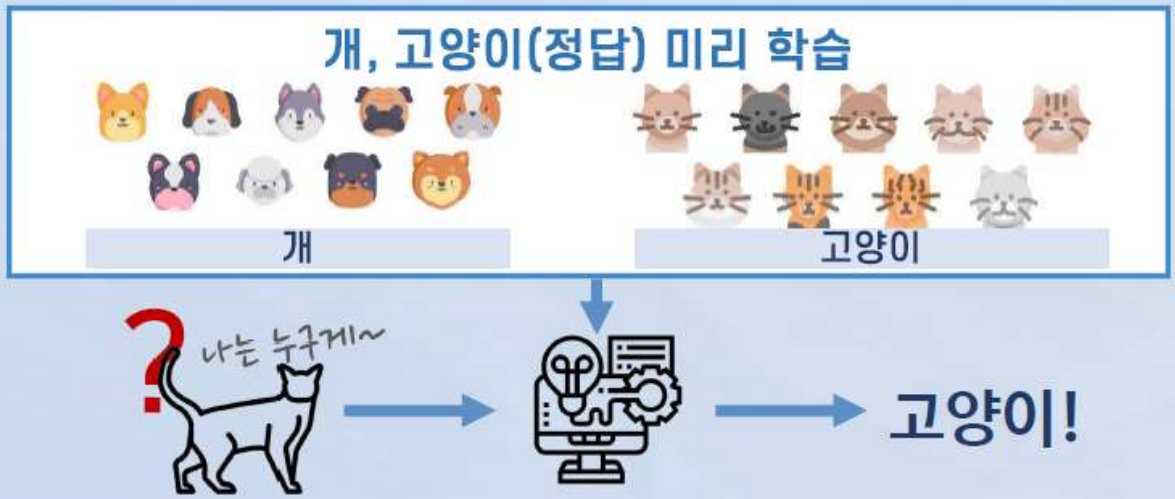
머신러닝

01 지도학습과 비지도학습

지도학습

(Supervised Learning)



장점 손쉽게 모델의 성능을 평가할 수 있음

단점 레이블이 없는 데이터는 레이블을 달기 위해 많은 시간을 투자해야 하는 단점 존재

대표적 예 분류, 회귀

01 지도학습과 비지도학습

비지도학습

(Unsupervised Learning)

