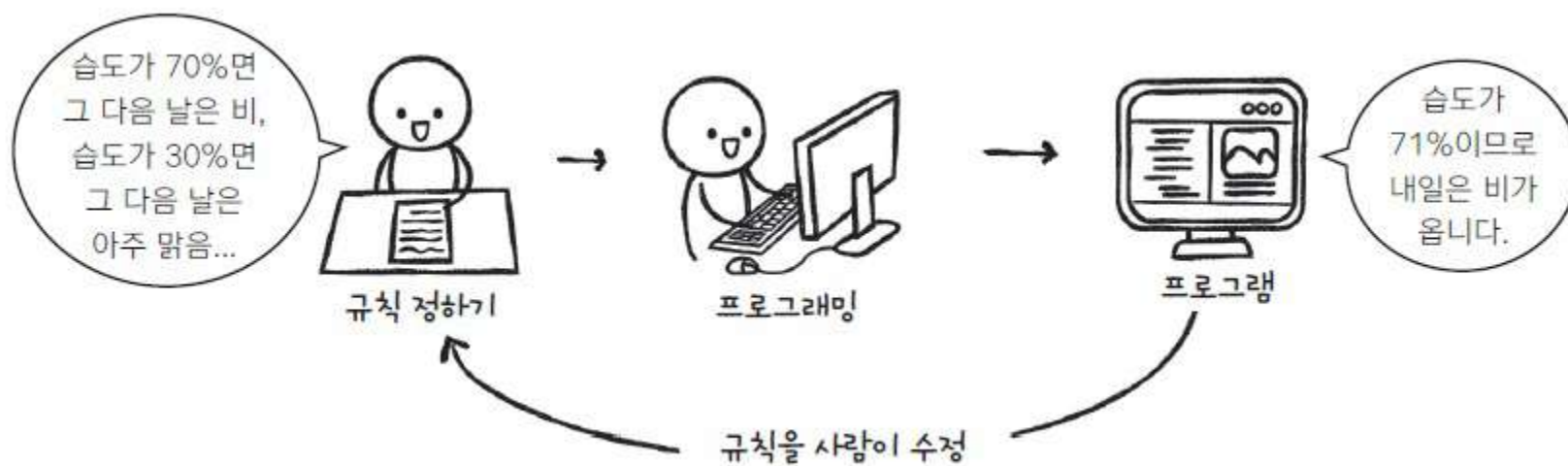


Artificial Intelligence

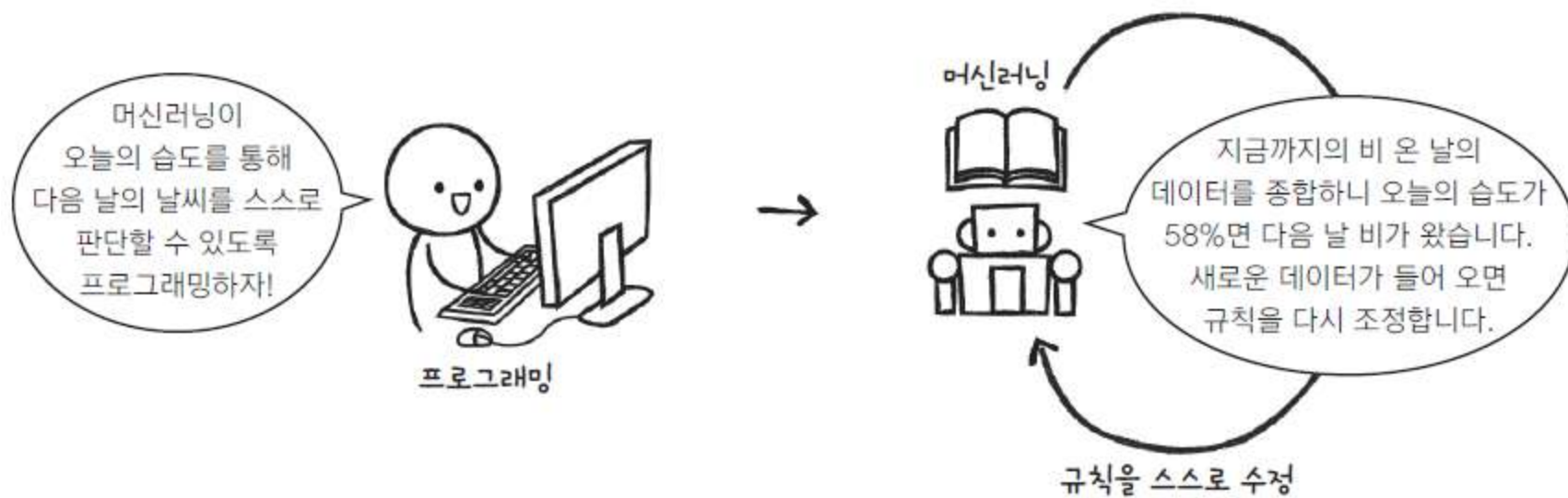
머신러닝의 기초



전통적인 프로그램은 사람이 규칙을 수정



머신러닝은 스스로 규칙을 수정

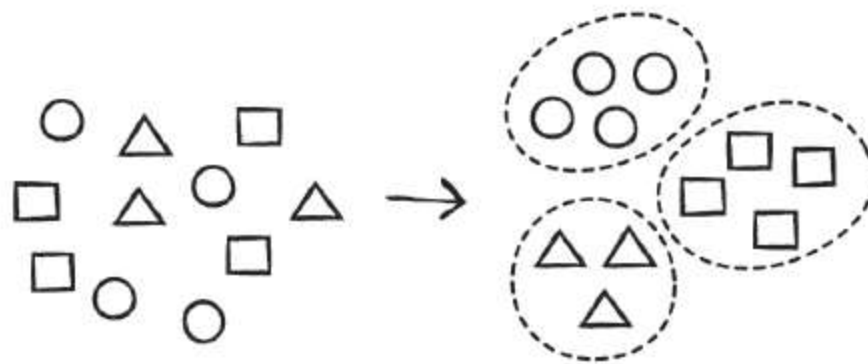


머신러닝은 학습 방식에 따라
지도 학습(supervised learning),
비지도 학습(unsupervised learning),
강화 학습(reinforcement learning)으로 분류

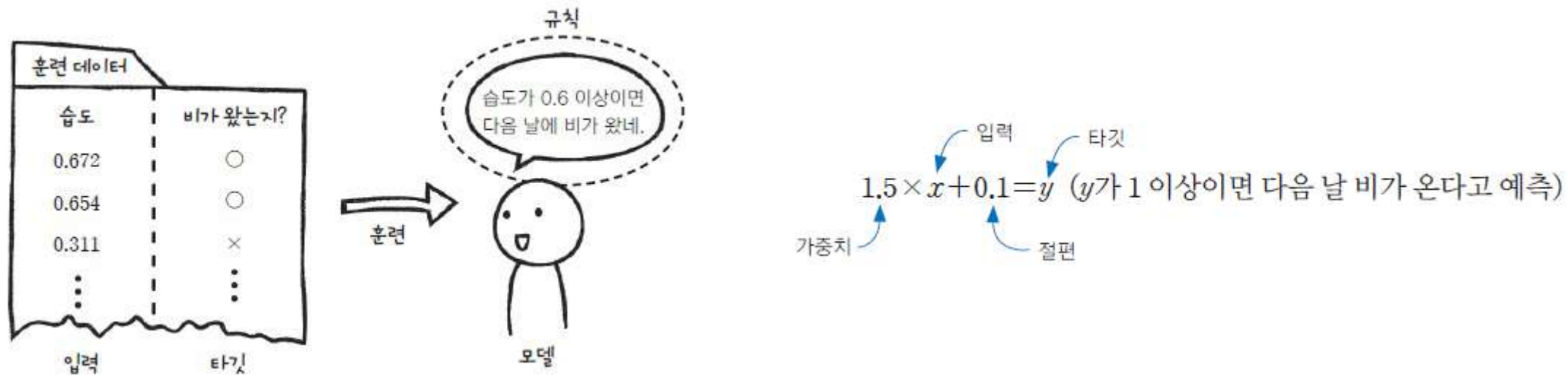
지도 학습은 입력과 타겟으로 모델을 훈련



비지도 학습은 타깃이 없는 데이터를 사용(ex. 군집(clustering))



