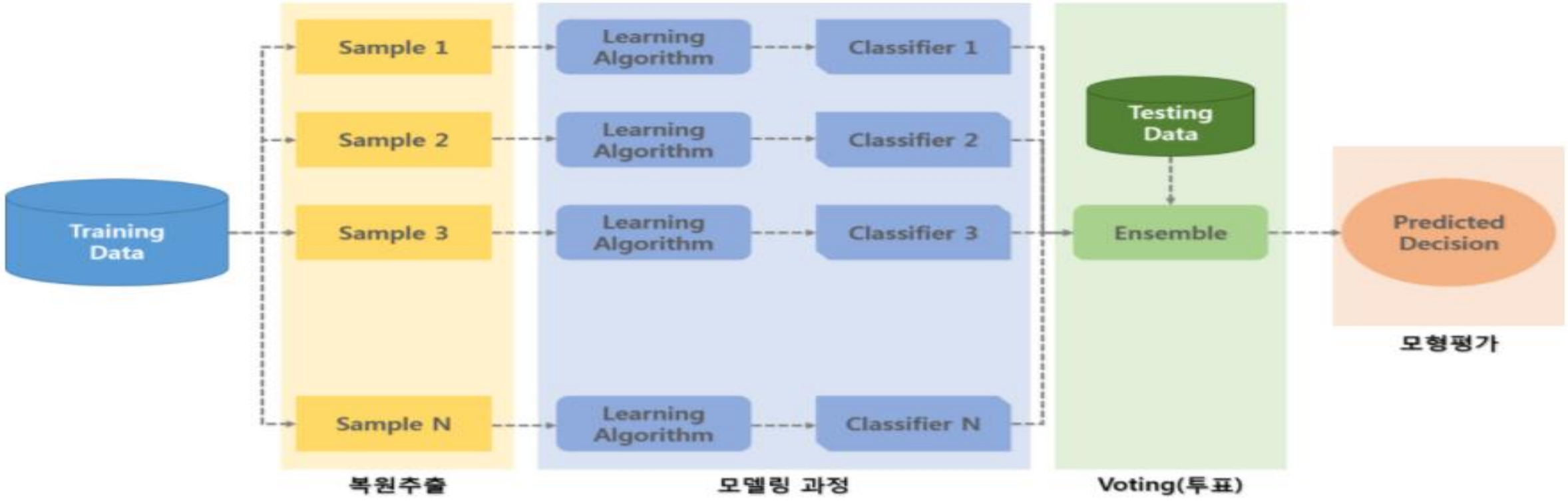
: 잘못 분류된 관측치에 집중하여 새로운 모델을 학습시키는 것을 반복

3. 랜덤포레스트(Random forest)

