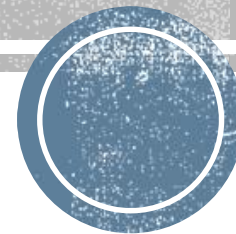
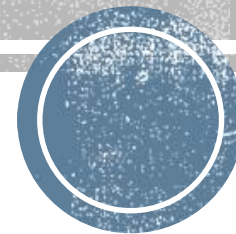


2. 머신 러닝의 기초



2. 머신 러닝의 기초



목차

- 머신 러닝 소개
- 머신 러닝의 종류
- 지도학습이란?
- 비지도학습이란?





머신 러닝 소개



머신 러닝의 정의

- 머신 러닝의 정의

- 명시적으로 프로그래밍하지 않아도 컴퓨터가 데이터를 통해 학습하여 패턴을 인식하고 이를 기반으로 예측이나 결정을 내리는 기술
- 다양한 알고리즘과 모델을 사용하여 데이터를 분석
- 스스로 성능을 개선

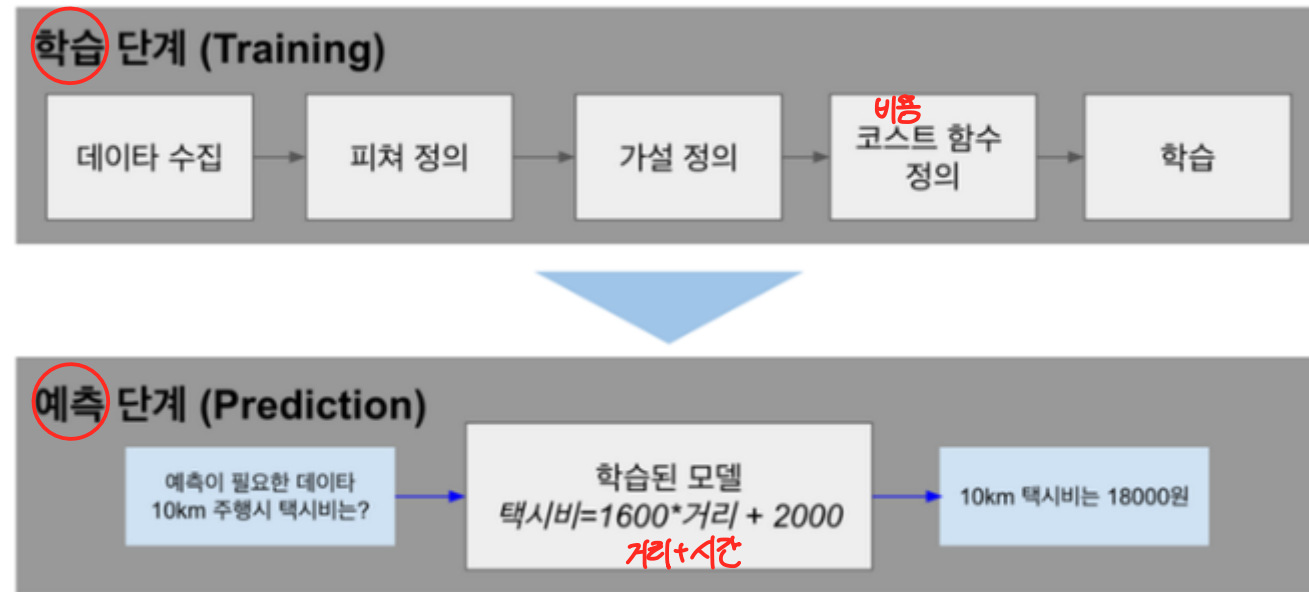
개발자가 세트를 하여 구체적으로
한계별로 코딩을 해주는 것.

CS (sw. coding)
|
AI



구체적인 정의 요소

- 데이터 기반 학습
 - 대량의 데이터를 분석하고 학습
 - 데이터의 패턴과 특징을 파악
- 명시적 프로그램 불필요
 - 명시적인 코딩이 불필요
 - 데이터를 기반으로 규칙과 모델을 학습
- 패턴 인식과 예측
 - 새로운 데이터에 대해 예측
 - 특정 작업을 수행
 - 이미지 인식, 음성 인식, 자연어 처리 등
- 성능 개선
 - 학습과정을 통해 지속적으로 성능 개선
 - 더 많은 데이터를 제공, 알고리즘 조정 등



epoch : 학습횟수



사례 : 스팸 필터링 (by human)

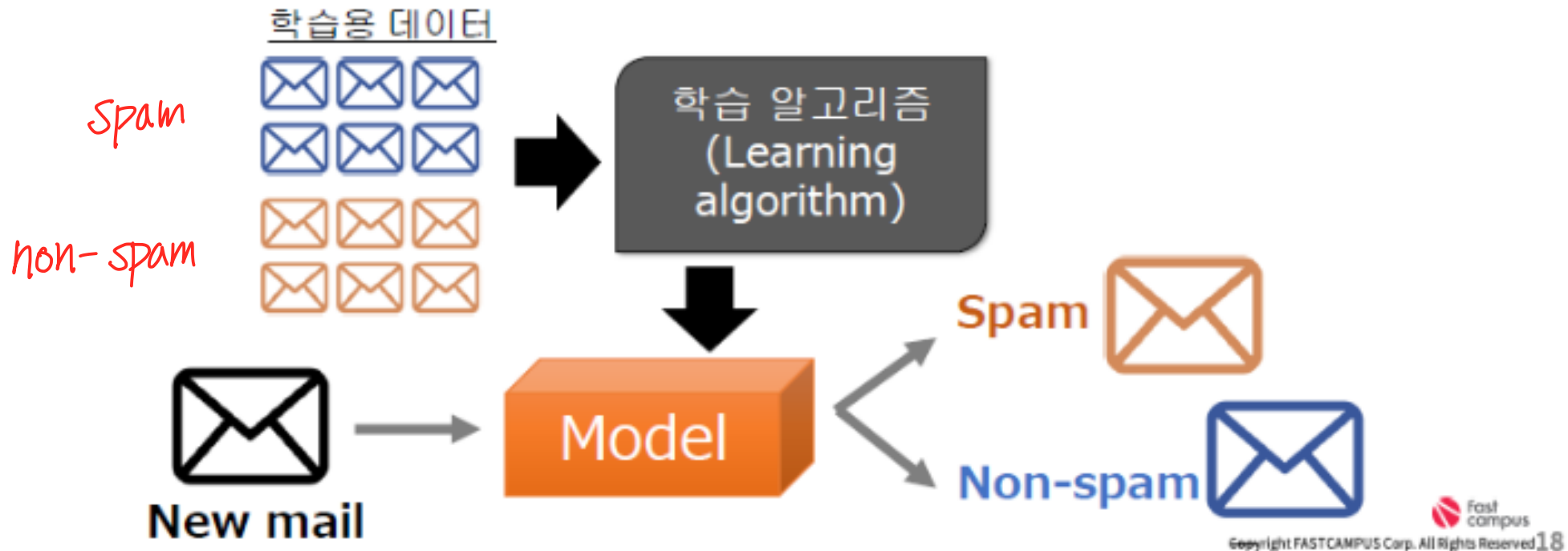


사례 : 스팸 필터링 (by rules)

- 전문가의 지식을 규칙으로 구현
 - 특정 단어의 포함 여부 확인 : 광고, 찜핑, 바둑이, 대출, 카지노 등등
 - 특수문자 개수 : 특수문자가 일정 개수 이상인 경우 스팸으로 처리
- 문제점
 - 전문가들의 지식을 규칙으로 완벽히 구현할 수 있는가?
 - 새로운 형태의 스팸은 어떻게 찾는가?