데이터의 특성을 스스로 파악

장점 별도로 레이블을 제공할 필요가 없으므로 시간 절약 가능

단점 레이블이 없으므로 모델의 성능을 평가하는 데 다소 어려움이 있음

대표적 예 클러스터링, 차원 축소

02 분류와 회귀

데이터가 입력되었을 때 지도 학습을 통해
미리 학습된 레이블 중 하나 또는 여러 개의 레이블로
예측하는 것

이진 분류

둘 중
하나의 값으로 분류

남 / 여 중 분류

다중 분류

여러 개 중
하나로 분류

0 1 2 3 4
5 6 7 8 9

0~9 중 하나

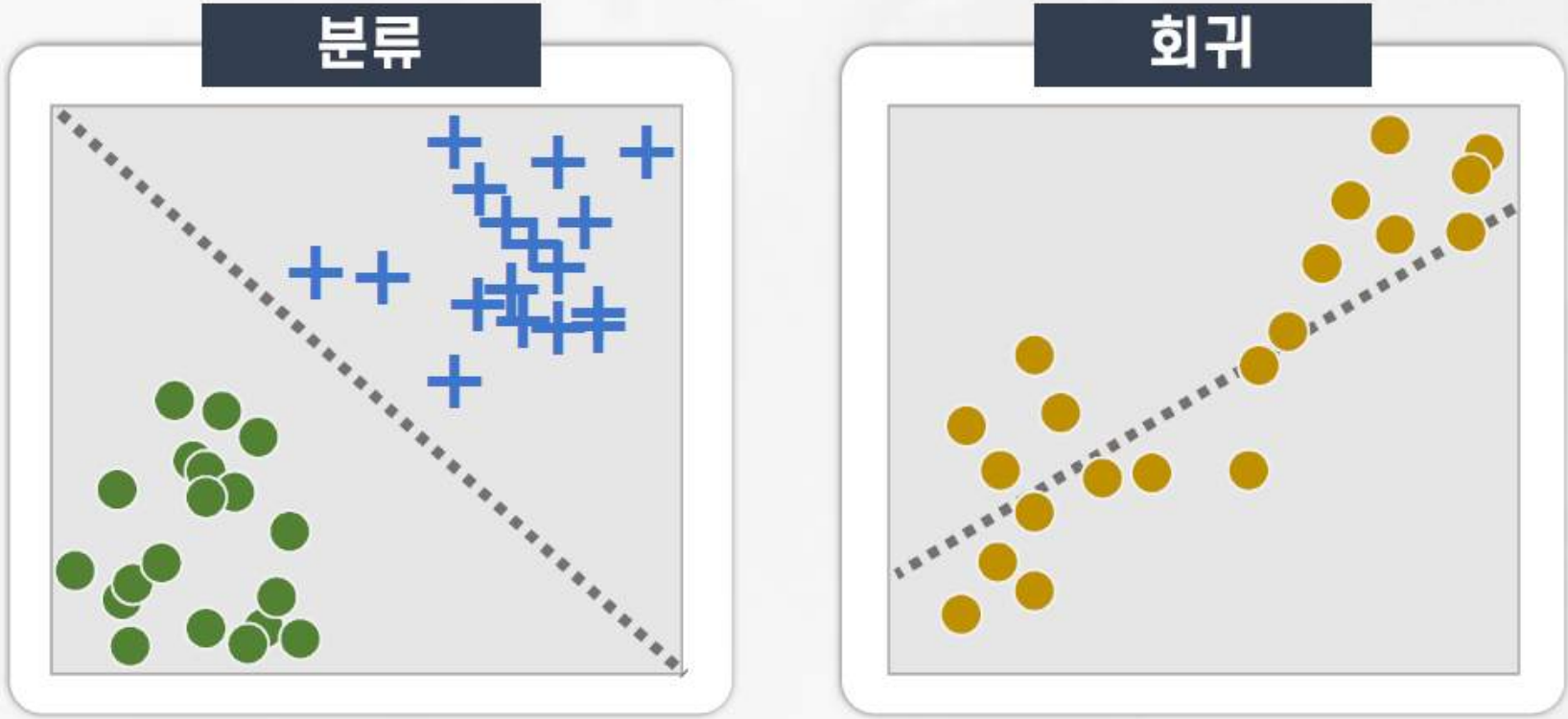
다중 레이블 분류

두 개 이상의
레이블로 분류

뉴스 카테고리 분류

02 분류와 회귀

분류와 회귀의 비교

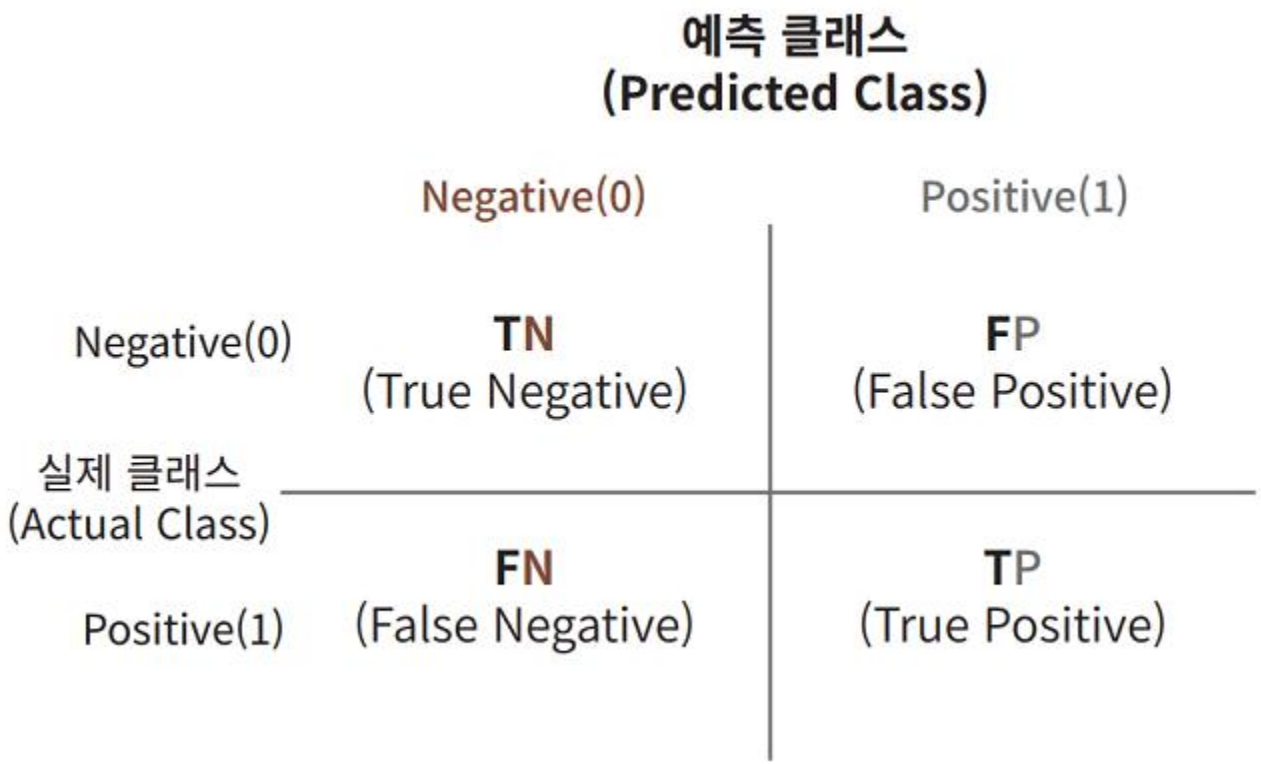


03 과대적합과 과소적합

과대적합(overfitting)

- 학습 데이터에 대한 정확도는 매우 높지만 테스트 데이터 또는 학습 데이터 외의 데이터에는 정확도가 낮게 나오는 것(분산이 높음)

04 혼동 행렬

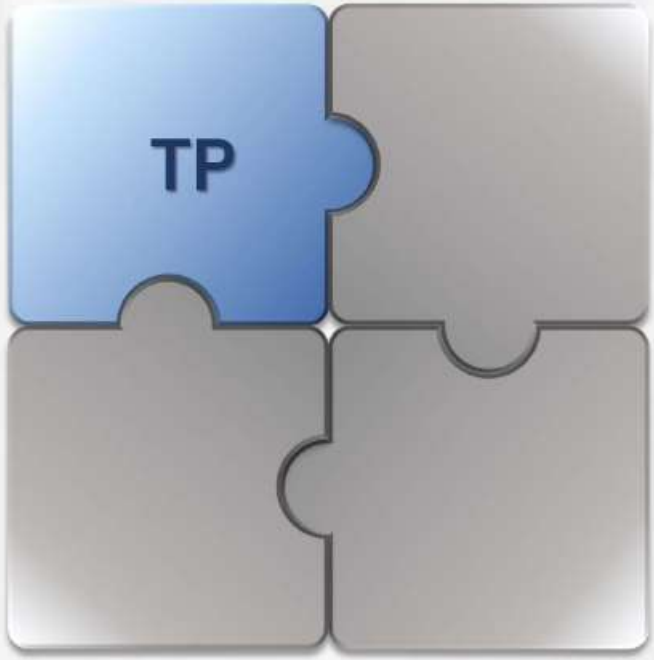

예측 클래스
(Predicted Class)
Negative(0)
Positive(1)
Negative(0)
TN