: 데이터와 속성의 임의 부분 집합을 사용하여 여러 의사결정나무들을 학습

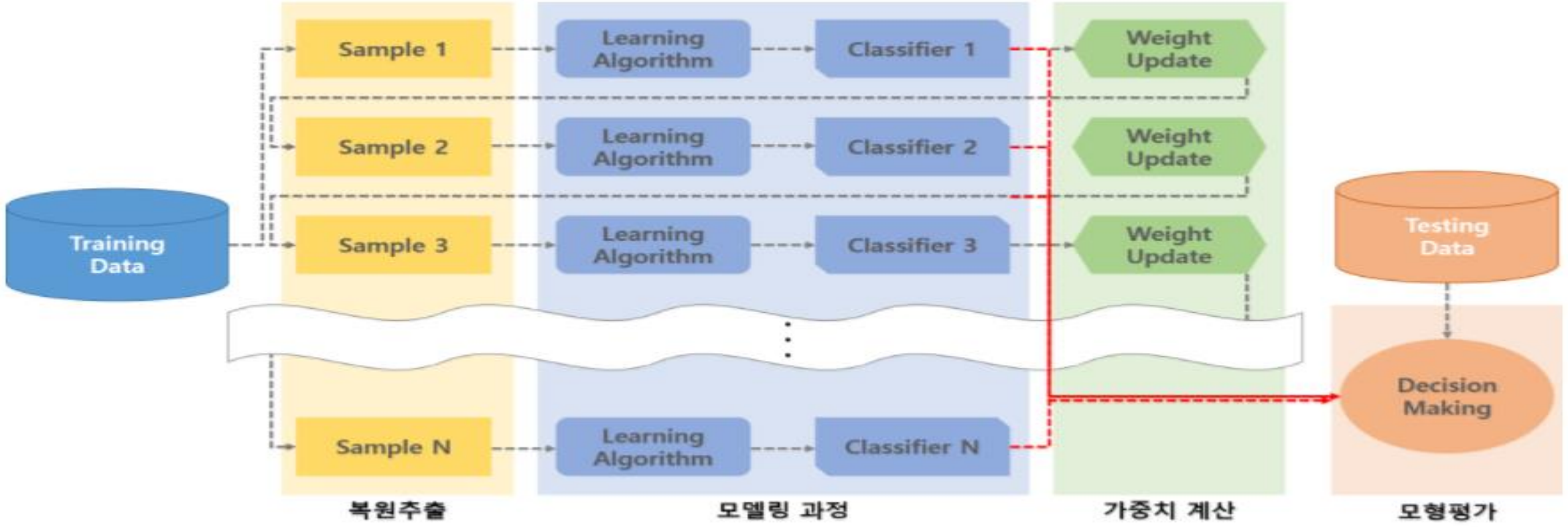
3. Ensemble method – 배깅

: 데이터에서 여러 개의 복원추출 샘플을 뽑아 각 샘플에 대한 모델을 학습하는 방법



3. Ensemble method – 부스팅

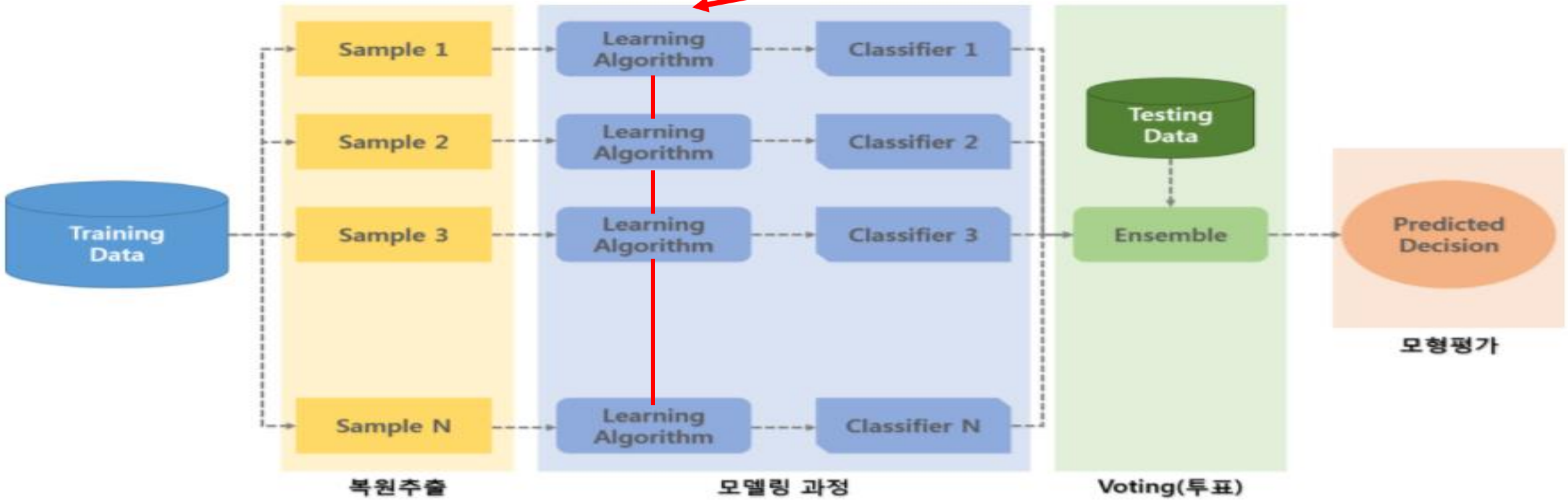
: 잘못 분류된 관측치에 집중하여 새로운 모델을 학습시키는 것을 반복해서 학습하는 방법