사례 : 스팸 필터링 (by machine learning)



머신러닝의 현재

인공지능기반 헬스케어 서비스



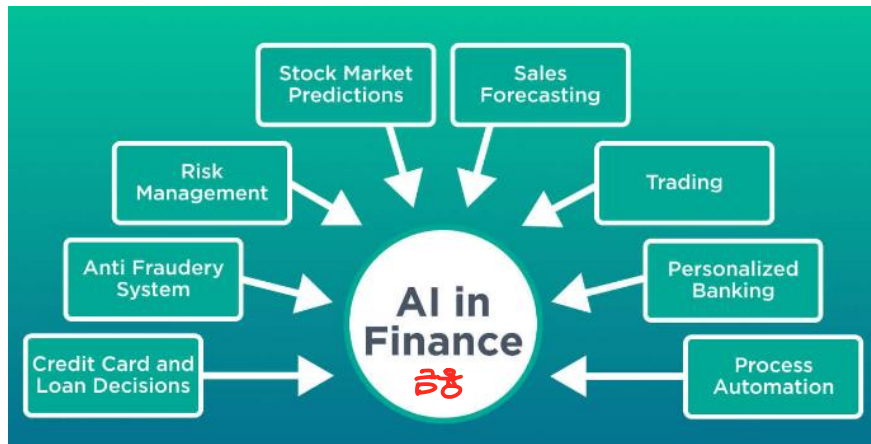
AI ① 선생님

② 변화

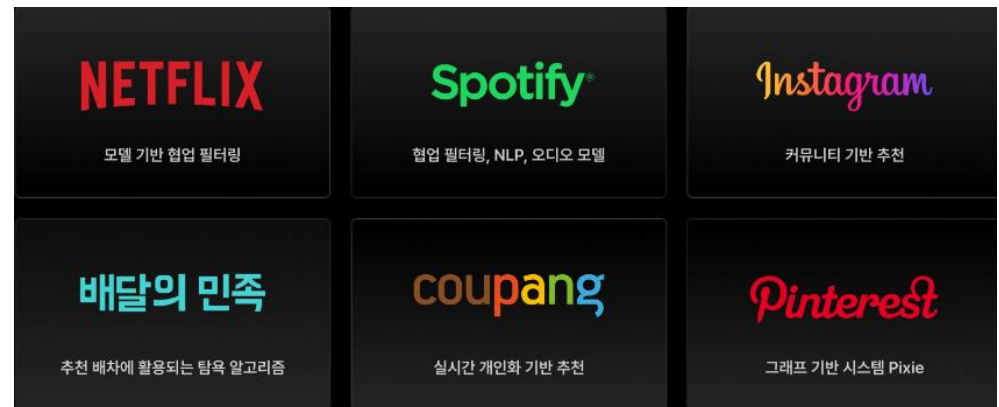
③ 의사

비용이 바뀐다

고객 트렌드 변화

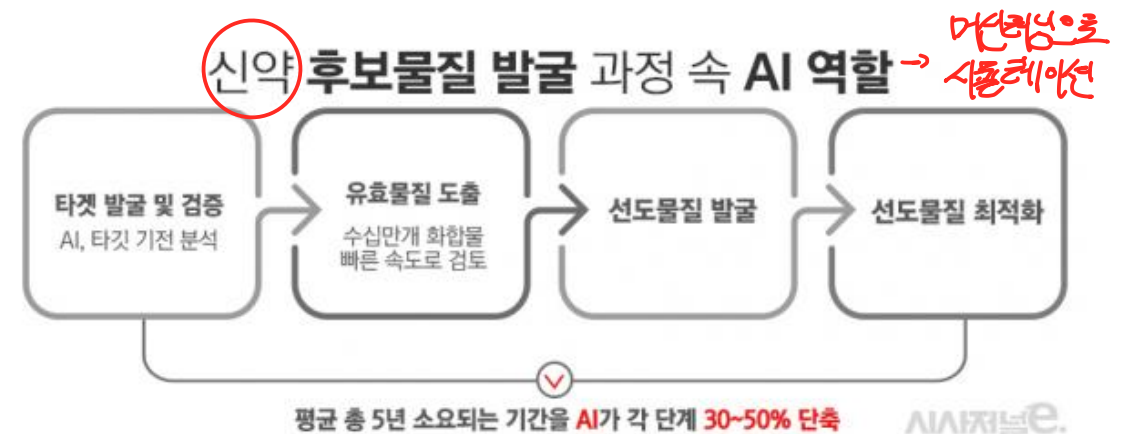
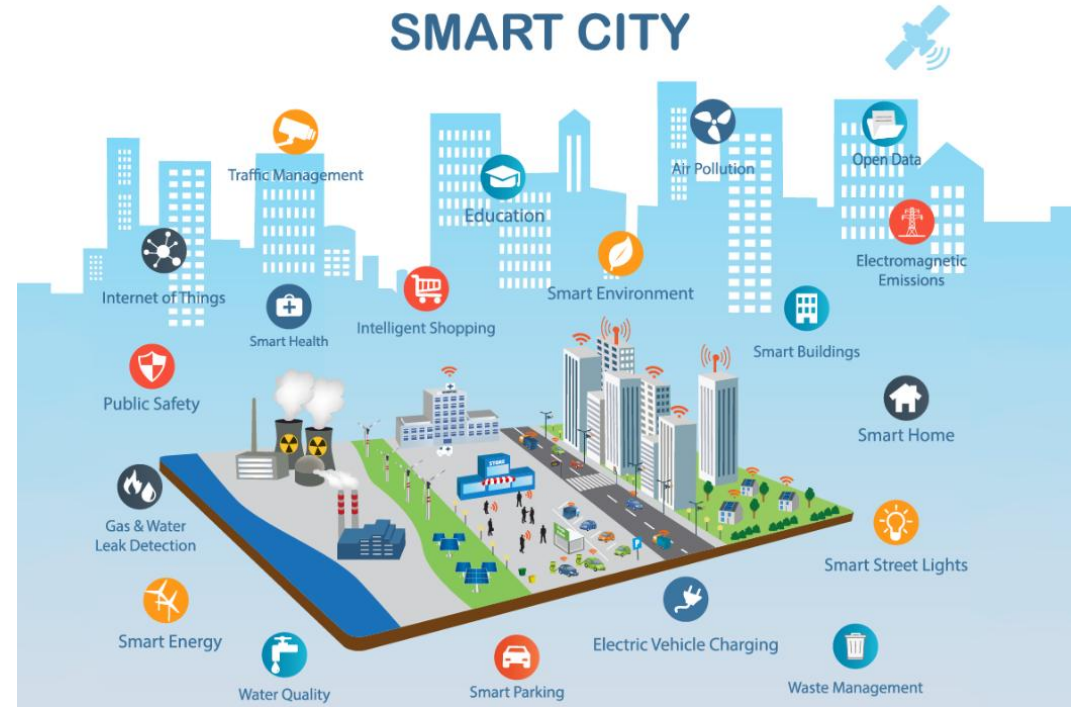
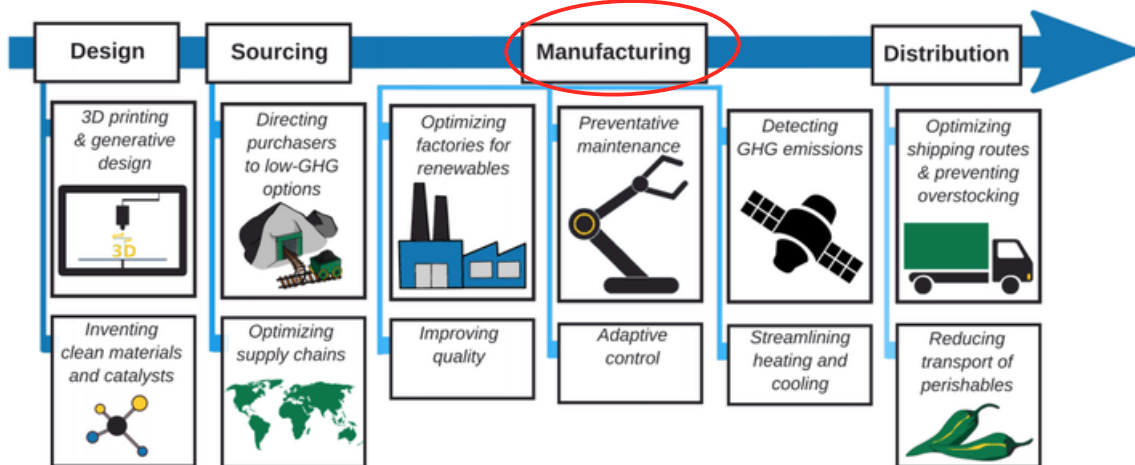


