머신러닝 개요

비타민 8기 1조

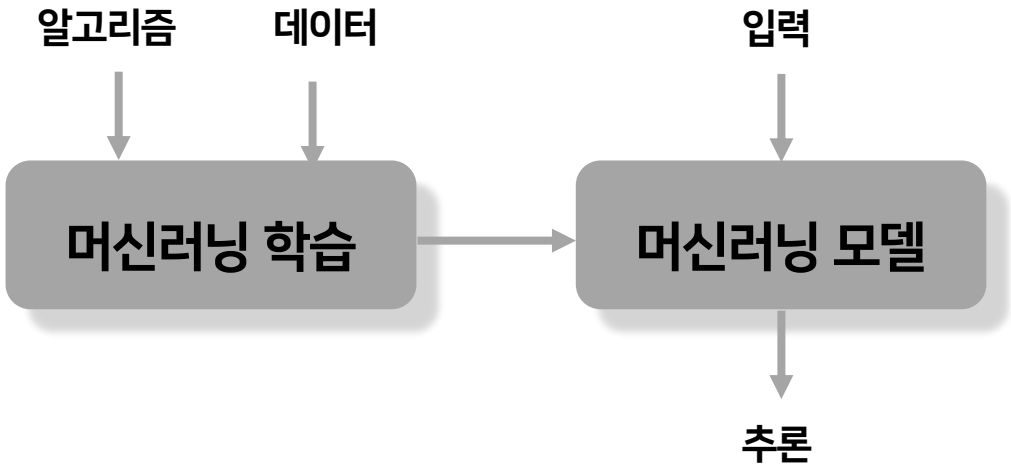
강호재 서진슬 석민정 우상백

INDEX

01 머신러닝

01 머신러닝

(1) 머신러닝이란?

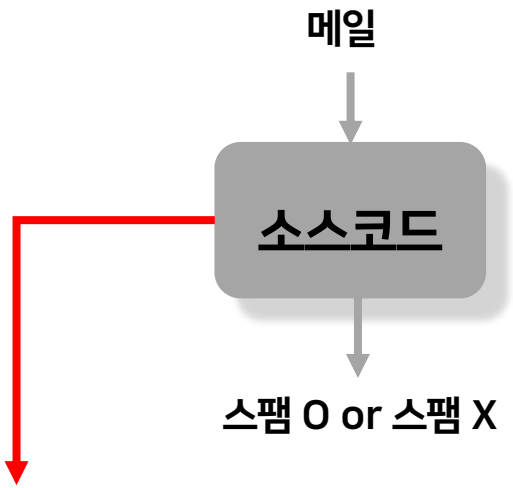


기계가 데이터를 바탕으로
스스로 패턴을 학습하고 결과를 추론하는
알고리즘 기법

01 머신러닝

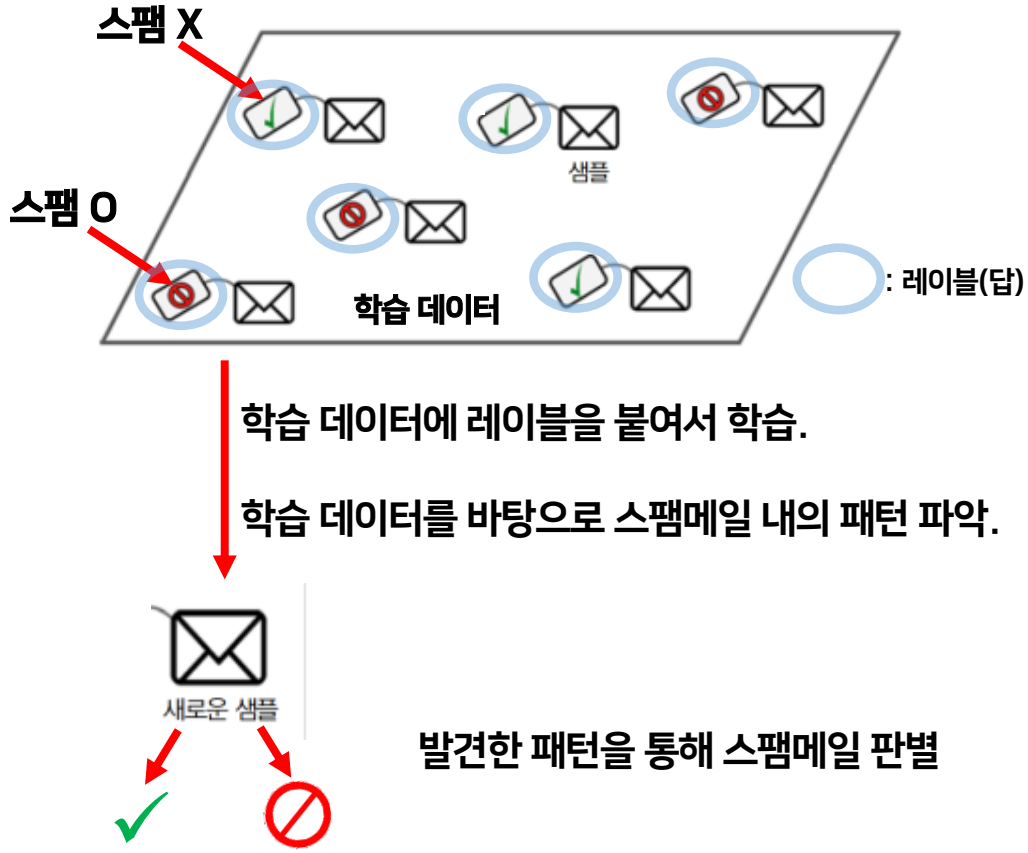
(2) 머신러닝 사용 예시 1 – 스팸메일 판별

전통적인 프로그램



- 스팸 문구들을 일일이 코드로 작성해야 하는 번거로움.
- 언어는 문맥에 따라 스팸메일을 판단해야 함.
  → 특정 단어가 포함되어 있다고 해서 이를 무조건 스팸메일로 볼 수 X
- 유지보수가 어려움.

머신러닝 방식 (지도학습 이용)

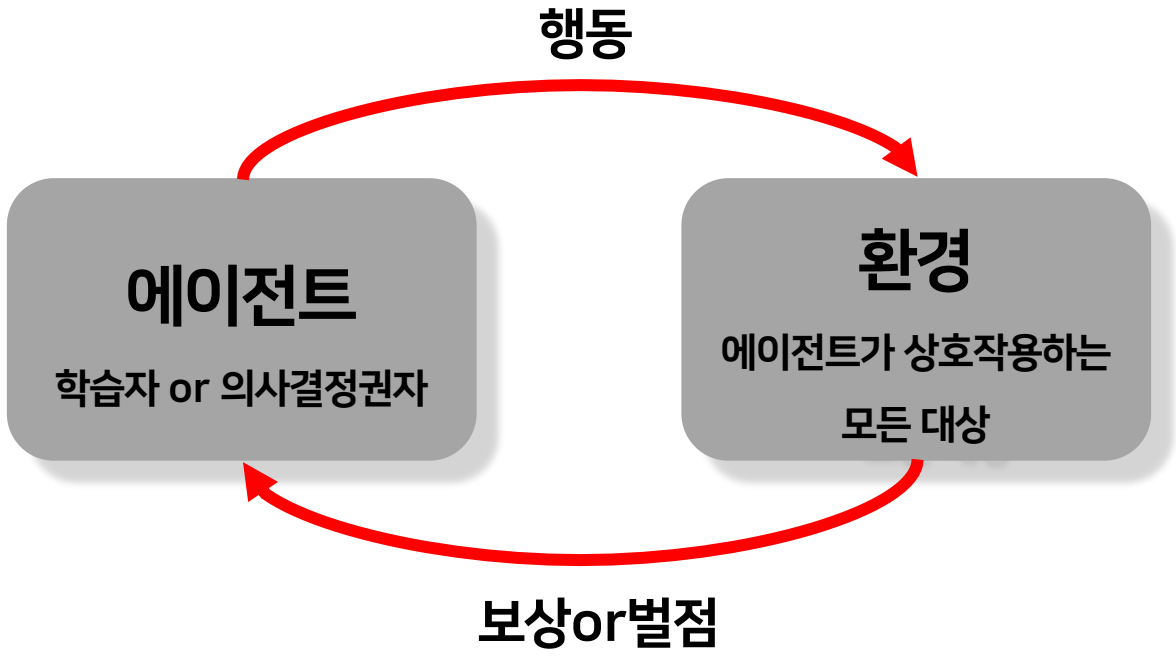


01 머신러닝

(2) 머신러닝 사용 예시 2 – 자율 주행 자동차 주차

강화학습이란 ?