(True Negative)
FP
(False Positive)
실제 클래스
(Actual Class)
Positive(1)
FN
(False Negative)
TP
(True Positive)

04 혼동 행렬

TP(True Positive)

- 맞는 것을 올바르게 예측한 것
- 암환자를 암환자로 예측

		예측 결과	
		암환자	일반인
실제 정답	암환자	5	2
	일반인	1	2



04 혼동 행렬

FP(False Positive)

- 틀린 것을 맞다고 잘못 예측한 것
- 일반인인데 암환자로 잘못 예측

		예측 결과	
		암환자	일반인
실제 정답	암환자	5	2
	일반인	1	2

05 머신러닝 모델의 성능 평가

정확도(Accuracy)

- 입력된 데이터에 대해 올바르게 예측한 비율
- 혼동 행렬 상에서는 대각선을 전체 셀로 나눈 값에 해당

$$\text{정확도} = \frac{TP + TN}{(TP + FN + FP + TN)}$$

05 머신러닝 모델의 성능 평가

정밀도(Precision)

- 모델의 예측 값이 얼마나 정확하게 예측됐는가를 나타내는 지표
- False를 True라고 판단하면 안되는 경우 중요

$$\text{정밀도} = \frac{TP}{(TP + FP)}$$

05 머신러닝 모델의 성능 평가

재현율(Recall)