추천 시스템



머신러닝의 미래

고위



머신 러닝의 기본 개념

■ 데이터

- 머신 러닝 모델이 학습하고 예측을 수행하는데 필요한 입력
- 좋은 품질의 데이터가 모델을 정확하게 학습
- 정형데이터(데이터베이스) 와 비정형데이터(텍스트, 이미지 등)

■ 모델

SQL(Table)

- 데이터를 기반으로 학습하여 예측을 수행하는 함수 또는 시스템
- 다양한 알고리즘을 사용하여 데이터를 분석하고 유의미한 패턴을 학습
- 지도학습, 비지도 학습, 강화 학습 등

■ 알고리즘

- 모델이 데이터를 학습하는 과정을 정의하는 절차나 방법
- 데이터와 모델 사이의 관계를 학습하고 새로운 데이터에 대한 예측을 수행
- 선형 알고리즘 : 선형 회귀, 로지스틱 회귀 등
- 비선형 알고리즘 : 결정 트리, 랜덤 포레스트, 서포트 벡터 머신 등

통계

- 확률적 알고리즘 : 나이브 베이즈, 히든 마르코프 모델 등
- 신경망 알고리즘 : 인공 신경망, 합성곱 신경망, 순환 신경망 등



머신 러닝의 기본 개념

■ 학습 (Training)

- 목적 : 모델이 데이터의 **패턴**을 학습하여 예측 능력을 갖추는 것
- 방법 : 데이터 -> 예측된 출력과 실제 출력 간의 차이를 최소화하는 방향으로 **파라미터 조정**
- 과정 : 데이터를 학습/검증 데이터로 분리 -> 학습 데이터로 훈련 -> 검증 데이터로 성능 평가

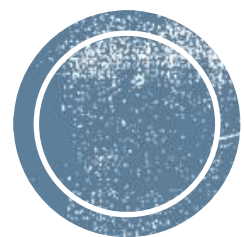
■ 예측 (Prediction)

- 목적 : 새로운 데이터에 대한 결과를 추정
- 방법 : 최적화된 모델 파라미터를 사용하여 입력 데이터를 처리
- 과정 : 새로운 입력 데이터 -> 예측값 출력

■ 평가 (Evaluation)

- 목적 : 모델의 정확도, 정밀도, 재현율 등 성능 지표를 측정
- 방법 : 테스트 데이터 세트로 모델의 예측 결과와 실제 결과를 비교
- 과정 : 혼동 행렬(confusion matrix) 등을 사용하여 성능 지표를 계산하고 모델의 강점과 약점을 파악