규칙이란 가중치와 절편을 의미



손실 함수로 모델의 규칙을 수정

훈련 데이터	
습도	비가 왔는지?
0.672	○
0.654	○
0.311	×
0.472	○

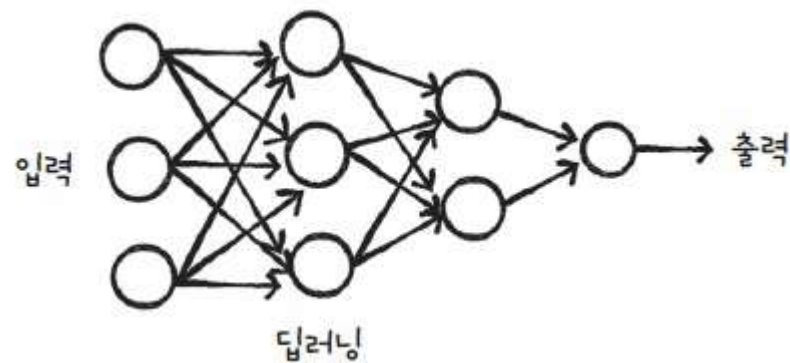
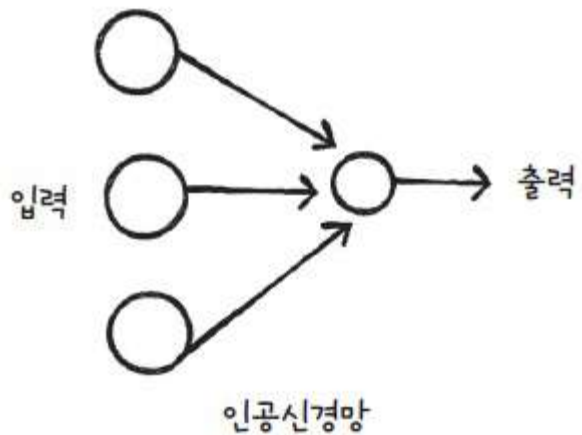
새로운 입력과 출력

$$1.5 \times x + 0.1 = y$$

(y가 1 이상이면 다음 날 비가 온다고 예측)

가중치 입력 타겟 절편

인공신경망을 여러 겹 쌓으면 딥러닝



딥러닝은 머신러닝이 처리하기 어려운 데이터를 더 잘 처리함(비정형 데이터)

딥러닝에 잘 맞는 데이터



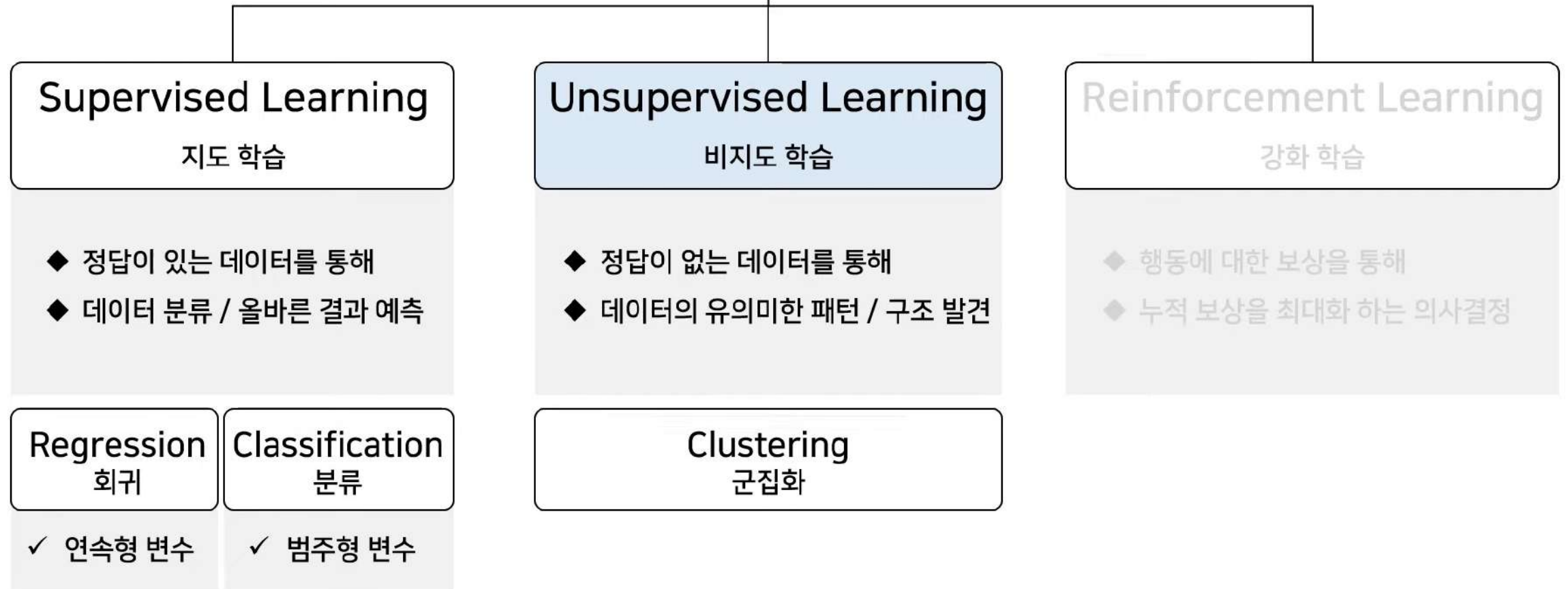
이미지/영상, 음성/소리, 텍스트/번역
등의 비정형 데이터

머신러닝에 잘 맞는 데이터



데이터베이스, 레코드 파일, 엑셀/CSV
등에 담긴 정형 데이터

Machine Learning



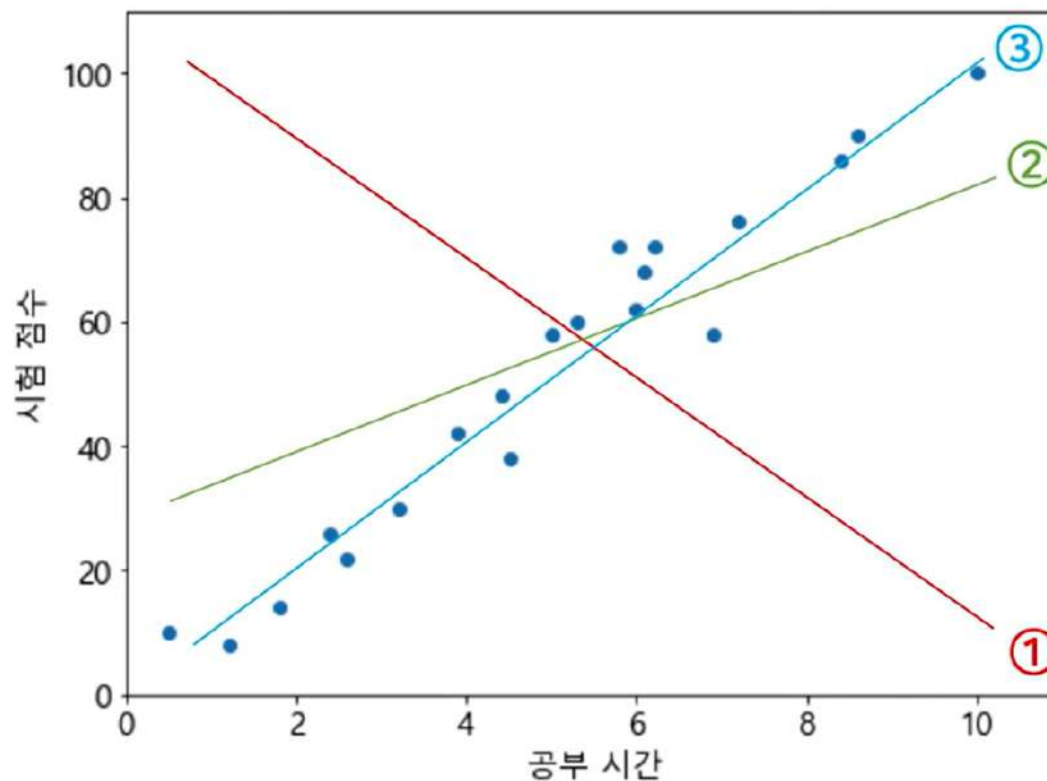
Linear Regression

선형 회귀

선형회귀(Linear Regression)

공부 시간	시험 점수
0.5	10
1.2	8
1.8	14
2.4	26
2.6	22
3.2	30
3.9	42
4.4	48
4.5	38
5	58
5.3	60
5.8	72
6	62
6.1	68
6.2	72
6.9	58
7.2	76
8.4	86
8.6	90
10	100