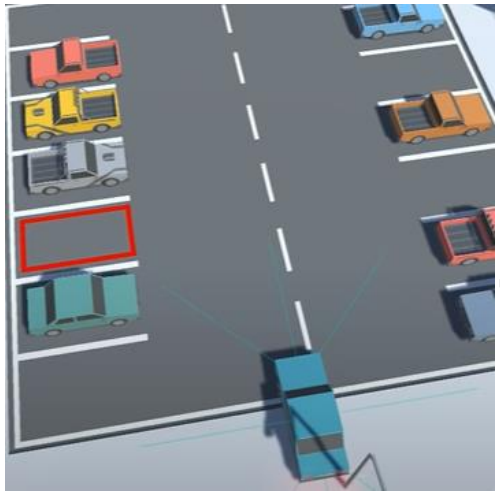
- 시행착오를 거쳐 보상을 극대화 하는 행동을 찾는 것
- 구성 요소



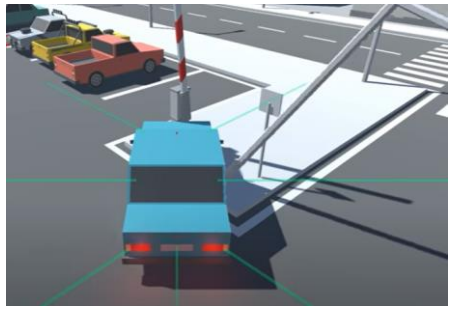
01 머신러닝

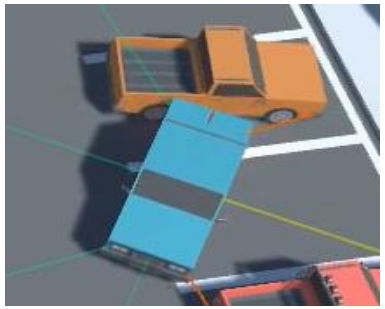
(2) 머신러닝 사용 예시 2 – 자율 주행 자동차 주차

- 목표 : 정확한 위치에 주차하는 것
- 에이전트 : 자동차(컴퓨터)
- 환경 : 주변 차량, 장애물, 날씨 등

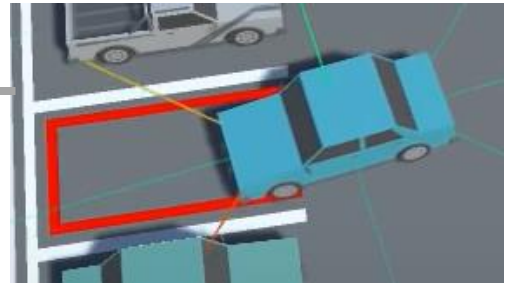


벌점 ex) 충돌할 경우





보상 ex) 주차 위치와 가까워질 경우



수많은 시도

영상

01 머신러닝

(3) 머신러닝 분류

01 머신러닝

(4) 머신러닝 분석절차

문제 정의 → 데이터 수집 → 데이터 탐색 → 데이터 전처리 → 모델링 → 평가

01 머신러닝

(4) 머신러닝 분석절차 – 데이터 탐색

- 기존 통계학: 가설을 세우고 가설을 검정하는 방법론에 치우침
  → 데이터 본래의 정보, 자료가 가지고 있는 본연의 의미를 파악하기 어려움

본연의 데이터 탐색에 집중하자!

탐색적 데이터 분석(EDA)

- 데이터를 다각도에서 관찰하고 전체적으로 데이터를 이해하는 과정

01 머신러닝

(4) 머신러닝 분석절차 – 데이터 탐색

탐색적 데이터 분석(EDA)

- 순서

  1. 데이터의 크기 확인

     - head(), tail() : 상위, 하위 5개 데이터 확인
     - shape() : 행과 열의 개수 확인
     - info() : 열별 결측치 개수 및 데이터 type 확인

     ⋮

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 15 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   survived     891 non-null    int64
 1   pclass       891 non-null    int64
 2   sex          891 non-null    object
 3   age          714 non-null    float64
 4   sibsp        891 non-null    int64
 5   parch        891 non-null    int64
 6   fare         891 non-null    float64
 7   embarked     889 non-null    object
 8   class        891 non-null    category
 9   who          891 non-null    object
 10  adult_male   891 non-null    bool
 11  deck         203 non-null    category
 12  embark_town  889 non-null    object
 13  alive        891 non-null    object
 14  alone        891 non-null    bool
dtypes: bool(2), category(2), float64(2), int64(4), object(5)
memory usage: 80.6+ KB
```

01 머신러닝

(4) 머신러닝 분석절차 – 데이터 탐색

탐색적 데이터 분석(EDA)

- 순서

  2. 데이터의 통계 값 확인

     - describe()
       수치형 데이터의 통계값 요약

     - describe(include = 'O')
       범주형 데이터의 통계값 요약

     - corr()
       각 변수간 상관계수 확인

     ⋮

수치형 데이터 요약

```
df.describe()
```

	survived	pclass	age	sibsp
count	891.000000	891.000000	714.000000	891.000000
mean	0.383838	2.308642	29.699118	0.523008
std	0.486592	0.836071	14.526497	1.102743
min	0.000000	1.000000	0.420000	0.000000
25%	0.000000	2.000000	20.125000	0.000000
50%	0.000000	3.000000	28.000000	0.000000
75%	1.000000	3.000000	38.000000	1.000000
max	1.000000	3.000000	80.000000	8.000000

범주형 데이터 요약

```
df.describe(include = 'O')
```

	sex	embarked	who	embark_town	alive
count	891	889	891	889	891
unique	2	3	3	3	2
top	male	S	man	Southampton	no
freq	577	644	537	644	549

상관계수 확인