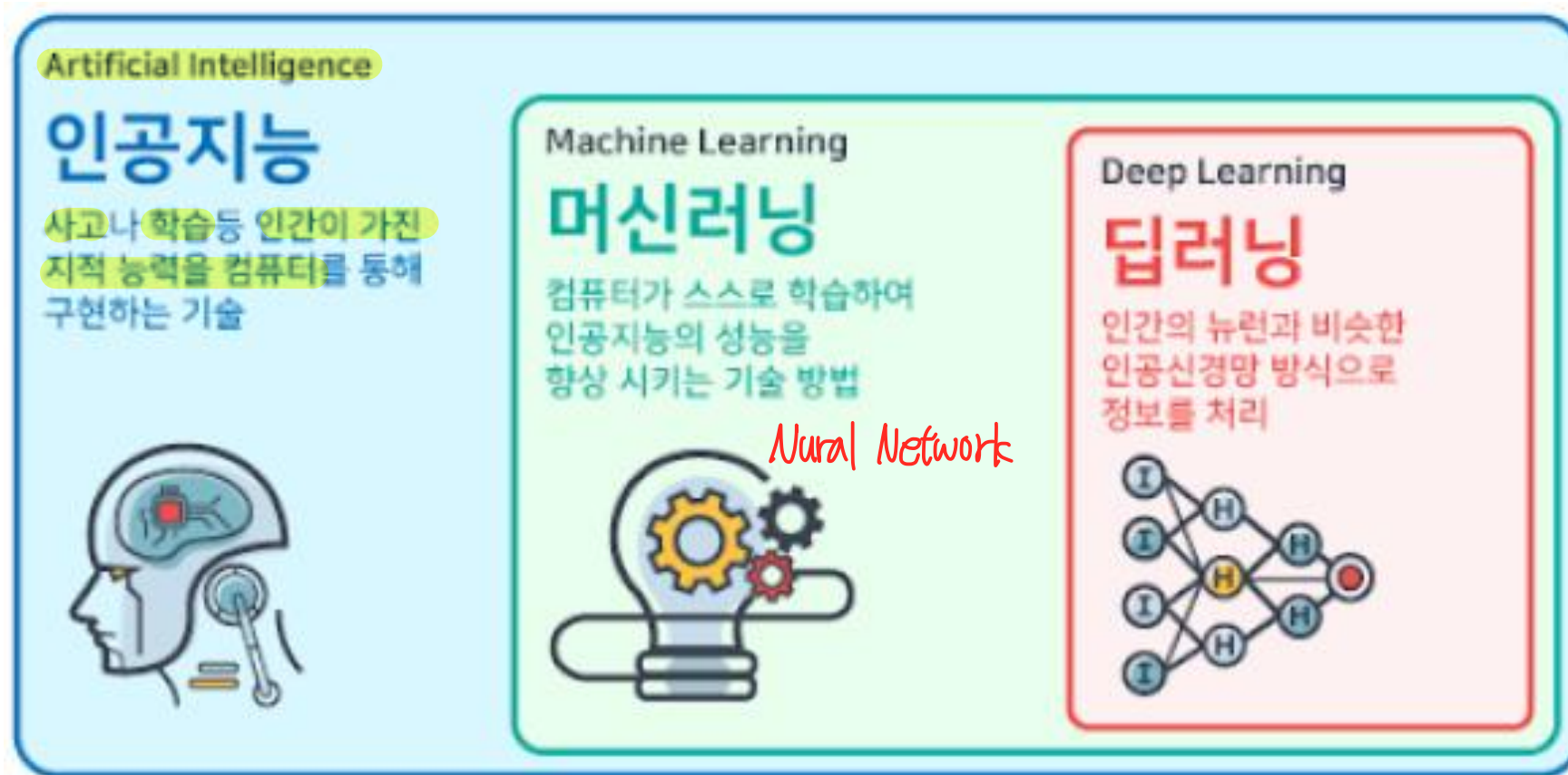
머신 러닝의 종류



인공지능 vs 머신러닝 vs 딥러닝

AGI

채널
→ 결과
모방



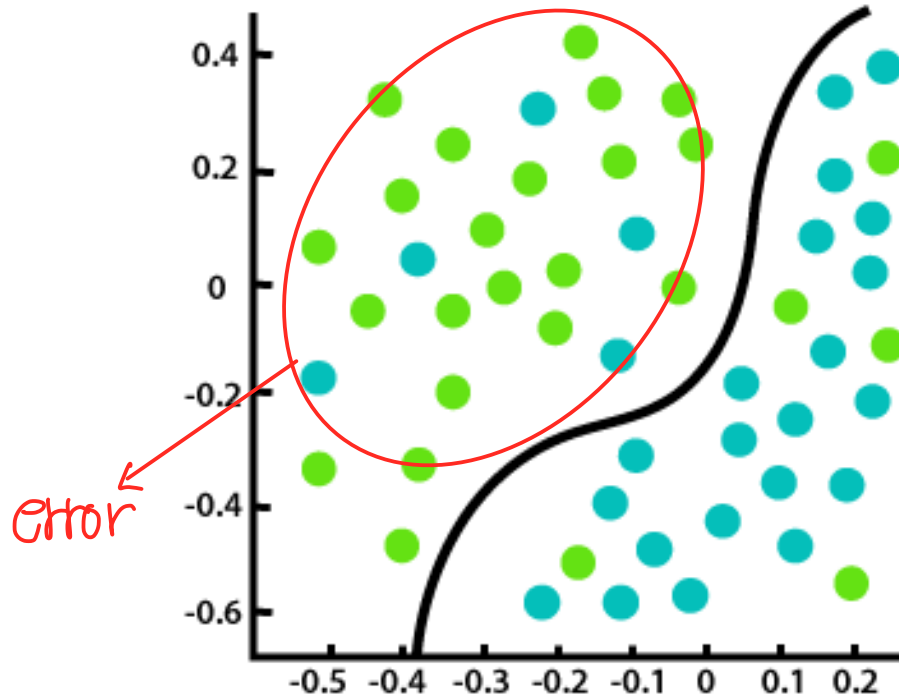
머신 러닝의 종류



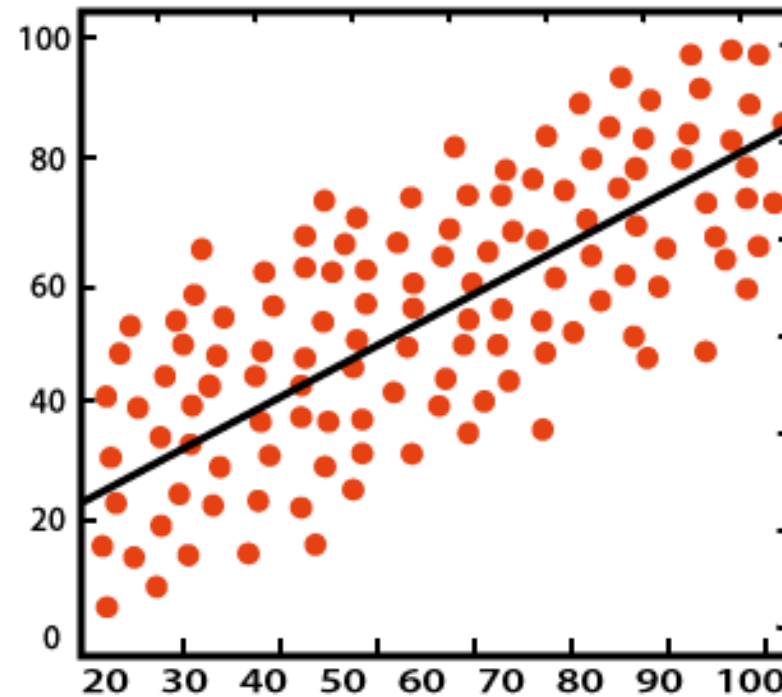
지도 학습 : Regression vs. Classification

회귀

분류



Classification



Regression

$$y = ax + b$$

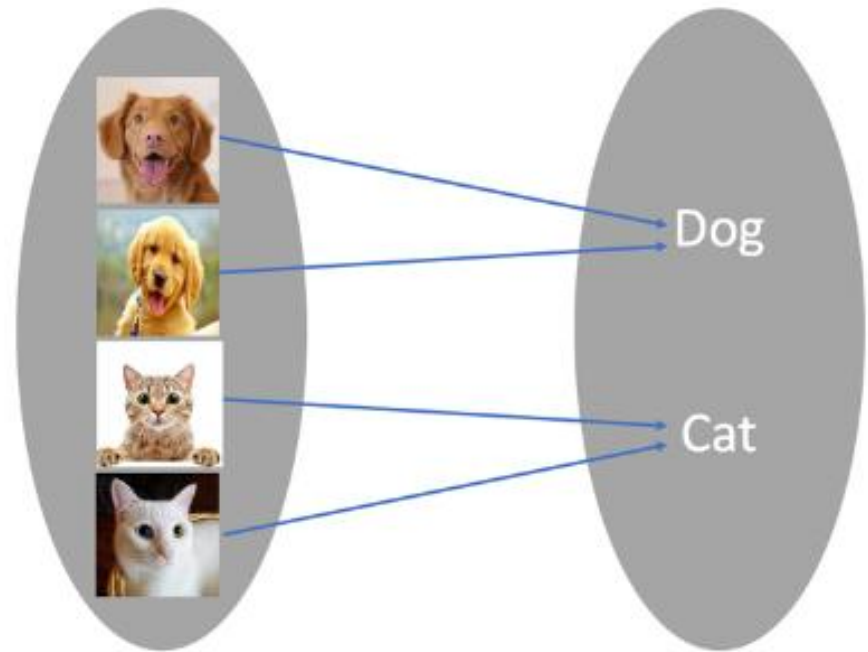