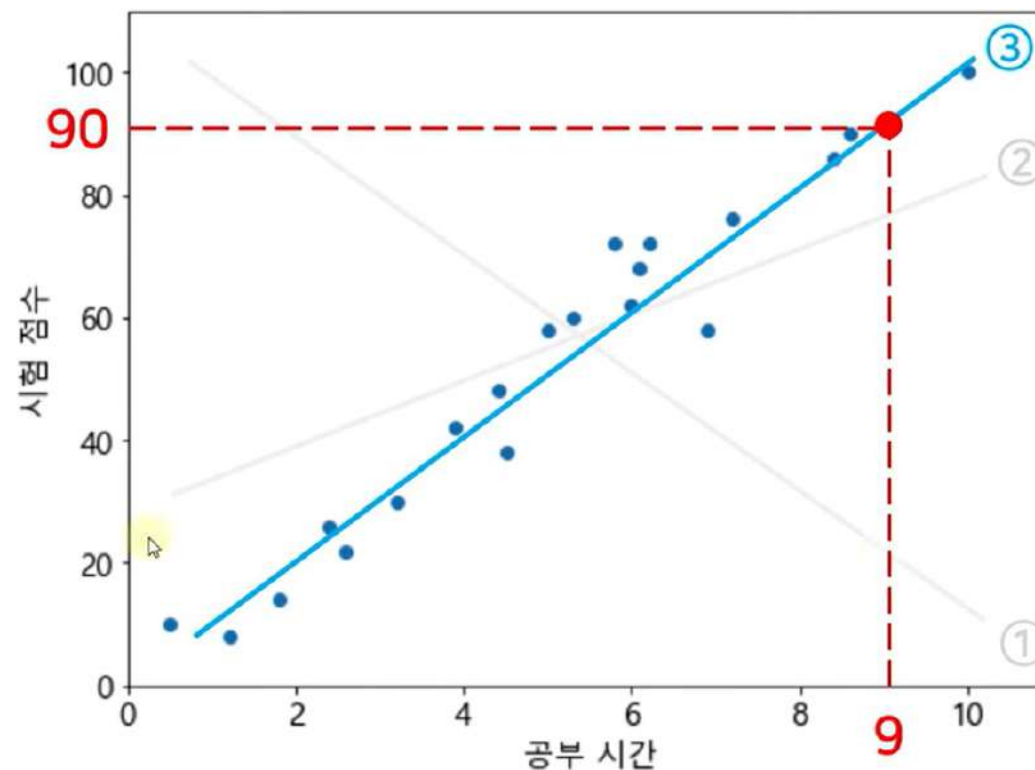
데이터를 가장 잘 표현하는 직선은?



선형회귀(Linear Regression)

공부 시간	시험 점수
0.5	10
1.2	8
1.8	14
2.4	26
2.6	22
3.2	30
3.9	42
4.4	48
4.5	38
5	58
5.3	60
5.8	72
6	62
6.1	68
6.2	72
6.9	58
7.2	76
8.4	86
8.6	90
10	100

9시간을 공부했을 때 예상 시험 점수는?



선형회귀(Linear Regression)

X

Independent variable
독립 변수 (원인)

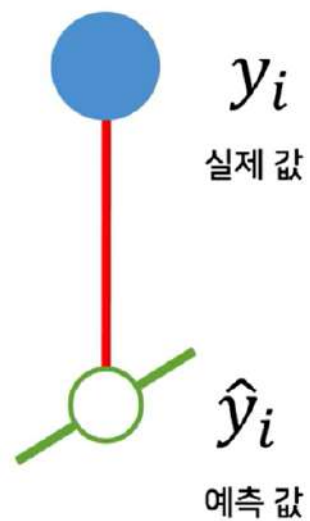
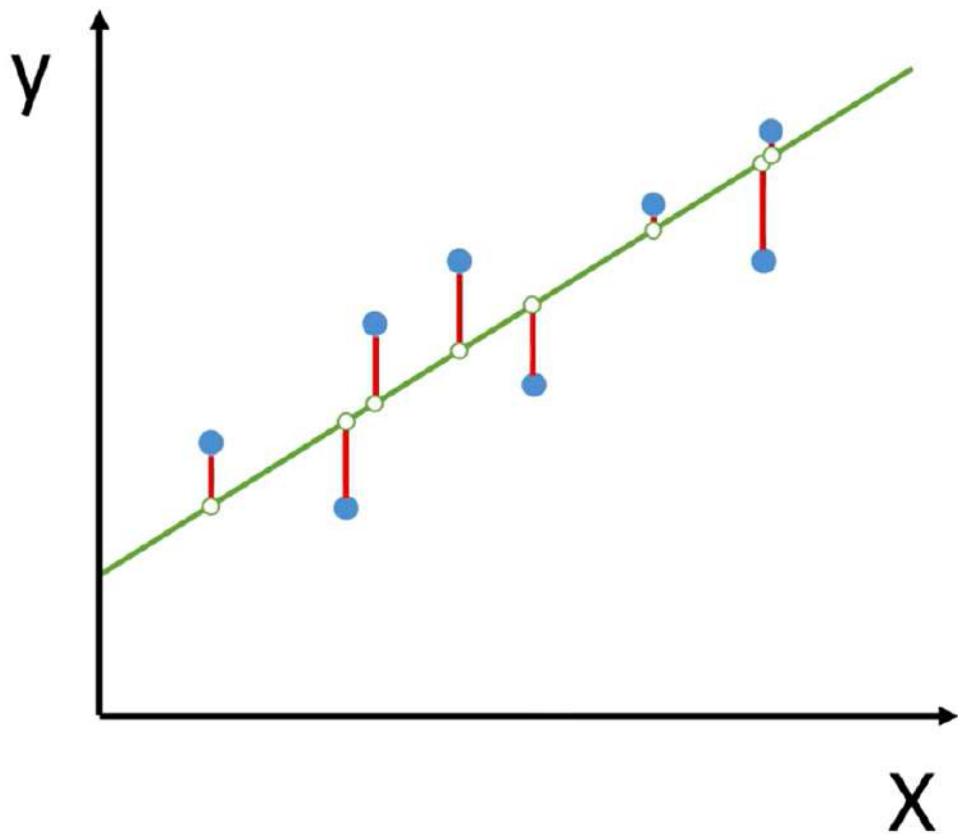
= 입력 변수, feature

y

Dependent variable
종속 변수 (결과)

= 출력 변수, target, label

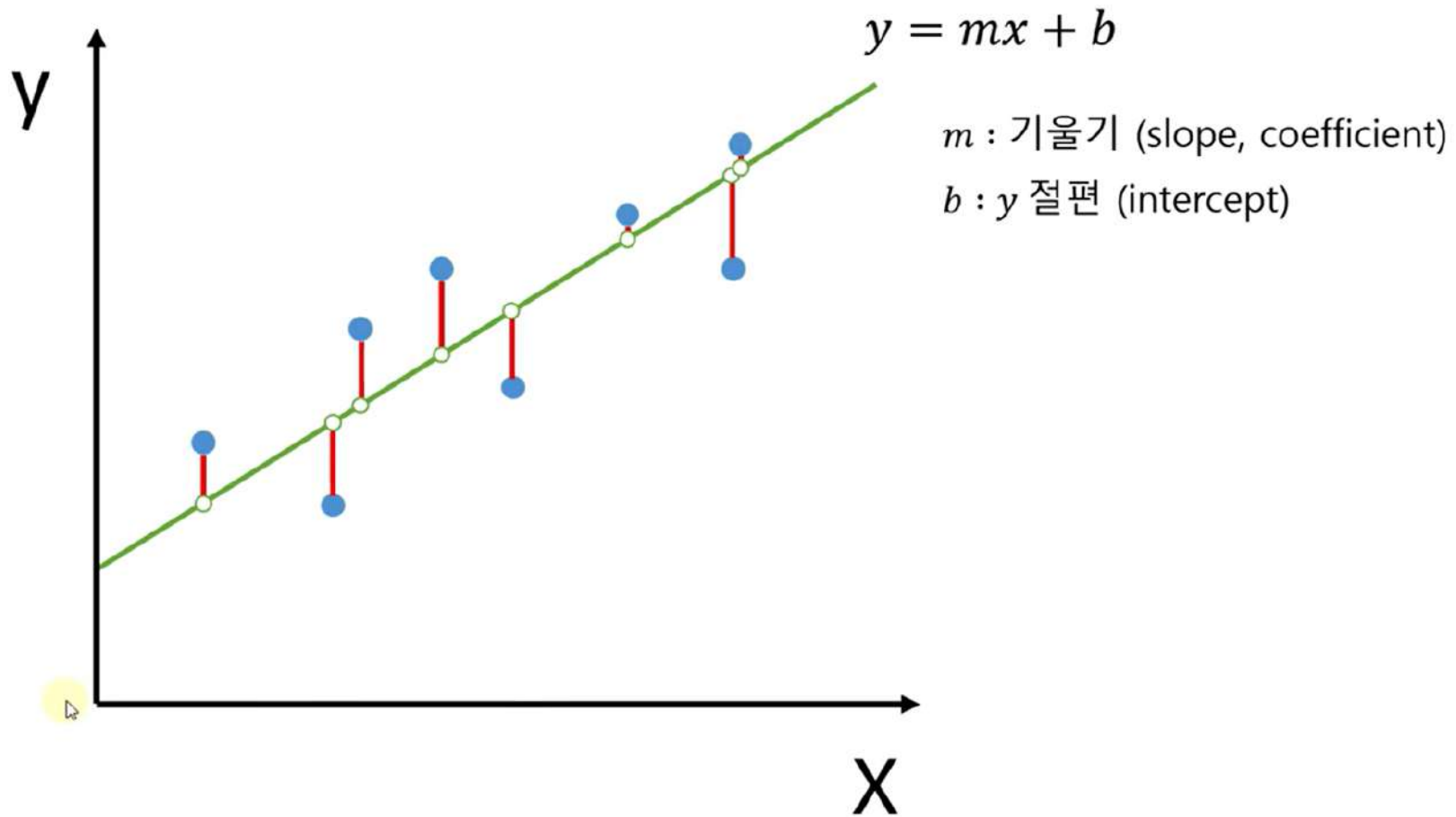
선형회귀(Linear Regression)