```
df.corr()
```

	survived	pclass	age	sibsp	parch	fare	adult_male	alone
survived	1.000000	-0.338481	-0.077221	-0.035322	0.081629	0.257307	-0.557080	-0.203367
pclass	-0.338481	1.000000	-0.369226	0.083081	0.018443	-0.549500	0.094035	0.135207
age	-0.077221	-0.369226	1.000000	-0.308247	-0.189119	0.096067	0.280328	0.198270
sibsp	-0.035322	0.083081	-0.308247	1.000000	0.414838	0.159651	-0.253586	-0.584471
parch	0.081629	0.018443	-0.189119	0.414838	1.000000	0.216225	-0.349943	-0.583398
fare	0.257307	-0.549500	0.096067	0.159651	0.216225	1.000000	-0.182024	-0.271832
adult_male	-0.557080	0.094035	0.280328	-0.253586	-0.349943	-0.182024	1.000000	0.404744
alone	-0.203367	0.135207	0.198270	-0.584471	-0.583398	-0.271832	0.404744	1.000000

01 머신러닝

(4) 머신러닝 분석절차 – 데이터 탐색

탐색적 데이터 분석(EDA)

- 순서

3. 그래프를 통해 시각화 하기

  - 막대그래프
  - 히스토그램
  - 산점도
  - 파이 차트
  - 선 그래프

  ⋮

01 머신러닝

(4) 머신러닝 분석절차 – 데이터 전처리

문제 정의 → 데이터 수집 → 데이터 탐색 → **데이터 전처리** → 모델링 → 평가

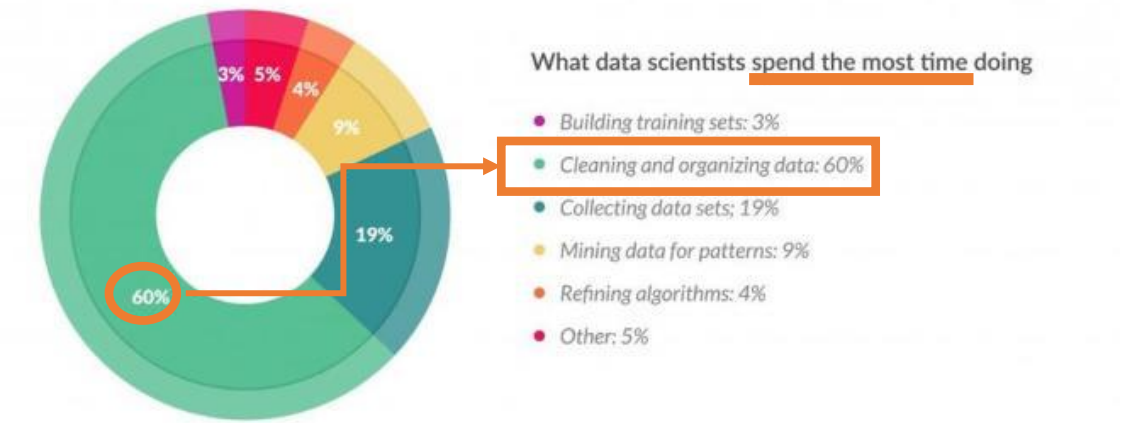
데이터 전처리

- 데이터를 분석하기 전, 분석에 적합한 형태로 데이터를 정돈하는 과정.
- 데이터 분석 단계 중 가장 많은 시간이 소요되는 단계.

방법

- 결측값 처리
- 이상치(outlier) 처리
- 데이터 변환
- 차원 축소

⋮



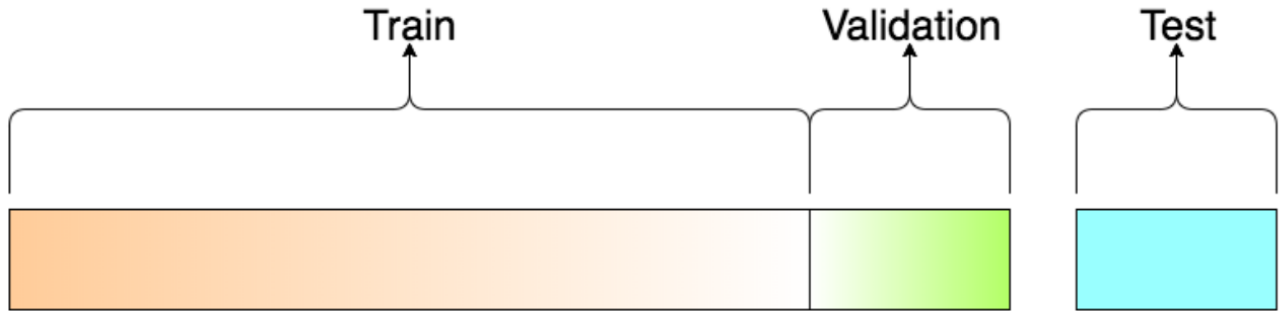
01 머신러닝

(4) 머신러닝 분석절차 – 데이터 전처리

문제 정의 → 데이터 수집 → 데이터 탐색 → **데이터 전처리** → 모델링 → 평가

데이터 분할

- 전처리한 데이터를 학습 데이터(training data), 검증 데이터(validation data), 테스트 데이터(test data)로 분할.
- 주로 전체 데이터의 20%를 테스트 데이터로, 나머지 데이터의 90%를 학습 데이터로, 남은 10%를 검증 데이터로 사용.
- 데이터가 충분하지 않을 경우 교차 검증을 통해 분할.



01 머신러닝

(4) 머신러닝 분석절차 – 모델링

문제 정의 → 데이터 수집 → 데이터 탐색 → 데이터 전처리 → **모델링** → 평가

모델링

- 학습 데이터를 바탕으로 여러 머신러닝 모델을 구현.