↓ ↓
가중치 y절편



지도 학습 사례 중첩이 있는 레이어



스팸 메일 필터링 { spam
non-spam

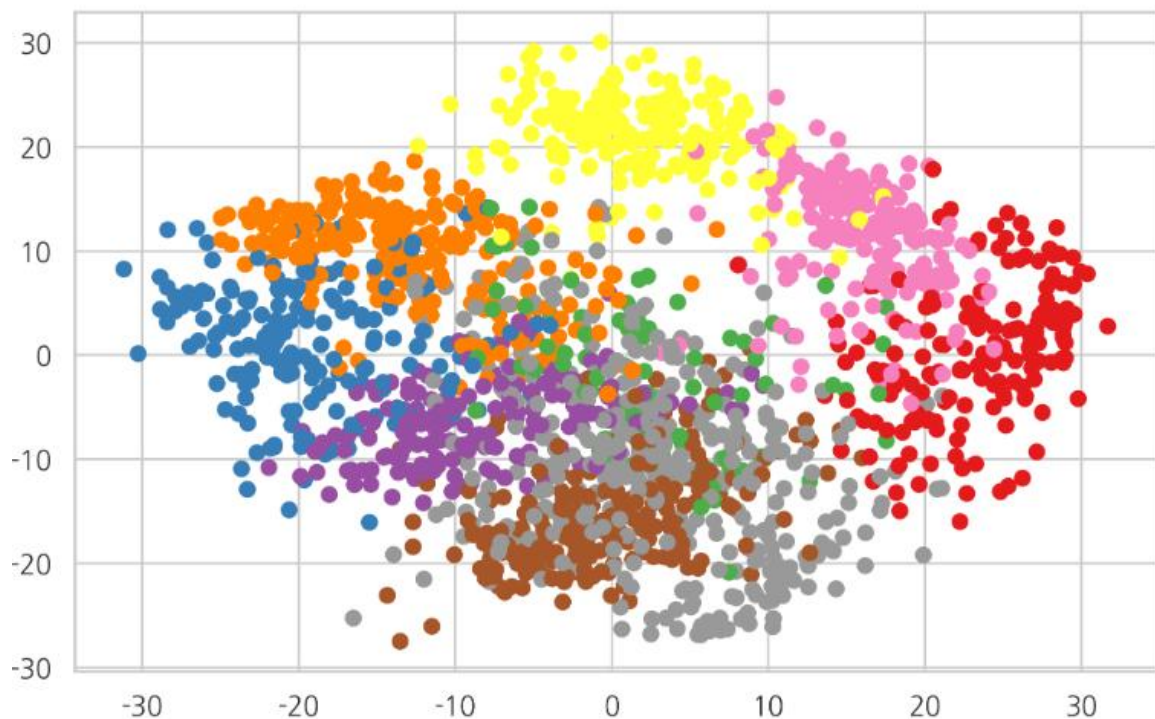


이미지 분류

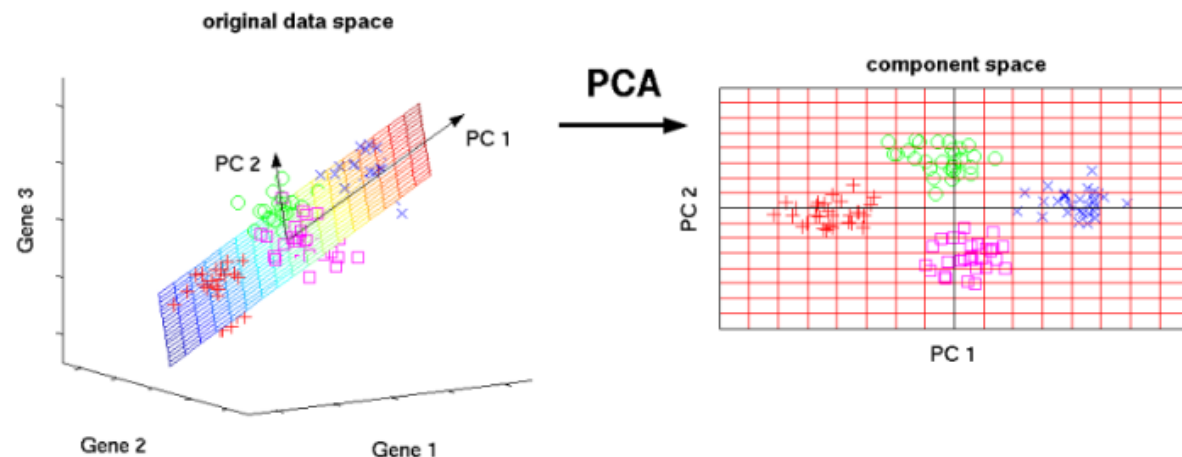


비지도 학습 : 군집화, 차원 축소

비슷한 특징끼리 모으는 것



Clustering



PCA

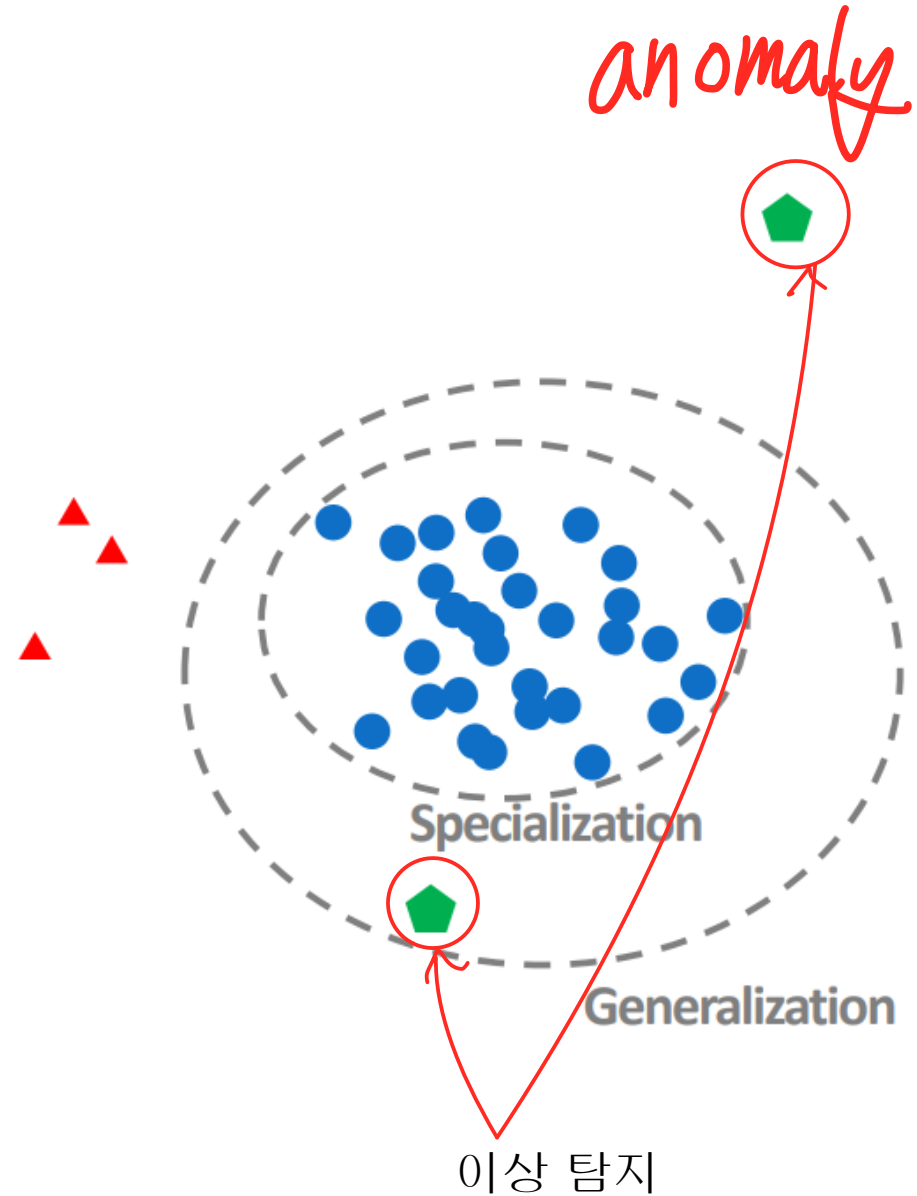
(Principal Component Analysis)