- 실제 값 중에서 모델이 검출한 실제 값의 비율을 나타내는 지표
- True를 False로 잘못 판단하면 큰일나는 경우 중요
- 정밀도와 재현율은 trade-off 관계

$$\text{재현율} = \frac{TP}{(TP + FN)}$$

05 머신러닝 모델의 성능 평가

F1 score

- 정밀도와 재현율 두 값을 조합하여 하나의 수치로 나타낸 지표
- 데이터의 레이블이 불균일하게 분포돼 있을 경우,
  정확도는 왜곡된 성능 평가로 이어질 수 있으므로 사용

$$\text{F1 score} = \frac{2*\text{precision}*\text{recall}}{\text{precision} + \text{recall}}$$

05 머신러닝 모델의 성능 평가

K-폴드 교차 검증(K-fold cross validation)

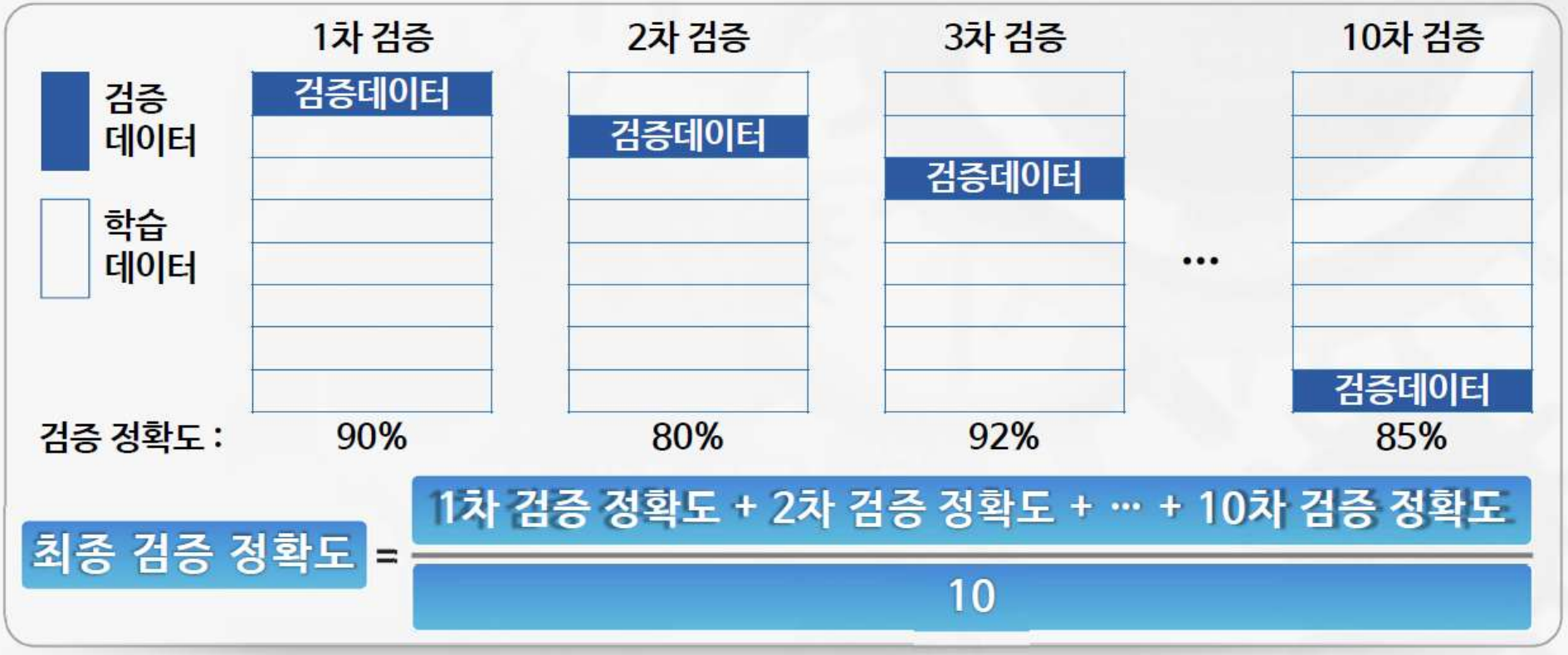
- 학습 데이터의 일정 부분을 검증데이터로 쓰되, n번의 검증 과정을 통해 학습 데이터의 모든 데이터를 한 번씩 검증데이터로 사용하는 방식
- Training Set, Validation Set, Test Set의 개념 구분