- 검증 데이터로 구현된 모델들의 성능을 검증.
- 하이퍼 파라미터를 튜닝함으로 써 최적화된 모델 생성.

* 하이퍼 파라미터(hyper parameter) : 머신러닝 알고리즘별로 최적의 학습을 위해 사용자가 직접 설정하는 파라미터들을 통칭. *

- 검증 결과가 가장 높은 모델을 최종 선택.

01 머신러닝

(4) 머신러닝 분석절차 – 평가

평가

- 선택된 모델의 성능을 테스트 데이터로 측정.
- 결과에 따라 프로젝트를 마무리할지, 이전 단계로 돌아가서 모델을 개선할지 결정.

- 평가지표

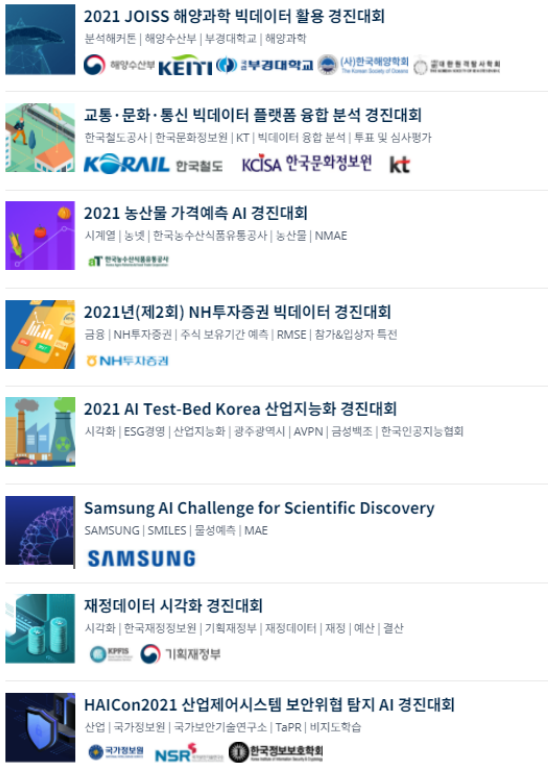
(1) 회귀
- R-square(R^2), MAE(mean absolute error), MAPE , MSE(mean square error), RMSE 등

(2) 분류
- 정확도(accuracy), 오차행렬, 정밀도, F1-score, ROC 곡선과 AUC 등

01 머신러닝

(5) 빅데이터 분석 공모전

DACON



사이트 이동

SEOUL METROPOLITAN GOVENIMENT BIG DATA CAMPUS
서울특별시 빅데이터 캠퍼스

사이트 이동

COMPAS

2021 빅콘테스트
2021 BIG CONTEST

⋮

02 지도학습 & 비지도학습

02 지도학습 & 비지도 학습

(1) 지도 학습 & 비지도 학습

지도 학습

- 지도학습이란 정답이 있는 데이터를 활용해 분석 모델을 학습시키는 것
- 컴퓨터가 학습을 할 때 입력 데이터에 따른 출력 데이터 모두가 필요한 학습 방법
- 손쉽게 모델의 성능을 평가할 수 있다는 장점이 있지만, 데이터마다 레이블을 달기 위해 많은 시간을 투자해야 함

→ 독립 변수에 따른 종속 변수가 존재

VS

비지도 학습

- 비지도학습은 지도학습과는 달리 정답을 알려주지 않고 학습
- 컴퓨터가 학습할 때 입력 데이터만 가지고 그 속에 숨겨진 패턴을 찾아내는 학습 방법
- 레이블이 없기 때문에 모델 성능을 평가하는 데는 다소 어려움이 존재하지만 따로 레이블을 제공할 필요가 없다는 장점

→ 독립 변수에 따른 종속 변수가 없으면 비지도 학습이라 할 수 있다.

02 지도학습 & 비지도 학습

(2) 지도 학습

고양이 강아지

정답
레이블

02 지도학습 & 비지도 학습

(2) 지도 학습

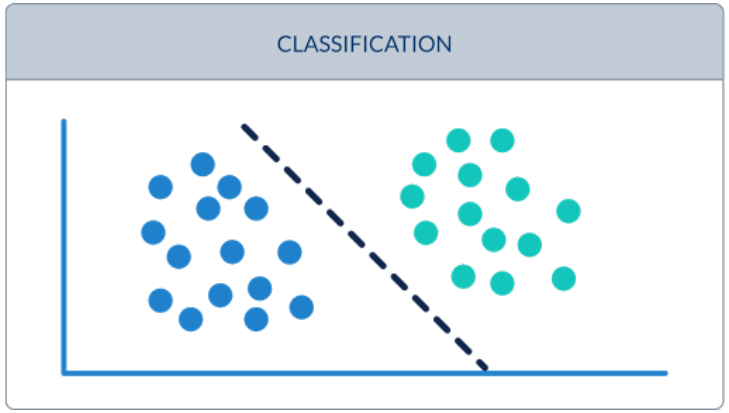
02 지도학습 & 비지도 학습

(2) 지도 학습 - 예시

분류(Classification)분석

대표적인 지도 학습 중 하나로 데이터가 어느 그룹에 속하는지 판별하고자 하는 분석 기법

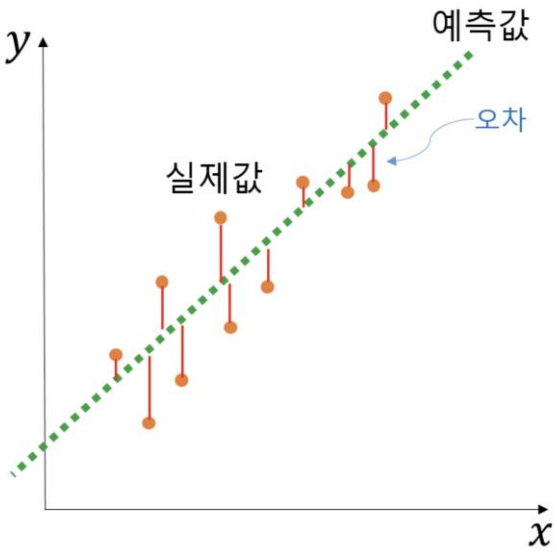
Ex) 의사결정나무, 앙상블 분석, 인공신경망



회귀(Regression)분석

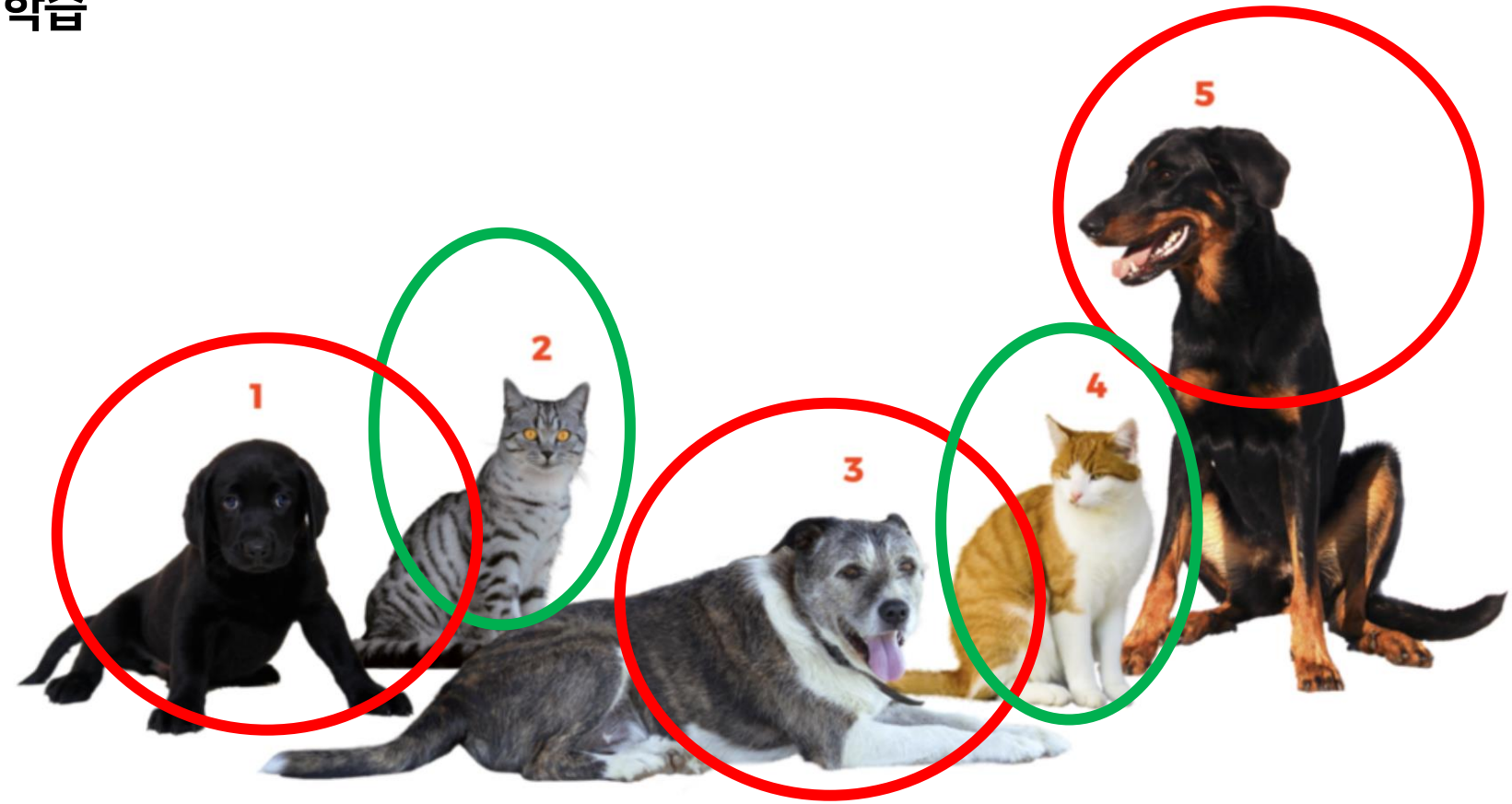
관찰된 연속형 데이터들의 특징(feature)을 토대로 하여 모형을 구축하고 값을 예측하는 분석 기법

Ex) 선형회귀분석, 릿지, 라쏘



02 지도학습 & 비지도 학습

(3) 비지도 학습



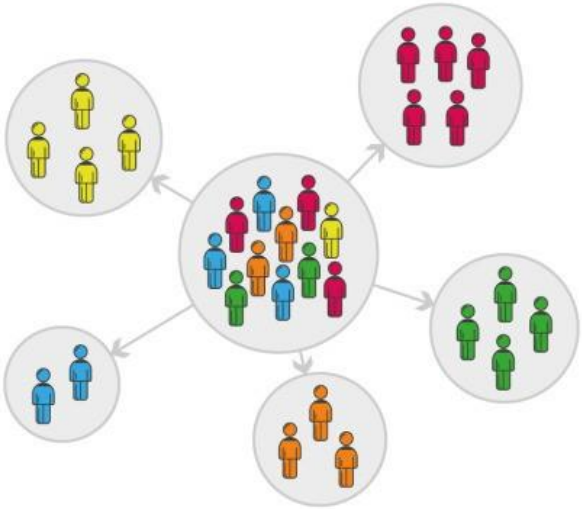