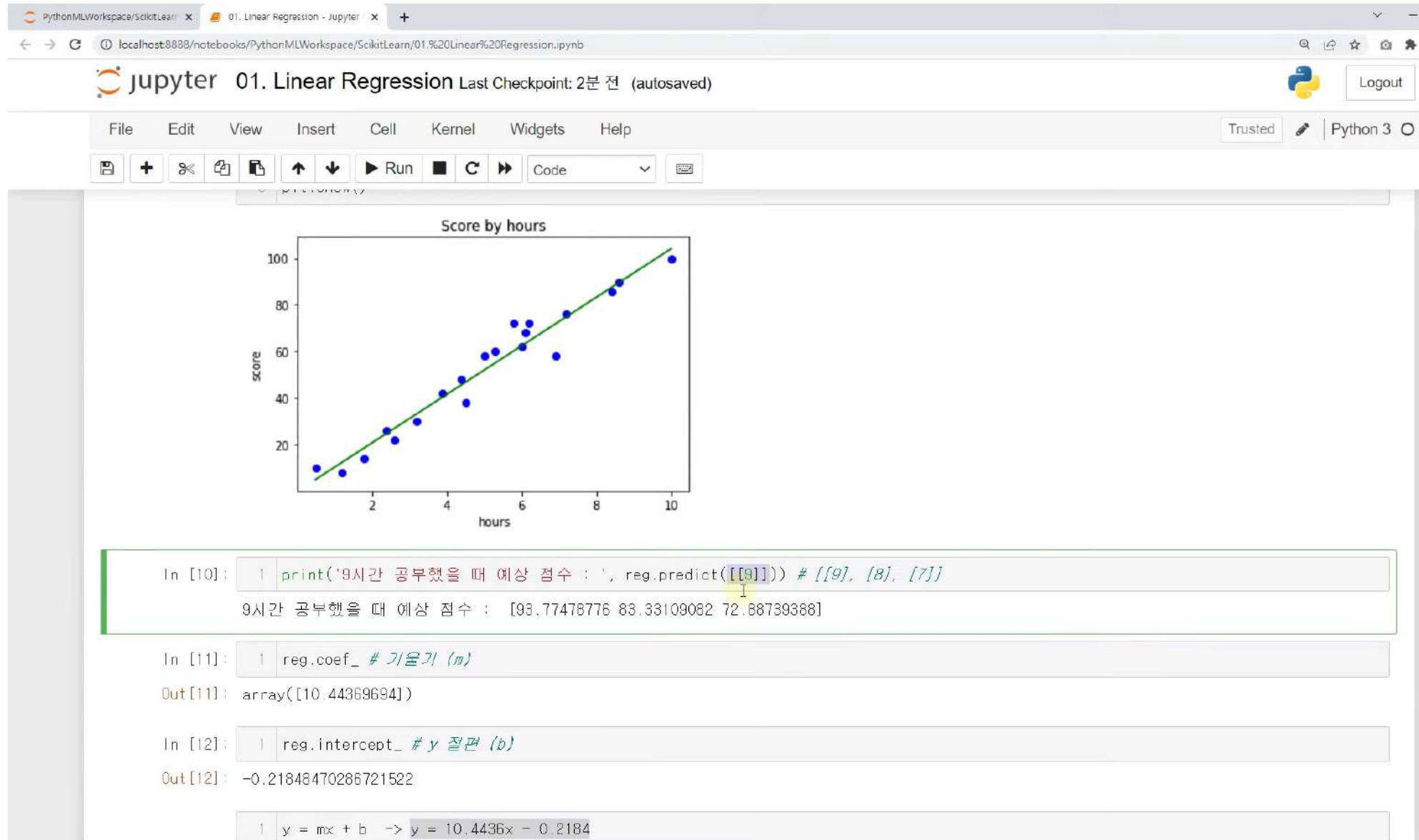
실제 값과 예측 값 차이의 제곱의 합을 최소화

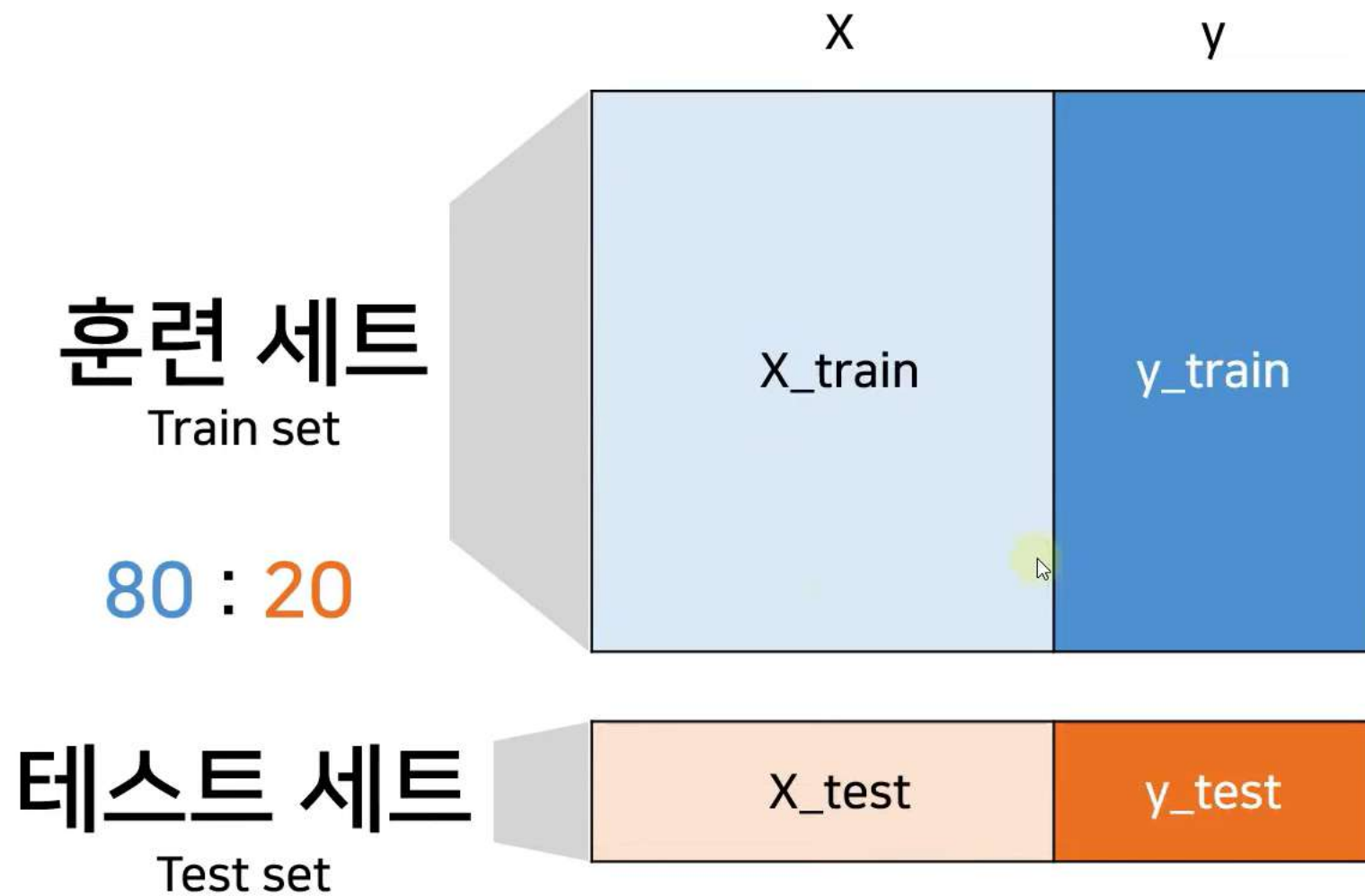
$$\sum (y - \hat{y})^2$$

선형회귀(Linear Regression)