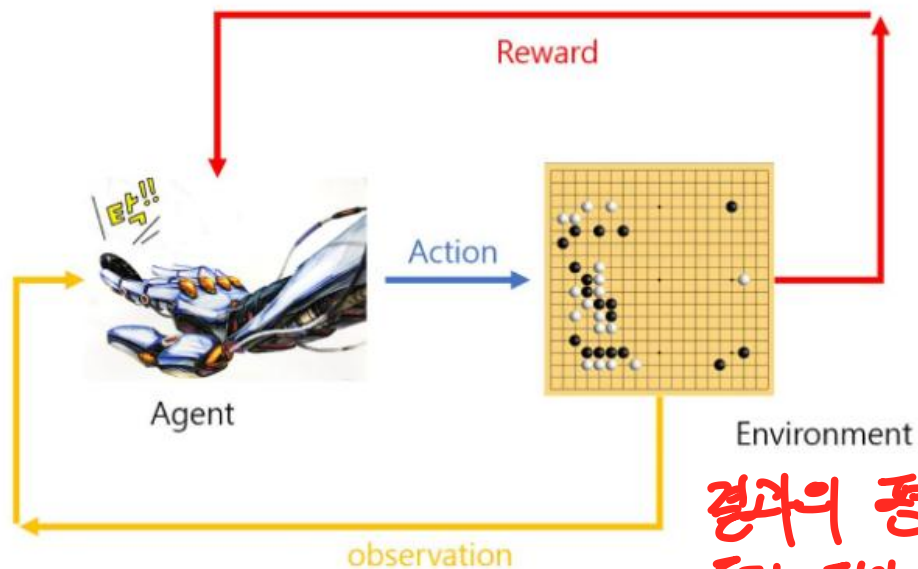
비지도 학습 사례



고객 세분화



강화학습 : q-learning, 정책 경사

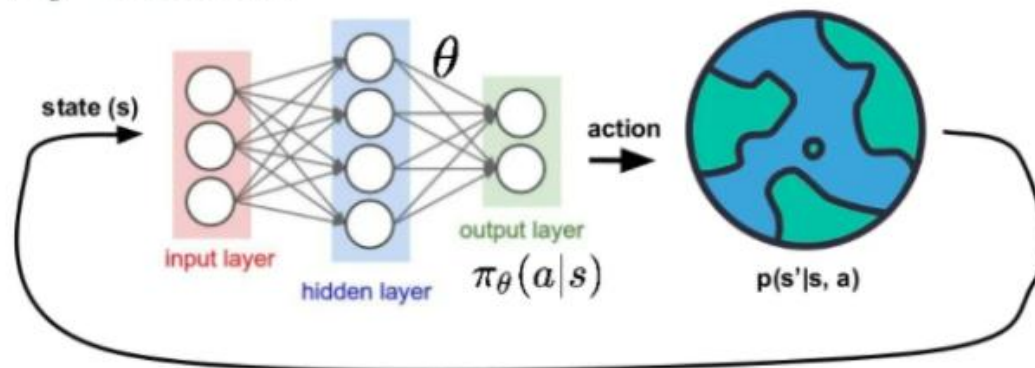


결과에 평가 점수를
통해 평가 점수가 좋은 결과를 선택

$$Q_{\pi}(s, a) = E_{\pi}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots | S_t = s, A_t = a]$$