장점

- 검증 결과가 일정 데이터에 치우치지 않고 모든 데이터에 대한 결과이므로 신빙성이 높음
- 별도로 검증 데이터를 분리하지 않아도 됨

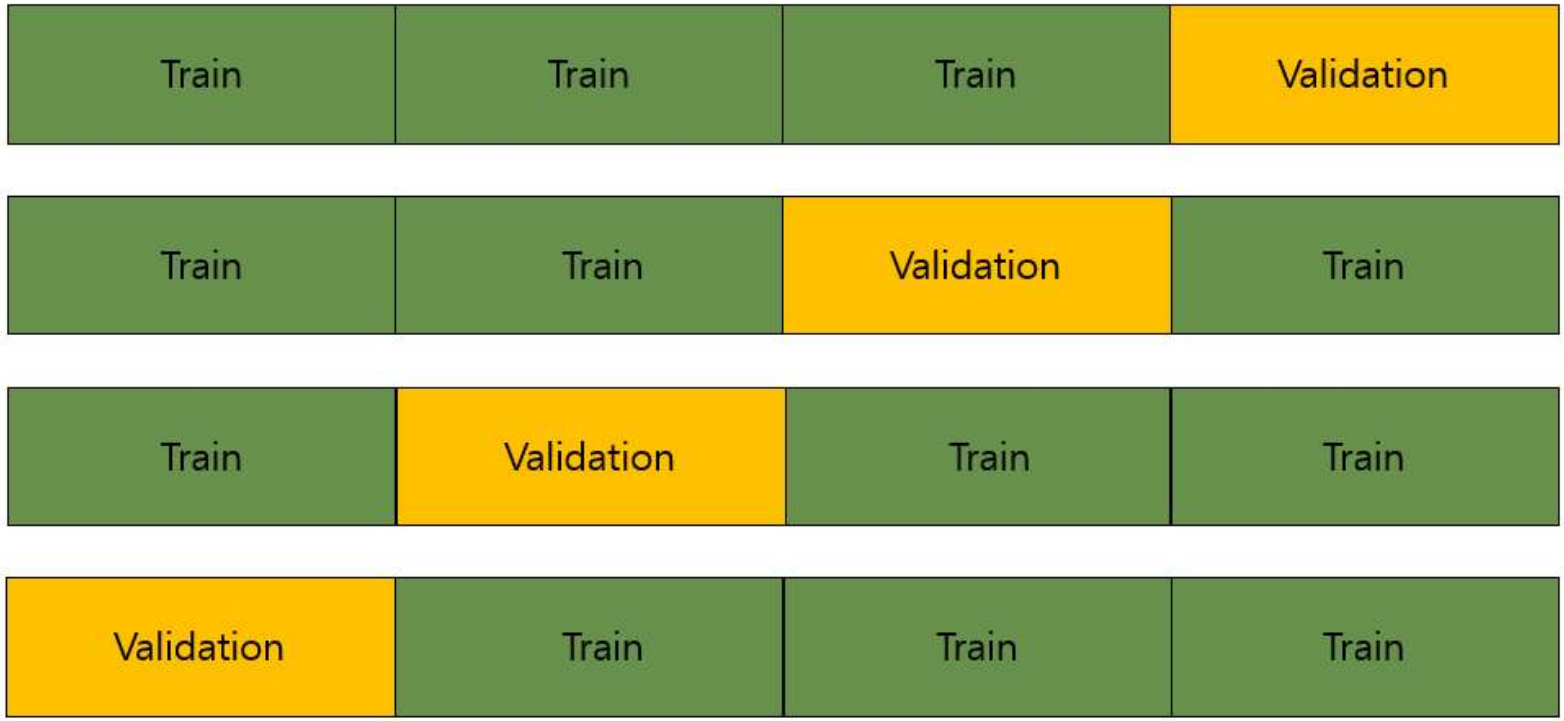
05 머신러닝 모델의 성능 평가

K-폴드 교차 검증(K-fold cross validation)



$$\text{최종 검증 정확도} = \frac{\text{1차 검증 정확도} + \text{2차 검증 정확도} + \cdots + \text{10차 검증 정확도}}{10}$$

K-폴드 교차 검증


Train
Train
Train
Validation
Train
Train
Validation
Train
Train
Validation
Train
Train
Validation
Train
Train

앙상블(Ensemble)
: 결정트리 기반 알고리즘을 결합하여 구현

[앙상블 종류 요약]