선형회귀(Linear Regression)





선형회귀(Linear Regression)

jupyter 01. Linear Regression Last Checkpoint: 17분 전 (unsaved changes)



Logout

File Edit View Insert Cell Kernel Widgets Help

Notebook saved

Not Trusted

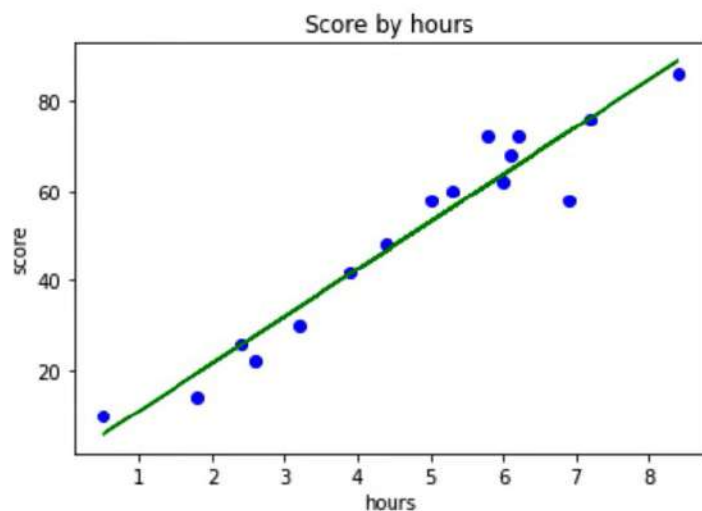


Python 3

Run Code

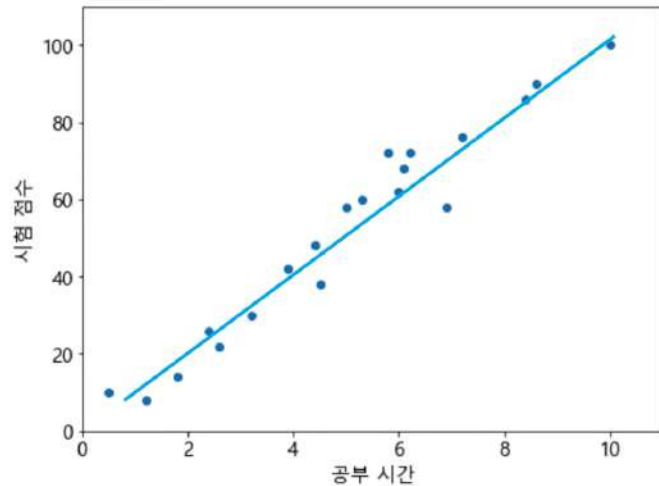
데이터 시각화 (훈련 세트)

```
In [14]: 1 plt.scatter(X_train, y_train, color='blue') # 산점도
          2 plt.plot(X_train, reg.predict(X_train), color='green') # 선 그래프
          3 plt.title('Score by hours (train data)') # 제목
          4 plt.xlabel('hours') # X 축 이름
          5 plt.ylabel('score') # Y 축 이름
          6 plt.show()
```

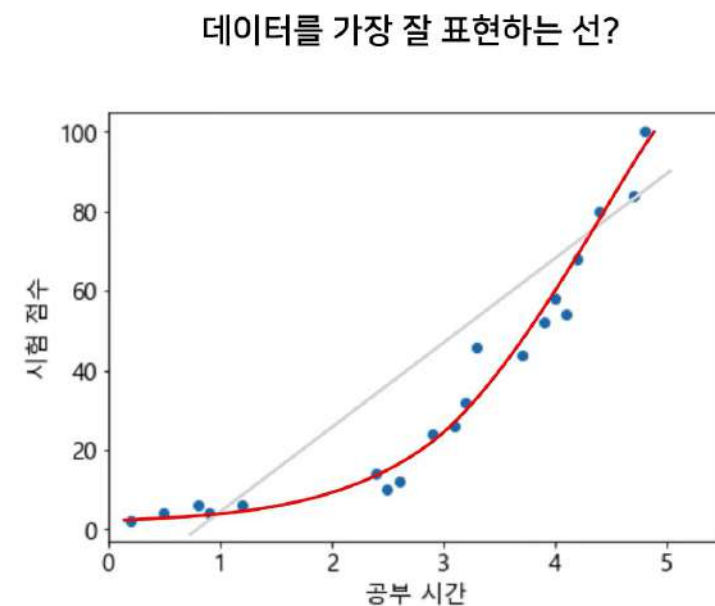


다항회귀(Polynomial Regression)

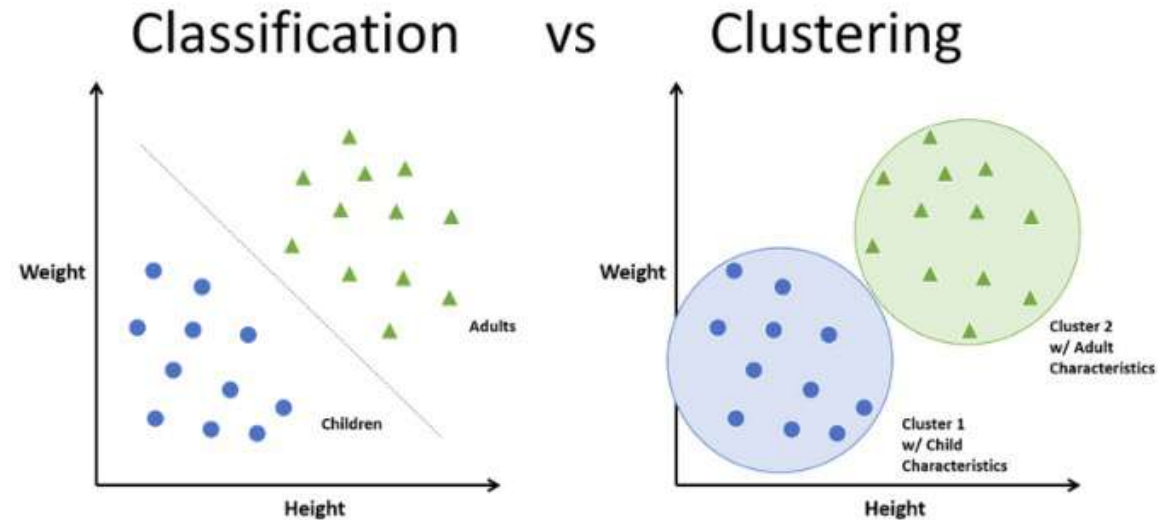
공부 시간	시험 점수
0.5	10
1.2	8
1.8	14
2.4	26
2.6	22
3.2	30
3.9	42
4.4	48
4.5	38
5	58
5.3	60
5.8	72
6	62
6.1	68
6.2	72
6.9	58
7.2	76
8.4	86
8.6	90
10	100



공부 시간	시험 점수
0.2	2
0.5	4
0.8	6
0.9	4
1.2	6
2.4	14
2.5	10
2.6	12
2.9	24
3.1	26
3.2	32
3.3	46
3.7	44
3.9	52
4	58
4.1	54
4.2	68
4.4	80
4.7	84
4.8	100