3. Ensemble method – 랜덤포레스트

: 데이터와 속성의 임의 부분 집합을 사용하여 여러 의사결정나무들을 학습

Feature 도 Random하게 sampling



1. 차원축소(Dimension Reduction) – PCA(Principal Components Analysis)

: 변수들간의 Auto-correlation을 줄이거나, 변수의 양이 많아 축소하기 위해 사용되는 방법

X_1 X_2 X_3 X_4

$$P_1 = -0.074\,X_1 - 0.303\,X_2 + 0.9501\,X_3$$
$$P_2 = 0.8193\,X_1 + 0.5247\,X_2 + 0.2312\,X_3$$

-P개의 변수를 가진 데이터에 대해 p개의 linear combination으로 표현되는 p개의 주성분을 생성한 후, 이 중 데이터의 70~90% 정도를 설명할 수 있는 k개의 주성분을 p개의 변수 대신 사용한다.

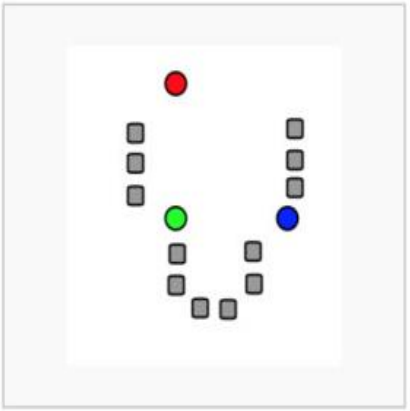
-계수 추정 : Variance를 최대화하는 계수 (eigen vector)

-k 결정 : 전체 Variance 의 70~90%를 설명하는 최소 k / screen plot 의 elbow point 직전의 k

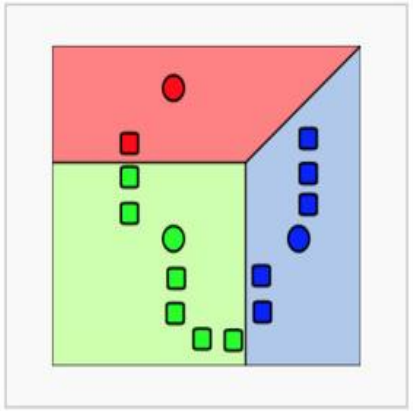
- 주성분 생성 이후 : 주성분 회귀분석

2. 군집(Clustering) – K-means

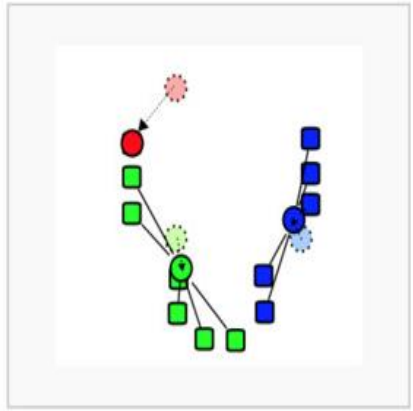
: 초기 k 개의 중심점(centroid)을 정한 후, 중심점과 모든 데이터 간의 거리를 비교해 가까운 중심점의 군집으로 분류하는 과정을 반복해 군집 내 거리를 최소화하는 k개의 군집으로 데이터를 분류하는 방법



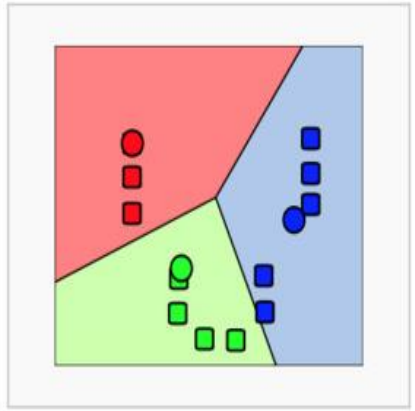
1. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).



2. *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.



3. The centroid of each of the *k* clusters becomes the new mean.



4. Steps 2 and 3 are repeated until convergence has been reached.

-초기 centroid 결정:
여러 번의 초기 중심점 random설정 후 최적의 초기 중심점 결정

-k 결정 : 주어진 k 이용 /
데이터 수의 제곱근 이용 /
elbow point

들어주셔서 감사합니다.