•(1) Voting : 서로 다른 알고리즘 가진 분류기를 결합, 사이킷런은 VotingClassifier 클래스를 제공함

<1> 하드보팅(Hard Voting) : 분류기들이 예측한 결과 값을 다수결로 결정

<2> 소프트 보팅(Soft Voting) : 각 분류기들이 예측 값을 확률로 구하면 이를 평균 내어 확률이 가장 높은 값을 결과 값으로 결정

(2) Bagging : 같은 유형의 알고리즘을 결합, 데이터 샘플링시 서로 다르게 가져가면서 학습, RandomForest 가 대표적, Bootstrapping Aggregation 줄임말

(Bootstrapping : 여러개의 데이터 세트를 중첩되게 분리하는 분할 방식

(3) Boosting : 여러개의 분류기가 순차적으로 학습하면서 가중치를 부스팅한다, XGBoost(캐글 대회 상위 석권),LightGBM

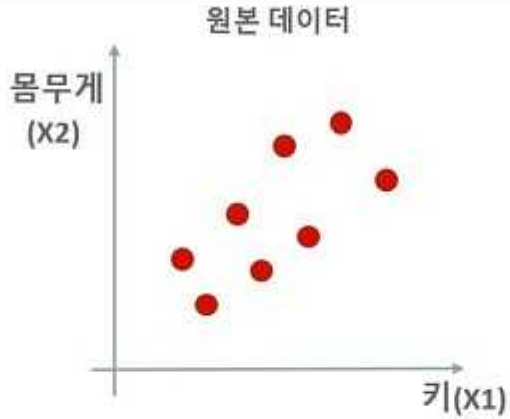
AdaBoost 알고리즘 참고사이트: https://dohk.tistory.com/217

랜덤포레스트(RandomForest)

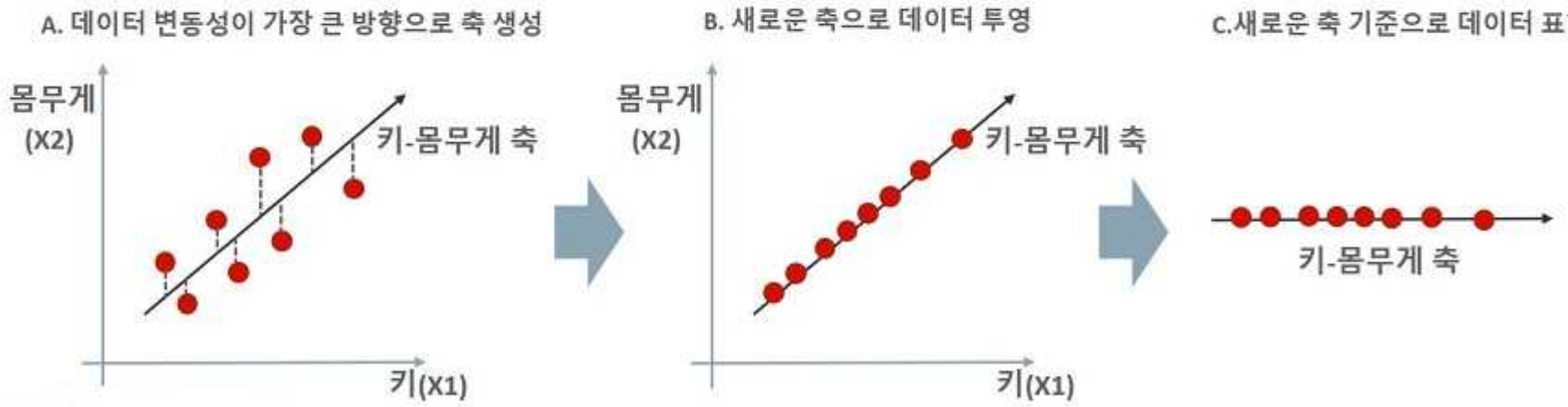
- 의사 결정 트리 기반(Decision Tree) 기반 분류 알고리즘

- 앙상블(Ensemble), 같은 결정트리를 여러개 사용, 비교적 빠른 수행
- 현재의 랜덤 포레스트의 개념은 레오 브레이먼(Leo Breimen)의 논문에서 만들어짐, 이 논문은 랜덤 노드 최적화(Randomized Node Optimization,RNO)와 배깅(bagging)을 결합한 방법과 같은 CART(Classification And Regression Tree)를 사용해 상관관계가 없는 트리들로 포레스트를 구성하는 방법을 제시했다

PCA(Principal Component Analysis)의 이해



PCA는 원본 데이터 변동성이 가장 큰 방향으로 순차적으로 축들을 생성하고, 이렇게 생성된 축으로 데이터를 투영하는 방식입니다.



K-Means Clustering

군집 중심점(Centroid) 기반 클러스터링