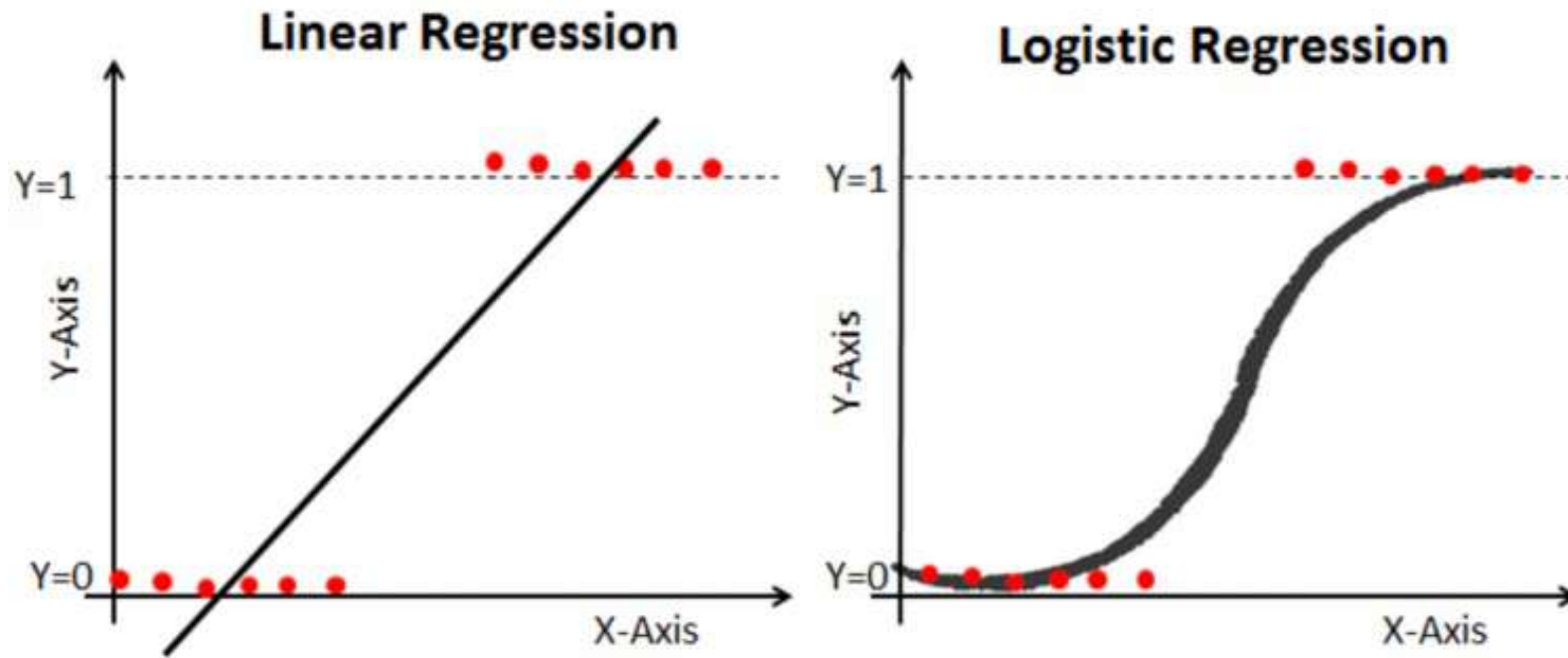


Classification vs Clustering



	Classification	Clustering
Class를 미리 아는가(사전정보)	Yes	No
사용	새로운 샘플/데이터를(이미 알고 있는)Class로 분류	데이터 패턴을 찾은 뒤 Class에 그룹화 제안
알고리즘	Decision Tree, Bayesian, KNN, Random Forest, Naive Bayes	K-Means, Fuzzy, EM, GMMM, 계층분석
데이터 조건	데이터 라벨링되어 있어야 함	예) 사진, 게시물, 비디오
학습	Supervised	Unsupervised
분석 방법	학습(Train) 모델로 데이터 학습	자체 데이터 학습

선형회귀 vs 로지스틱회귀



| Clustering

지도학습		비지도학습
Classification	≠	Clustering
분류		군집화

“유사한 특징을 가지는 데이터들을 그룹화”

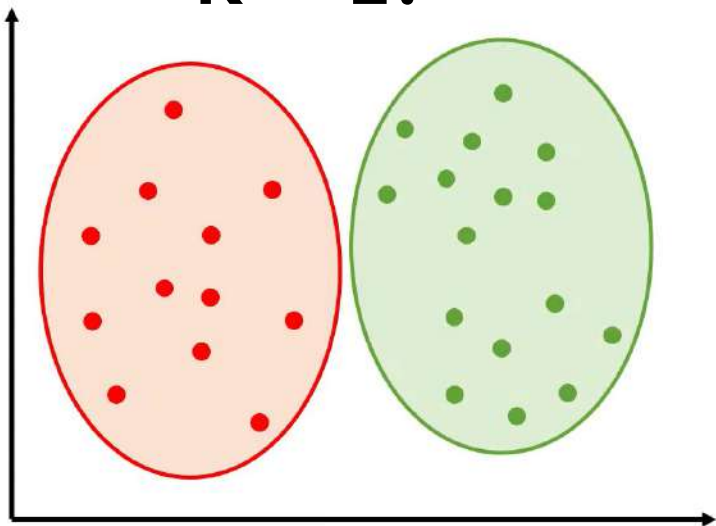
(예) 고객 세분화, 소셜 네트워크 분석, 기사 그룹 분류, ...

K-Means

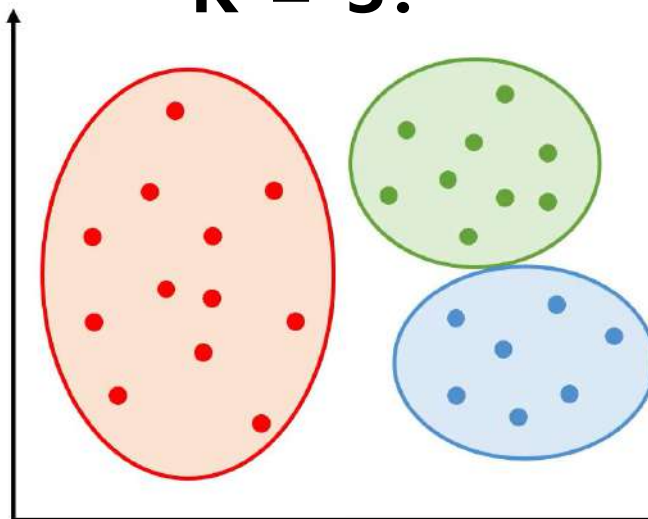
“데이터를 K 개의 클러스터(그룹)로 군집화하는 알고리즘,
각 데이터로부터 이들이 속한 클러스터의 중심점까지의 평균 거리를 계산”

중심점 : Centroid

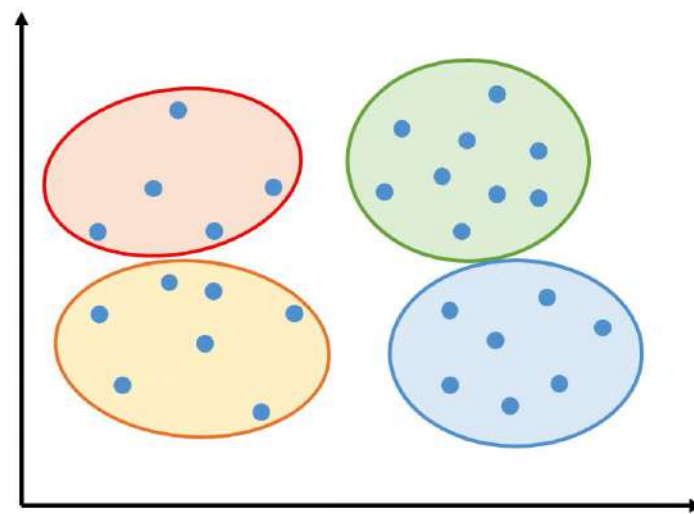
K = 2?



K = 3?

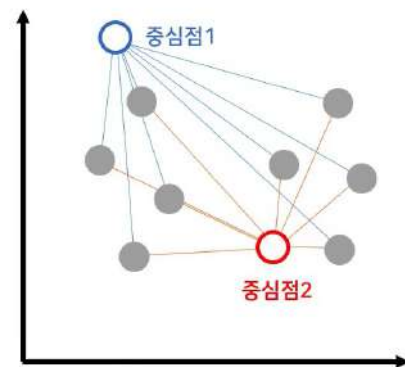


K = 4?