02 지도학습 & 비지도 학습

(3) 비지도 학습 - 예시

군집(Clustering)분석

비지도 학습 중 하나로 여러 이질적인 데이터들 사이의 유사성을 측정하여, 유사성이 높은 객체끼리 하나의 그룹으로 묶는 분석 방법

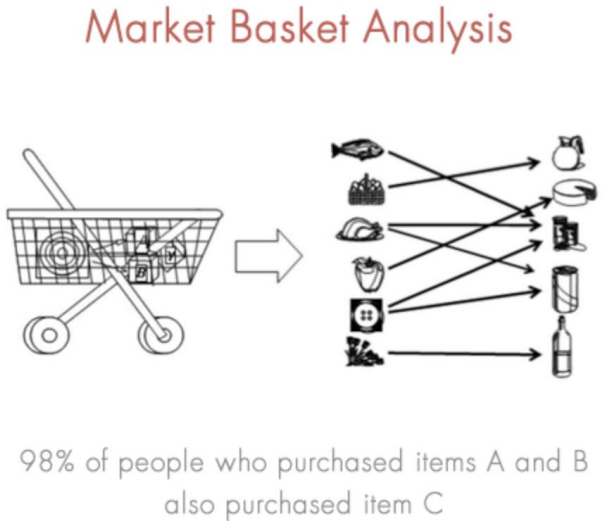
Ex) 병합적 방법, 분할적 방법, K-평균 군집



연관(Association)분석

비지도 학습 중 하나로 데이터의 연관성을 파악하는 분석 방법

Ex) '맥주를 사는 고객은 기저귀를 살 가능성이 높다'와 같이 상품이나 판촉행사 등을 위한 목적으로 사용될 수 있다.



02 지도학습 & 비지도 학습

(4) 지도 학습 & 비지도 학습

구분	지도 학습 (Supervised Learning)	비지도 학습 (Unsupervised Learning)
목적	종속변수 예측 또는 주요 인자 발견	데이터 특징, 성향 파악
기준	종속변수가 있는 문제	종속변수가 없는 문제
특징	대체로 분석 목적이 명확	대체로 분석 목적이 불명확
방법	회귀, 분류	그룹화, 차원축소

02 지도학습 & 비지도 학습

(4) 지도 학습 & 비지도 학습

지도 학습		비지도 학습	
회귀 (연속형)	**선형회귀분석** 의사결정나무 **SVM** 신경망 모형 릿지 라쏘	**군집**	K-means SOM DBSCAN 병합군집 계층군집
		연관	Apriori
분류 (범주형)	**로지스틱 회귀분석** 신경망 모형 의사결정나무 **k-NN** 앙상블모형 **SVM** 나이브 베이즈 분류	**차원 축소**	**PCA(주성분분석)** **LDA(선형판별분석)** **SVD(특이값 분해)** **MDS(다차원 척도법)**

03 분류 & 회귀

03 분류 & 회귀

(1) 분류(classification)

- 분류는 이산 값을 포함하는 유한 집합에서 레이블(각 데이터에 정해진 특징)을 예측하는 것
- 쉽게 말하면, 과일 중, 사과와 딸기를 분류하는 것.
- ex) Binary / Multi - class classification (이진분류/ 다중분류)
- ex) Multi – label classification(다중 레이블 분류)

03 분류 & 회귀

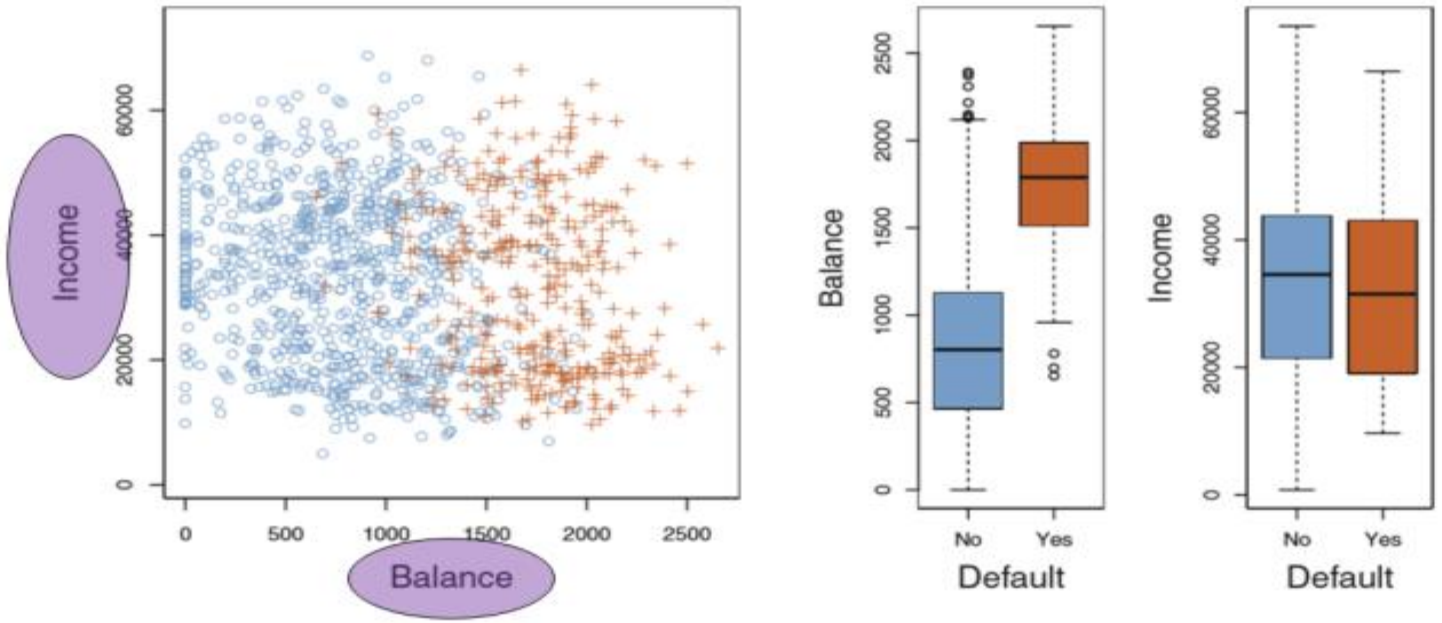
(1) 분류(classification) - 예시

	독립변수	종속변수	방법론 예시
이진분류	공부시간	합격 여부 (합격/불합격)	학생들의 공부시간을 입력받고, 최종 합격여부 확인.
이진분류	X-ray 사진과 영상 속 종양의 크기, 두께	악성 종양 여부 (양성/음성)	의학적으로 양성과 음성이 정확하게 확인된 사진과 영상 데이터를 모은 뒤, 실제 진료에서 양성 판별.
다중분류	품종, 산도, 당도, 지역, 연도	와인의 등급 (1/2/3/4/5등급)	소믈리에를 통해서 등급이 확인된 와인을 가지고 품종, 산도 등의 독립변수를 정하고 기록.
다중 레이블 분류	대상자1의 직업, 대상자2의 직업	뉴스 카테고리 (스포츠/정치/연예/시사/세계)	결혼 뉴스의 대상자1과 대상자2의 직업을 입력받고, 뉴스 카테고리(ex. (스포츠,연예)) 분류.

03 분류 & 회귀