Q-learning

Policy Gradient



$$J(\theta) = E_{\tau \sim \pi_{\theta}(\tau)}[r(\tau)]$$

정책 경사





지도 학습 심화

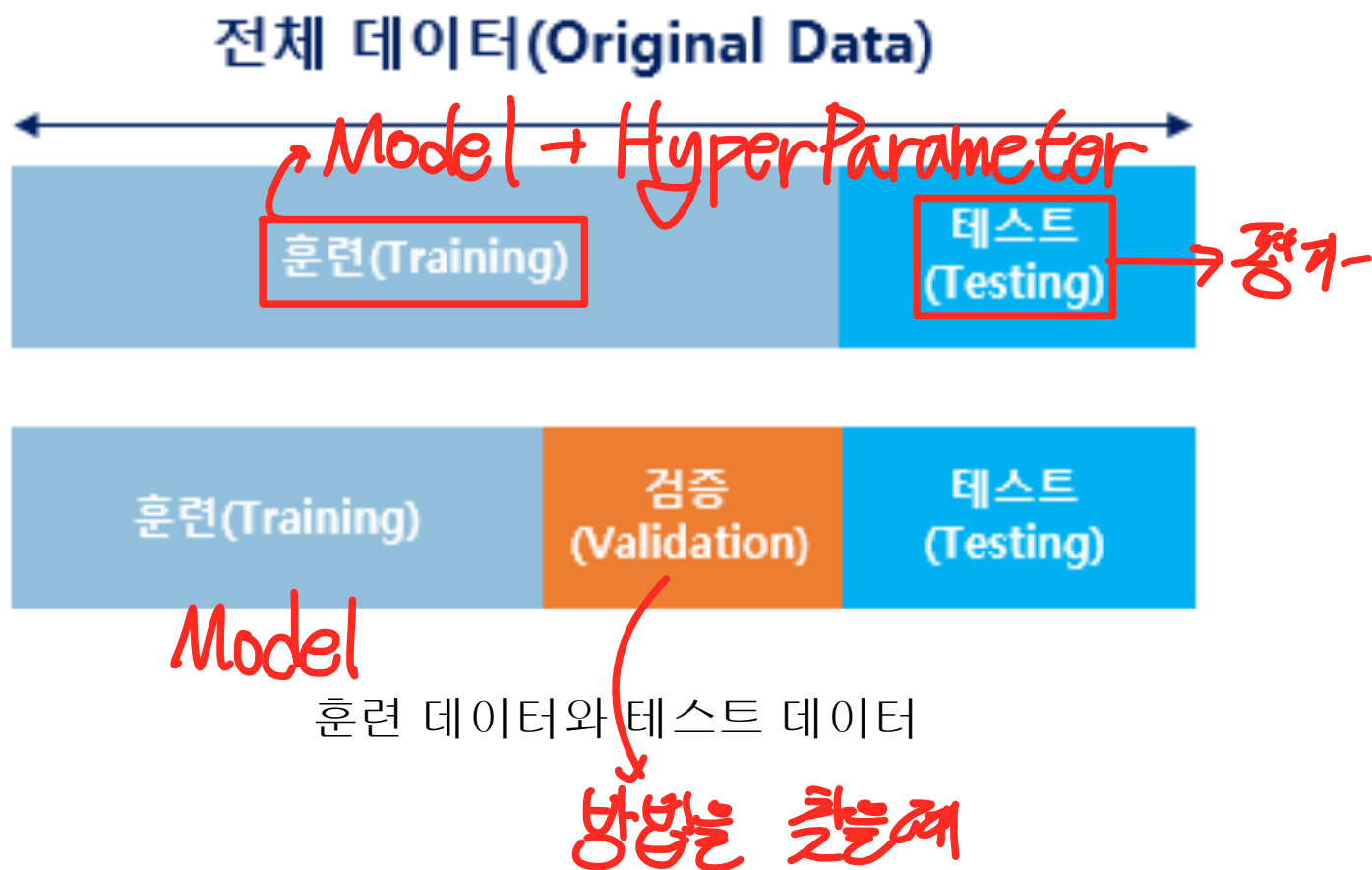


데이터 전처리

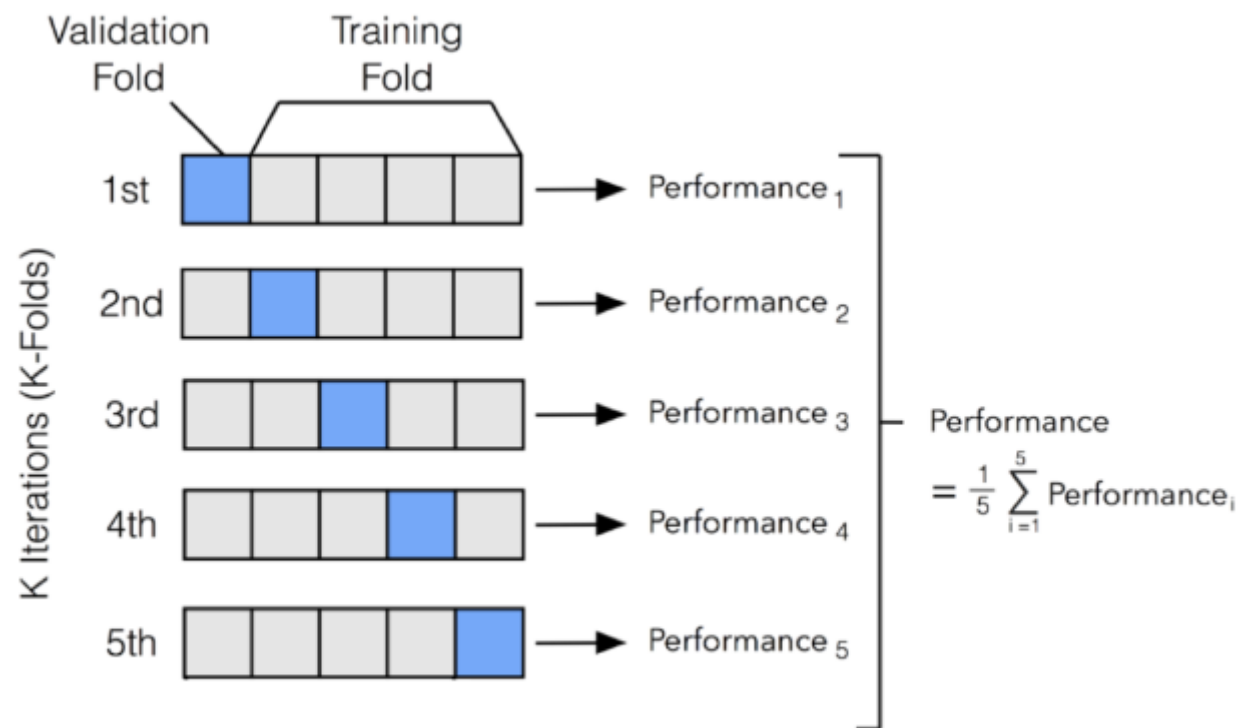
- 결측값 처리 (Handling Missing Value)
 - 데이터셋에서 누락된 값을 다루는 과정
 - 제거 : 결측값을 포함하는 행이나 열을 삭제
 - 대체 : 결측값을 다른 값으로 대체 (평균/중앙값/최빈값 등)
- 정규화 (Normalization)
 - 데이터 값을 일정한 범위로 조정하는 과정 (주로 0과 1사이)
 - 최소-최대 정규화 (Min-Max Normalization)
 - 이상치가 데이터를 왜곡시킬 수 있음
- 특성 스케일링 (Feature Scaling)
 - 데이터의 스케일을 조정하여 모든 특성값이 같은 범위에 위치하도록 하는 과정
 - 표준화 (Standardization) : 데이터의 평균을 0, 표준편차를 1로 조정
 - 데이터의 실제 범위가 없어져 직관적 해석이 어려움



모델 학습 및 평가



모델 학습 및 평가



교차 검증



모델 학습 및 평가

		Actual	
		Positive	Negative
Predicted	Positive	TP	FP
	Negative	FN	TN

Figure 1. Confusion matrix

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

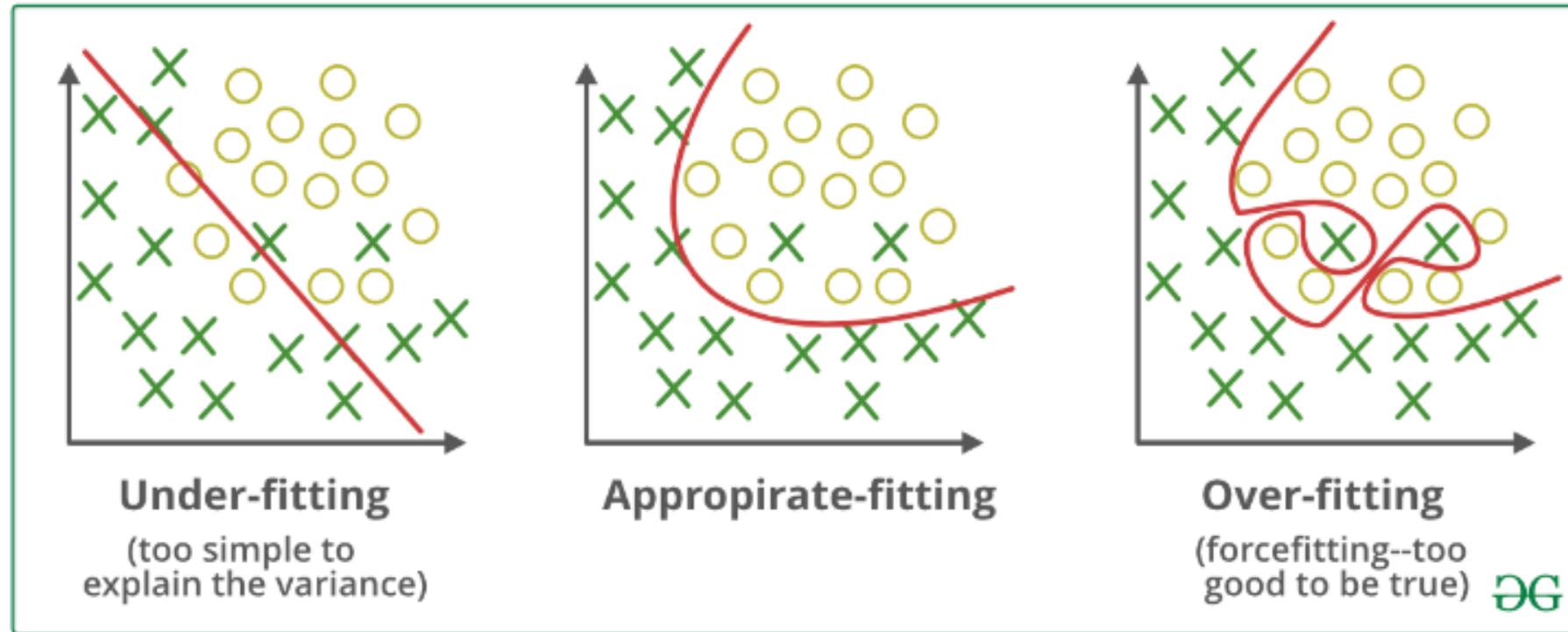
$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$