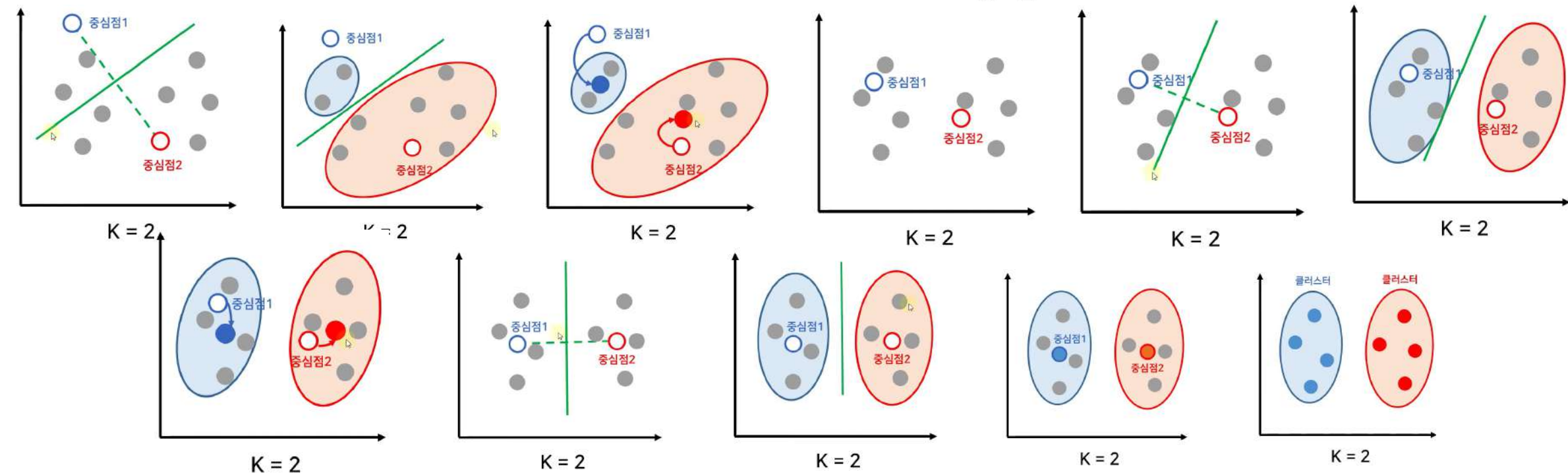


K-Means 동작순서

1. K값 설정
2. 지정된 K개 만큼의 랜덤좌표 설정
3. 모든 데이터로부터 가장 가까운 중심점 선택
4. 데이터들의 평균 중심점으로 중심점 이동
5. 중심점이 더 이상 이동되지 않을 때까지 반복



K = 2

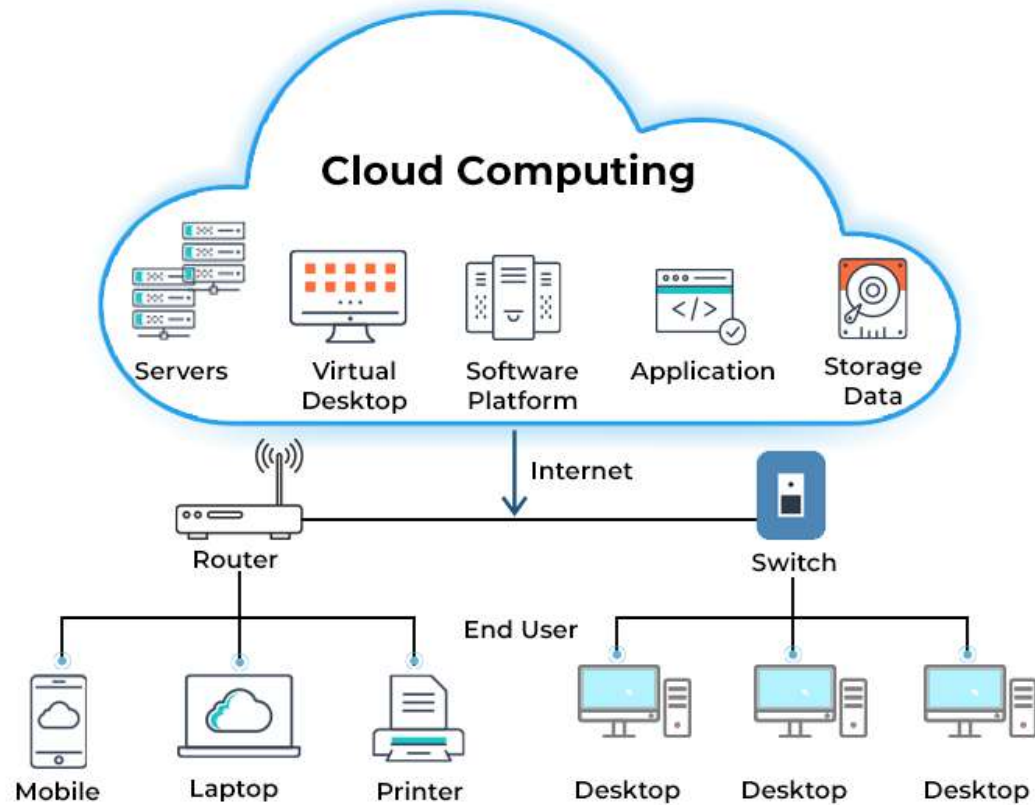


Information Technology



IaaS, PaaS, SaaS and Cloud

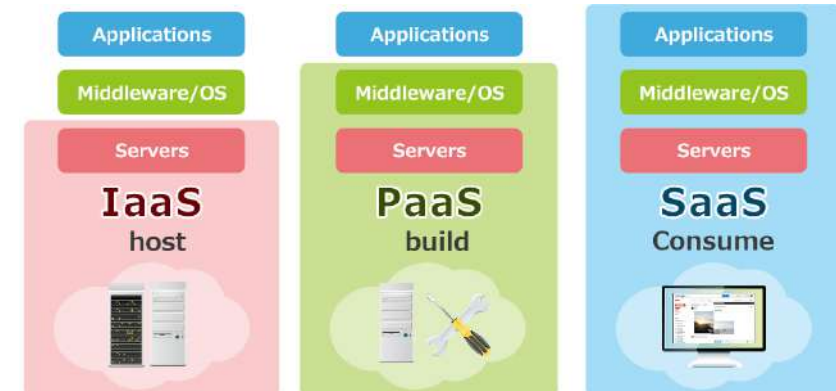
CLOUD COMPUTING ARCHITECTURE








On-site	IaaS	PaaS	SaaS
Applications	Applications	Applications	Applications
Data	Data	Data	Data
Runtime	Runtime	Runtime	Runtime
Middleware	Middleware	Middleware	Middleware
O/S	O/S	O/S	O/S
Virtualization	Virtualization	Virtualization	Virtualization
Servers	Servers	Servers	Servers
Storage	Storage	Storage	Storage
Networking	Networking	Networking	Networking

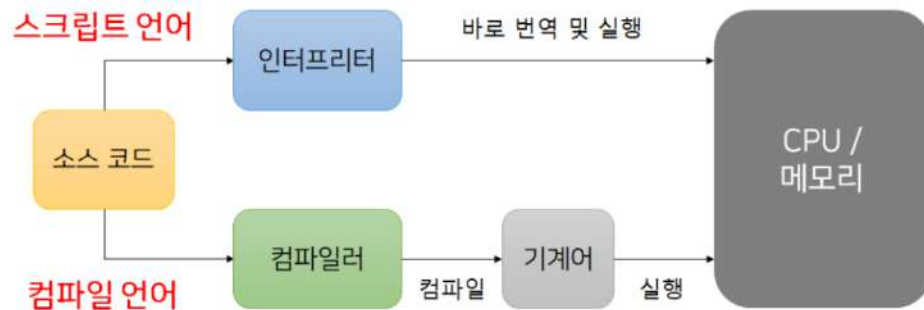
■ You manage

■ Service provider manages



Compiler vs Interpreter

	Interpreter	Compiler
종류	 	  
작동 방식	소스코드를 실행 시 마다 해석	소스 코드를 한번에 기계어로 변환
실행속도	느림	빠름
보안	낮음	높음
메모리 사용량	큼	적음
디버깅	비교적 용이	어려움
활용	웹개발, 프로토타입 제작 등	운영체제, 펌웨어, 게임엔진, 그래픽처리 등 성능이 중요한 애플리케이션 또는 시스템 프로그래밍



Interactive Mode vs Script Mode



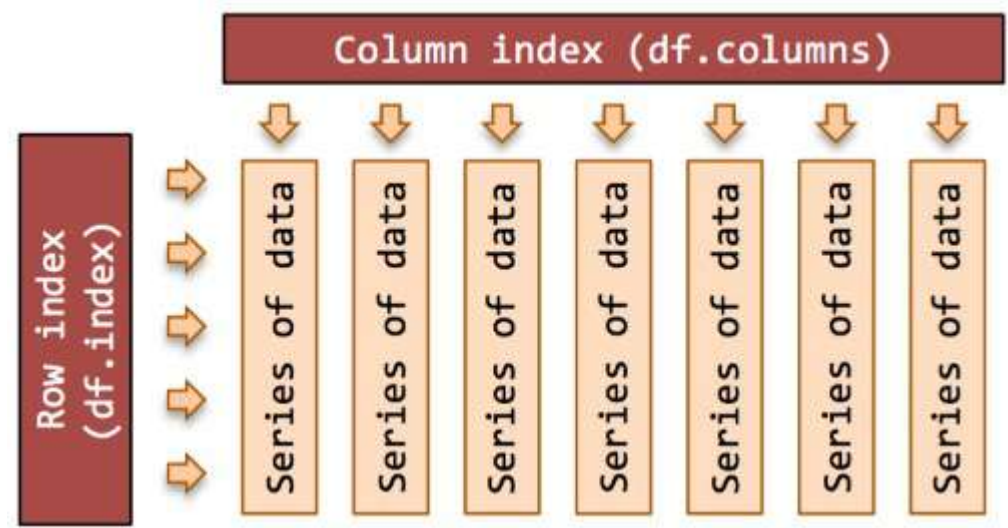
Interactive Mode	Script Mode
 Notebook	 Visual Studio Code Microsoft
<pre>In [1]: import pandas as pd</pre> <pre>In [2]: import matplotlib.pyplot as plt</pre>	>>> Editor Screen
한줄 한줄 바로 결과를 표시	전체 코드 작성 후 Run