(1) 분류 – 이진분류(Binary Classification) 예시

예시 : 잔고, 수입 현황 → 채무 불이행 Yes or No



SW

'잔고'가 많은 사람이 '채무 불이행'을 더 많이 하는 것으로 '예측' 가능!
('수입'과는 크게 관련 없는 것으로 예측.)

03 분류 & 회귀

(1) 분류 – 다중분류(Multi-class Classification) 예시

그림 속의 대상이 되는 객체는 **'한 개'**만 있어야 하고,
그 객체는 **'2개 이상의 카테고리'**(고양이/개/토끼/앵무새)에 속하는 경우

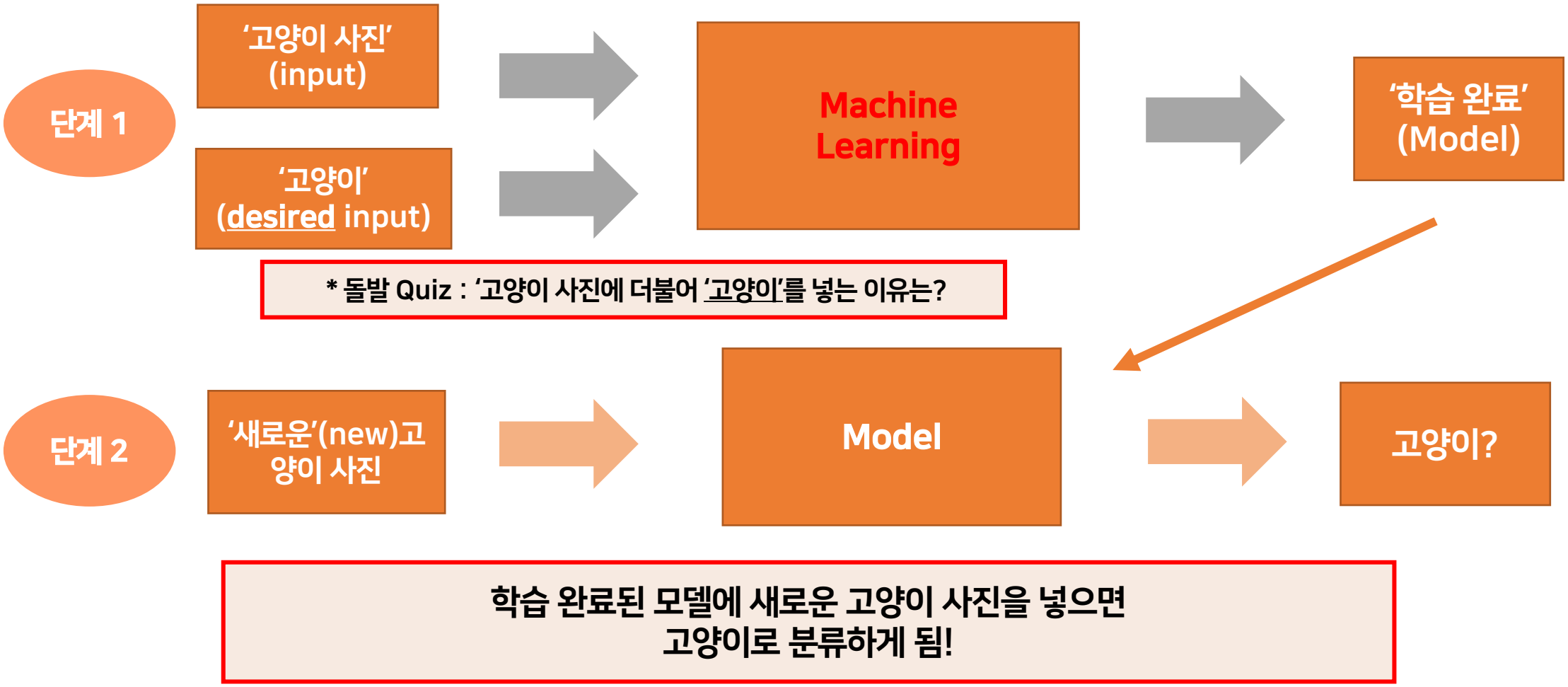
03 분류 & 회귀

(1) 분류 – 다중 레이블 분류(Multi – label Classification) 예시

그림 속의 대상이 되는 객체는 **'여러 개'**가 있어야 하고, 그 **객체는 '2개 이상의 카테고리'**(고양이/개/토끼/앵무새)에 속하는 경우

03 분류 & 회귀

(1) 분류 – 모델링



학습 완료된 모델에 새로운 고양이 사진을 넣으면
고양이로 분류하게 됨!

03 분류 & 회귀

(2) 분류 vs 군집

분류(Classfication)

- 지도(교사)학습(supervised)
- 즉, 입력값에 대한 '출력값' 인 명확한 목표, 정답 존재.
- 즉, 목표 값에 대하여 '예측' 하고자 함.
- 동물 사진을 넣고, 개인지 고양이 인지 머신러닝을 통하여 판별하는 문제

VS

군집(Clustering)

- 비지도(교사)학습(unsupervised)
- 즉, 입력값에 대한 '출력값' 인 명확한 목표, 정답이 존재하지 않음.
- 즉, 목표 값이 정해져 있지 않으므로, 데이터 셋 간 유의미한 '상관관계', '패턴'등을 찾아 내고자 함.
- 동물 사진을 넣고, 동물들의 유사성을 머신러닝을 통하여 분석하는 문제

03 분류 & 회귀

(3) 회귀(Regression)

- **회귀**는 **연속형 값을 포함하는 유한 집합**에서 **연속된 값을 예측**하는 것
- 쉽게 말하면, **직원들의 성과**를 토대로 **연봉** 예측
- ex) 단순선형회귀분석(Simple Linear Regression)

03 분류 & 회귀

(3) 회귀(Regression) – 예시

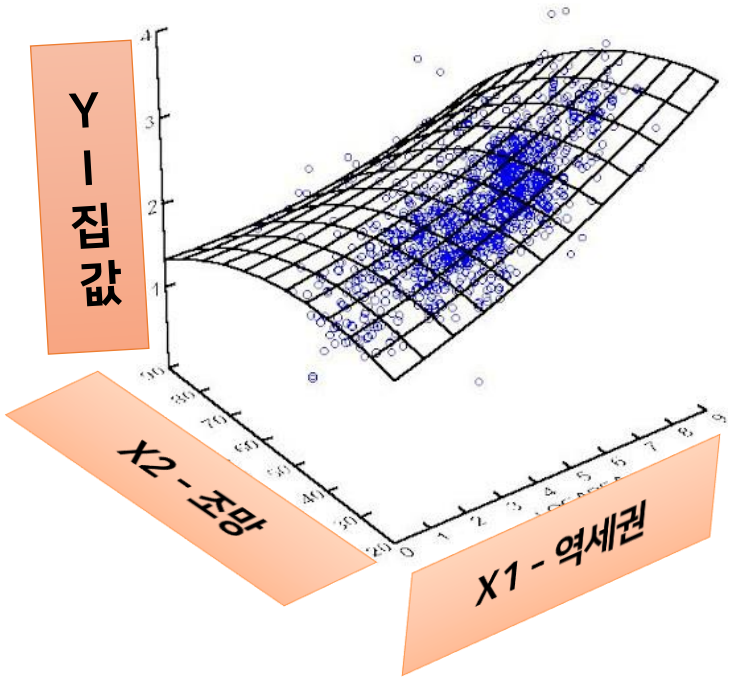
독립변수	종속변수	방법론 예시
공부시간	시험점수	학생들의 공부시간을 입력받고, 최종 점수를 확인.
역세권, 조망, 욕실 크기, 침실 크기	집값(매매액)	집과 역까지의 거리, 수치화된 조망의 평점 등, 욕실과 침실 크기를 통해 집 값을 확인.
나이	키	학생들의 나이에 따른 키를 예측.
자동차 속도	충돌 시 사망확률	충돌 시 속도와 사상자를 기록한다.

03 분류 & 회귀

(3) 회귀(Regression) – 다중회귀모형 예시

X

X1(집- 역 측정거리)
X2(조망 평점)



Y

Y(집 값 실매매액)

다중회귀모형 예시

03 분류 & 회귀

(4) 분류 & 회귀 결론

- **분류**는 **이산 값을 포함하는 유한 집합**에서 **레이블**(각 데이터에 정해진 특징)을 **예측**하는 것