Overfitting vs Underfitting



과소적합과 과적합





비지도 학습 심화



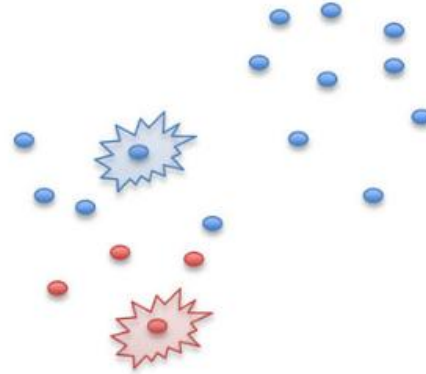
군집화

clustering

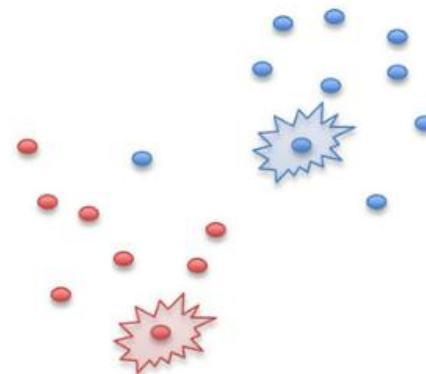
k-means

이진색의 클러스터가 가능한가?

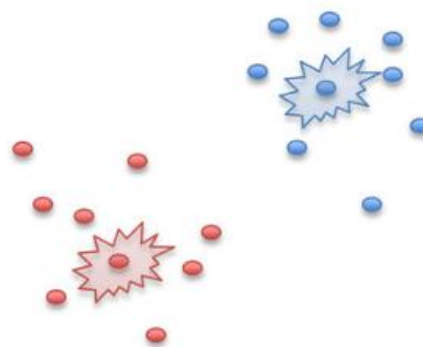
Initial Seeding



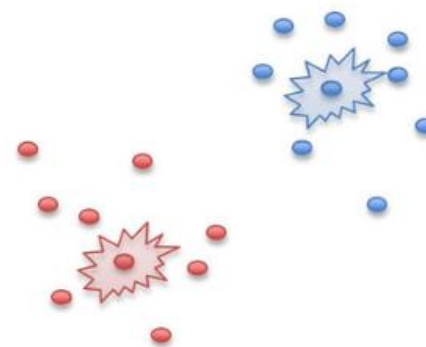
After Round 1



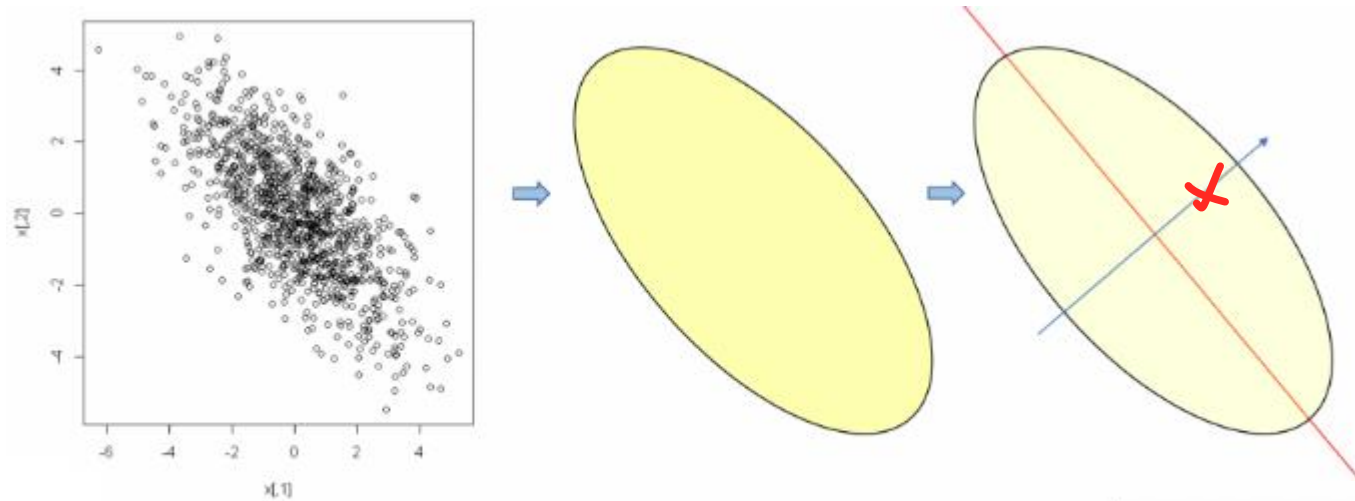
After Round 2



Final

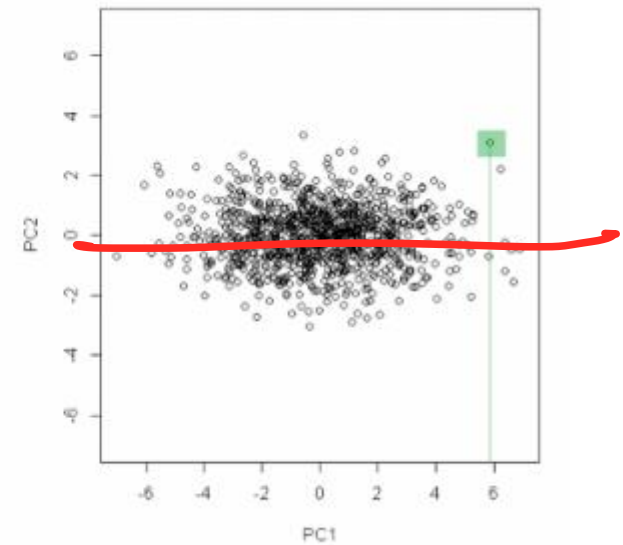
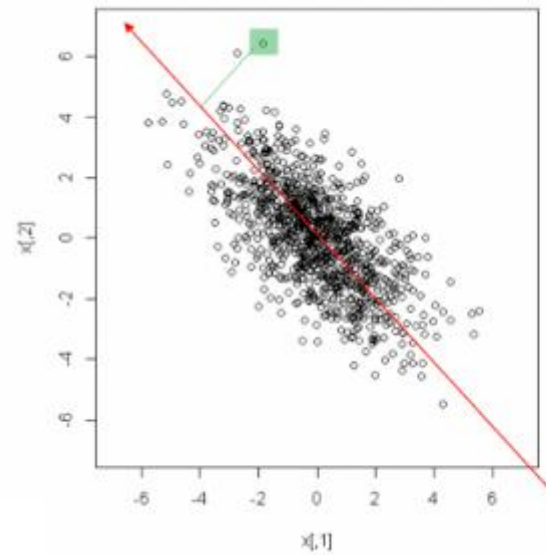


차원 축소



2차원 \rightarrow 1차원

Principal Components의 개념 - 02



SVD (Singular Value Decomposition)