- **회귀**는 **연속형 값을 포함하는 유한 집합**에서 **연속된 값**을 **예측**하는 것
- 무언가를 **'예측'** 하는 데에 있어서 공통점이자 핵심점.

04 과대적합 & 과소적합

04 과대적합 & 과소적합

(1) 과소적합(underfitting)

: 데이터에서 충분한 특징을 찾아내지 못하여 머신러닝 모델을 학습할 때 발생

: 너무 간단한 모델이 선택되는 것 (= 너무 편향되어(biased) 학습된 모델)

학습 데이터

사물	생김새	분류값
야구공	동그라미	공
농구공	동그라미	공
테니스공	동그라미	공
딸기	세모	과일

사과 -> 공?

데이터의 특징(feature)이 생김새 뿐 -> 생김새가 동그라미이면 공

-> 공을 구별할 수 있는 특징이 너무 적음 -> 높은 정확도를 가질 수 없음 -> 과소적합된 모델

04 과대적합 & 과소적합

(2) 과대적합(overfitting)

: 필요 이상의 특징을 발견해서

학습 데이터에 대한 정확도는 높지만, 테스트 데이터의 정확도가 낮게 나오는 모델

: 너무 복잡한 모델이 선택되는 것 (= variance가 높게 학습된 모델)

학습 데이터

사물	생김새	크기	줄무늬	분류값
야구공	원형	중간	있음	공
농구공	원형	큼	있음	공
테니스공	원형	중간	있음	공
딸기	세모	중간	없음	과일
포도알	원형	작음	없음	과일

"생김새가 원형이고
크기가 작지 않으며,
줄무늬가 있으면 공이다 "

04 과대적합 & 과소적합

(2) 과대적합(overfitting)

"생김새가 원형이고 크기가 작지않으며, 줄무늬가 있으면 공이다 "

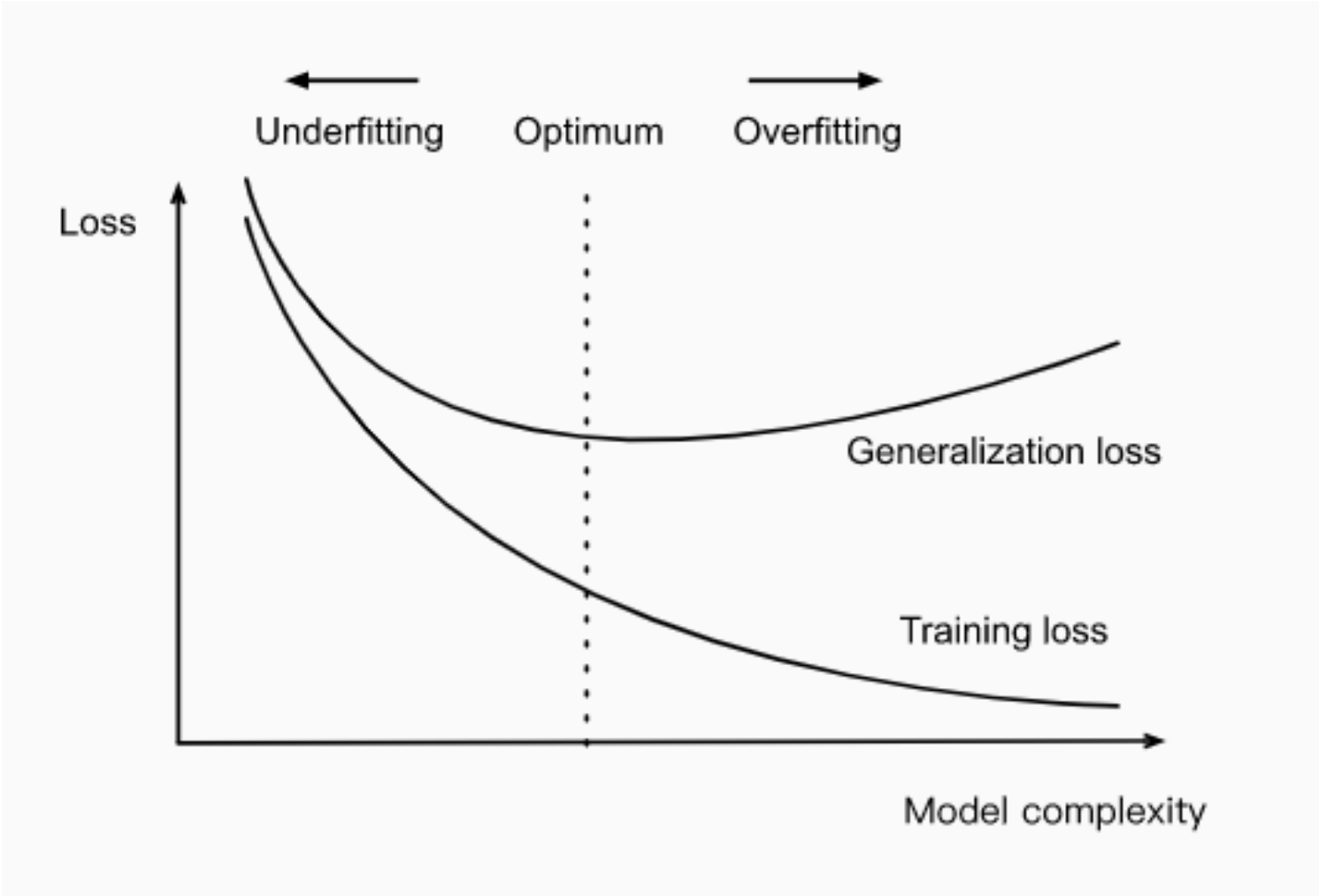
테스트 데이터

사물	생김새	크기	줄무늬
골프공	원형	작음	없음
수박	원형	큼	있음
당구공	원형	중간	없음
럭비공	타원형	큼	있음

테스트 데이터에 대한 정확도 : 0%

04 과대적합 & 과소적합

(3) 과대적합과 과소적합



04 과대적합 & 과소적합

(4) Bias와 Variance

학습데이터를 잘 설명하는 모델 = Training error를 minimize하는 모델

$$MSE_{(trainig)} = (Y - \hat{Y})^2$$

테스트 데이터를 잘 설명하는 모델 = 테스트 데이터에 대한 expected error가 낮은 모델

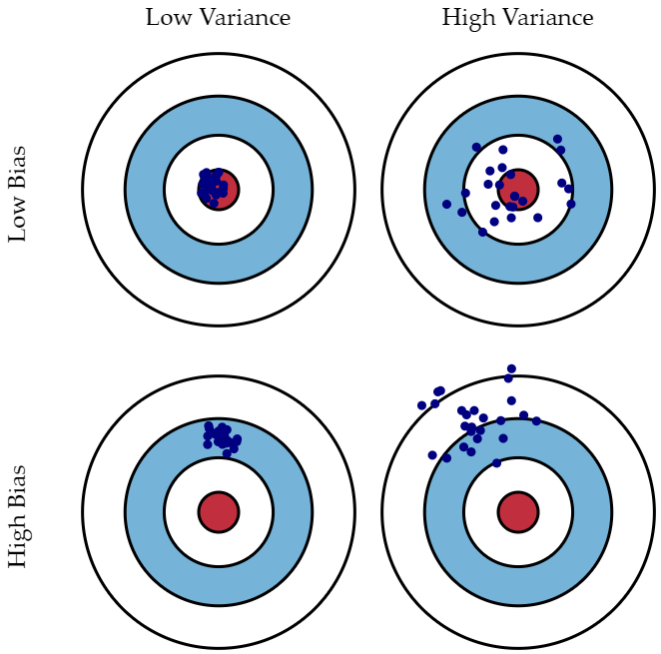
$$\begin{aligned}\text{Expected MSE} &= E\left[(Y - \hat{Y})^2 | X\right] \\ &= \sigma^2 + (E[\hat{Y}] - \hat{Y})^2 + E[\hat{Y} - \mathrm{E}[\hat{Y}]]^2 \\ &= \sigma^2 + \mathrm{Bias}^2(\hat{Y}) + \mathrm{Var}(\hat{Y}) \\ &= \text{Irreducible Error} + \mathrm{Bias}^2 + \text{Variance}\end{aligned}$$

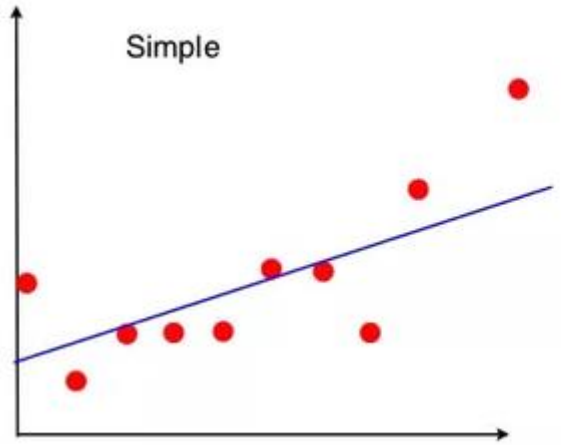
04 과대적합 & 과소적합

(4) Bias와 Variance의 직관적 해석

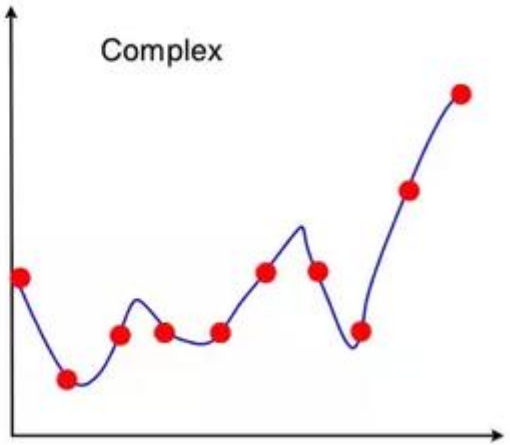
Bias : 예측된 값들이 실제 값에서 얼마나 멀어져 있는가

Variance : 예측된 값들이 서로간에 얼마나 멀리 떨어져 있는가





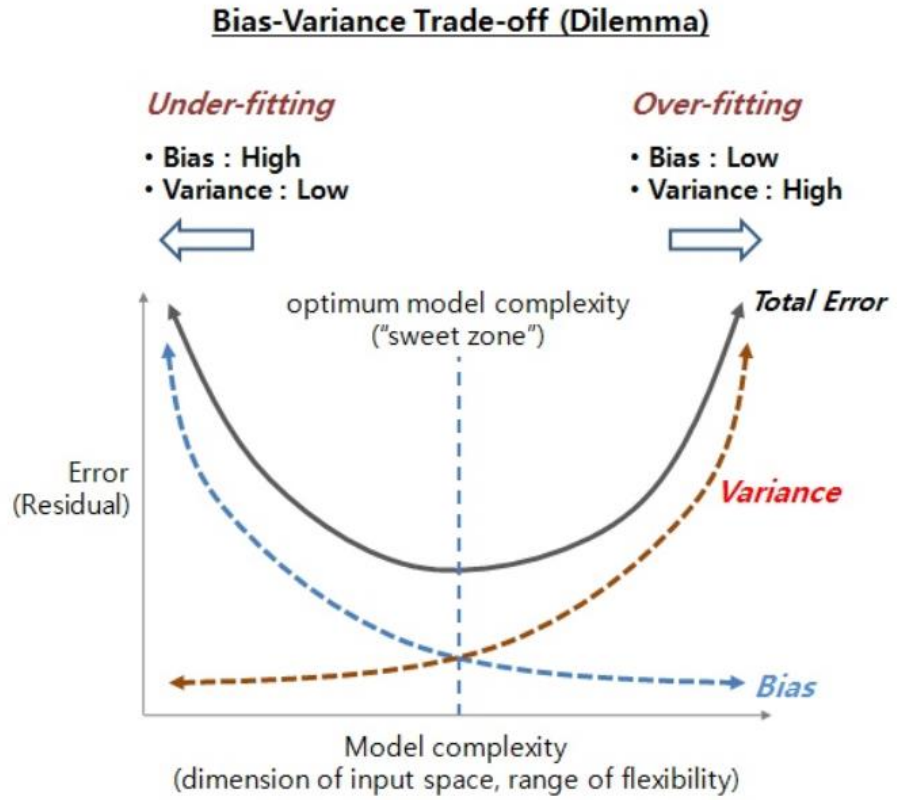
(high bias, low variance)



(low bias, high variance)

04 과대적합 & 과소적합

(4) Bias와 Variance의 trade-off



적절한 모델이라는 것은 결국
분산과 편향의 균형을 고려하여 한쪽으로 치우치지 않는 최적의 복잡도를 찾는 것

04 과대적합 & 과소적합

(5) 해결방법

과소적합 해결방법

- 특징을 더 찾는 것
- 과소적합된 모델은 bias가 높은 모델이므로, bias는 낮추고, variance를 높인다

  → Variance가 높은 모델을 사용해봄 (ex. Decision Tree, k-NN, SVM)

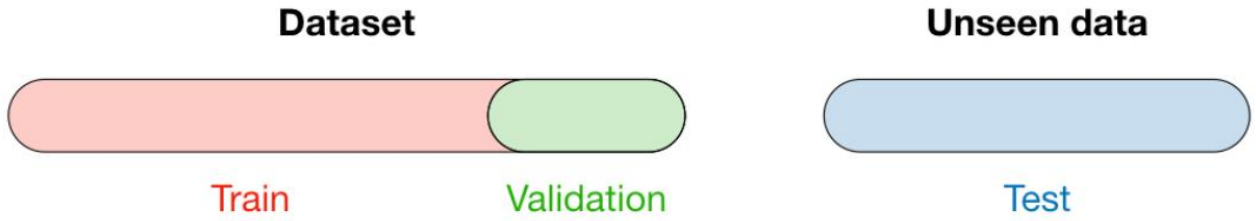
과대적합 해결방법

- 더 많은 학습데이터 확보
- 학습에 사용된 특징(feature)의 수 줄이기
- 특징(feature)들의 수치값을 정규화 → 특정 특징에 의한 편향(bias)을 줄이기
- 검증 데이터셋을 갖추는 것
- 딥러닝의 경우 , 조기 종료 및 드랍 아웃 사용

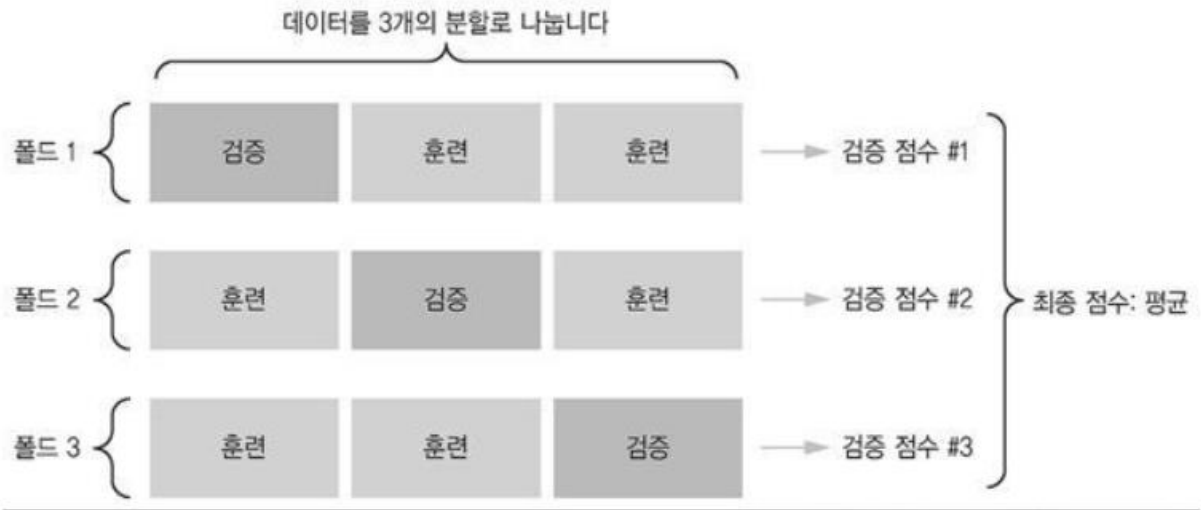
04 과대적합 & 과소적합

(5) 과대적합 해결방법 – 검증 데이터 셋 갖추기

홀드아웃 데이터셋(holdout dataset)

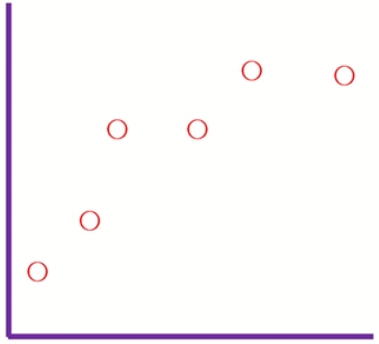


K-겹 검증 (K-fold validation)



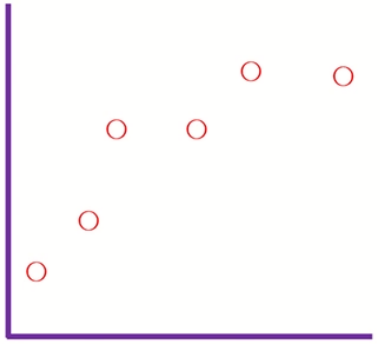
04 과대적합 & 과소적합

(5) 과대적합 해결방법 – 정규화 (Regularization)



$\beta_0 + \beta_1 x$

underfitted

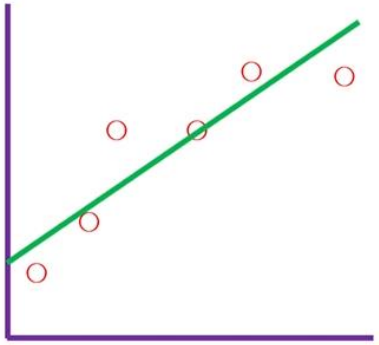


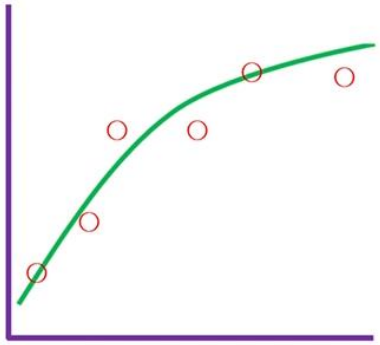
$\beta_0 + \beta_1 x + \beta_2 x^2$

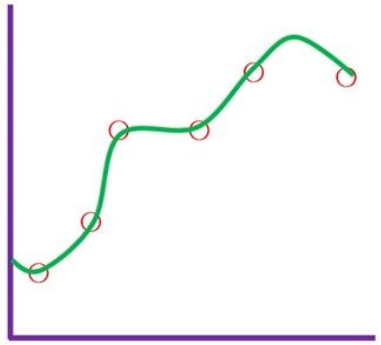
굿굿

$\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4$

overfitted

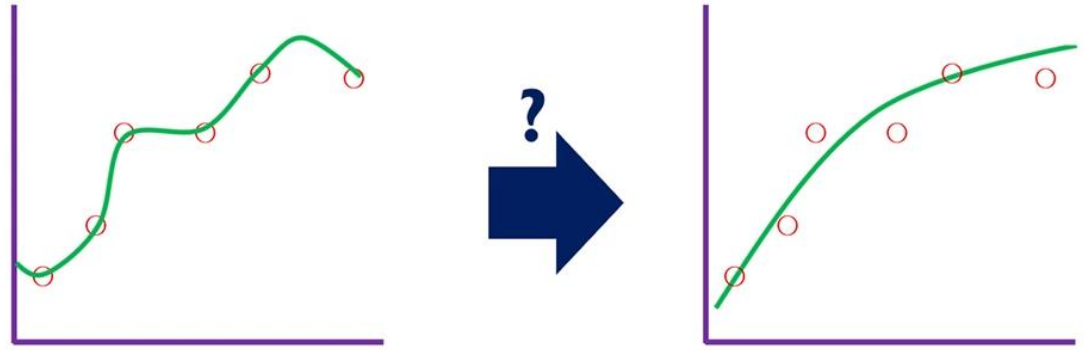






04 과대적합 & 과소적합

(5) 과대적합 해결방법 – 정규화 (Regularization)



$\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4$

$\beta_0 + \beta_1 x + \beta_2 x^2$

$$\min_{\beta} \sum_{i=1} (y_i - \hat{y}_i)^2 + 5000\beta_3{}^2 + 5000\beta_4{}^2$$

$\beta_3 \approx 0$ $\beta_4 \approx 0$

Q & A

감사합니다