통계 기초

■ Pandas

```
1 df1 = pd.DataFrame([[1, 2], [3, 4]], columns=["A", "B"], index=["x", "y"])
2 display(df1)
```

	A	B
x	1	2
y	3	4



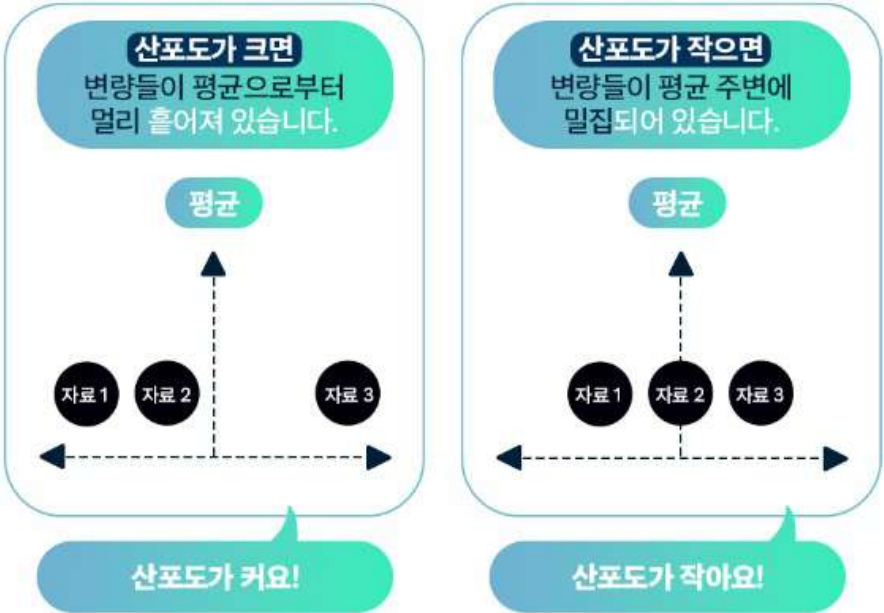
```
1 s3 = pd.Series({"A":1, "B":2, "C":3, "D":4})
2 display(s3)
```

```
A    1
B    2
C    3
D    4
dtype: int64
```

통계 기초

산포도

얼마나 많이 흩어져 있는지에 대한 정도를 하나의 수로 나타낸 값



분산

편차를 제공한 값의 평균

$$(분산) = \frac{(편차)^2 \text{의 총합}}{(변량) \text{의 개수}}$$

+ PLUS

편차 $\rightarrow (변량) - (평균)$

- 편차의 총합은 항상 0입니다.
- 편차의 절댓값이 클수록 그 변량은 평균에서 멀리 떨어져 있습니다.
- 편차의 절댓값이 작을수록 평균에 가까이 있습니다.

평균을 중심으로 흩어진 정도

▣ 통계 기초

표준편차

분산의 양의 제곱근

(표준편차) = $\sqrt{\text{분산}}$



표준편차를 구하는 순서

평균 ▶ 편차 ▶ 분산 ▶ 표준편차

- (편차) = (변량) - (평균)
- (분산) = (편차)² 의 평균
- (표준편차) = $\sqrt{\text{분산}}$

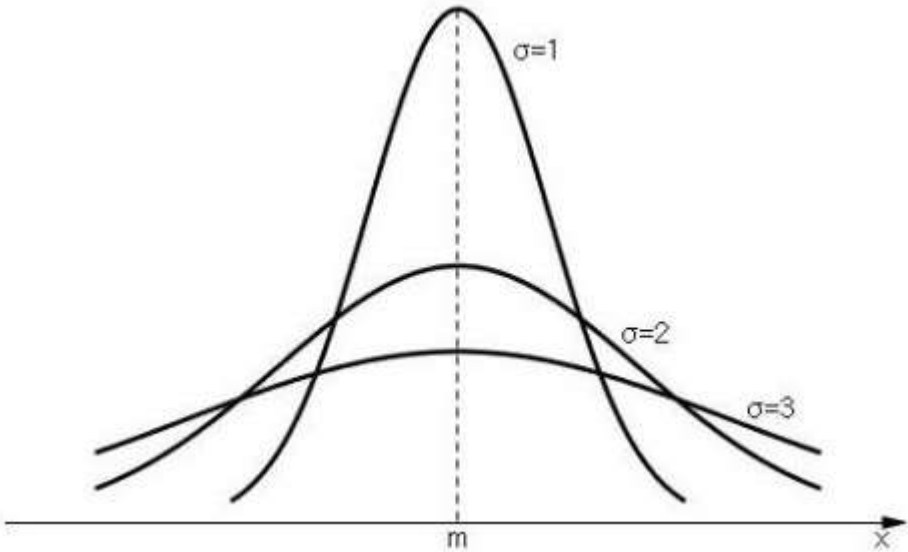
■ 통계 기초

변량	자료1 변량	(평균) 편차
		변량 - 평균
1	156	-4
2	157	-3
3	159	-1
4	163	3
5	165	5
총합	800	0
평균	$\frac{800}{5} = 160$	
		산포도 사용불가

	자료1 변량	(평균) 편차	편차 이용한 산포도	
		변량 - 평균	평균편차 변량 - 160	중앙값편차 변량 - 159
1	156	-4	4	3
2	157	-3	3	2
3	159	-1	1	0
4	163	3	3	4
5	165	5	5	6
총합	800	0	16	15
평균	$\frac{800}{5} = 160$			
		산포도 사용불가	산포도 사용불가	중앙값을 이용한 산포도 교육과정 밖

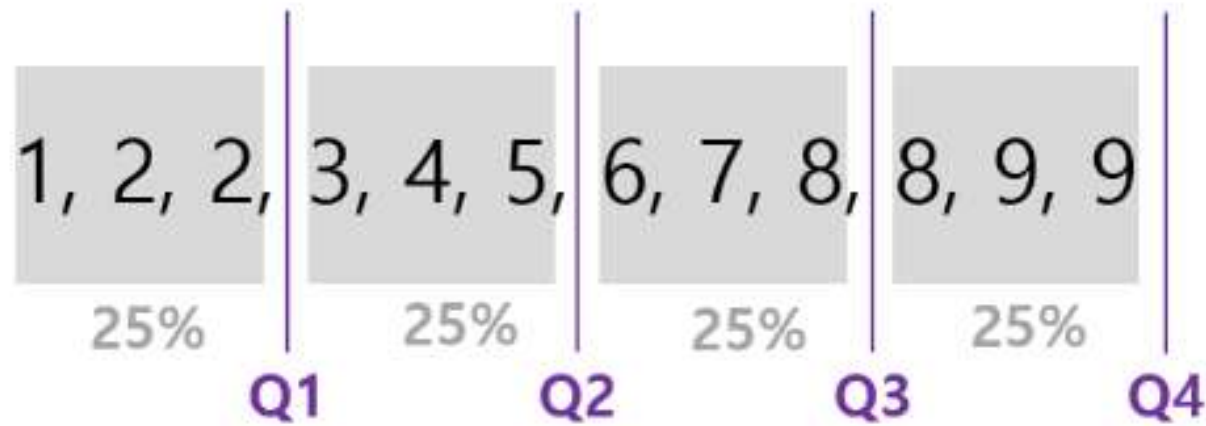
통계 기초

	자료1	(편차) ² 이용		자료2	(편차) ² 이용
		(변량-평균) ²			(변량-평균) ²
1	156	16	1	156	16
2	157	9	2	157	9
3	159	1	3	162	4
4	163	9	4	165	25
5	165	25			
총합	800	60	총합	640	54
평균	$\frac{800}{5} = 160$	분산 $\frac{60}{5} = 12$	평균	$\frac{640}{4} = 160$	분산 $\frac{54}{4} = 13.5$
		표준편차 $2\sqrt{3}$			표준편차 $\frac{3\sqrt{6}}{2}$



▣ 통계 기초

7, 2, 8, 3, 1, 9, 5, 2, 9, 6, 8, 4



InnerQuatile Range

$IQR = Q3 - Q1$

- $Q1 = (2+3) / 2 = 2.5$
- $Q2 = (5+6) / 2 = 5.5$
- $Q3 = (8+8) / 2 = 8$
- $Q4 = 9$

▣ 통계 기초

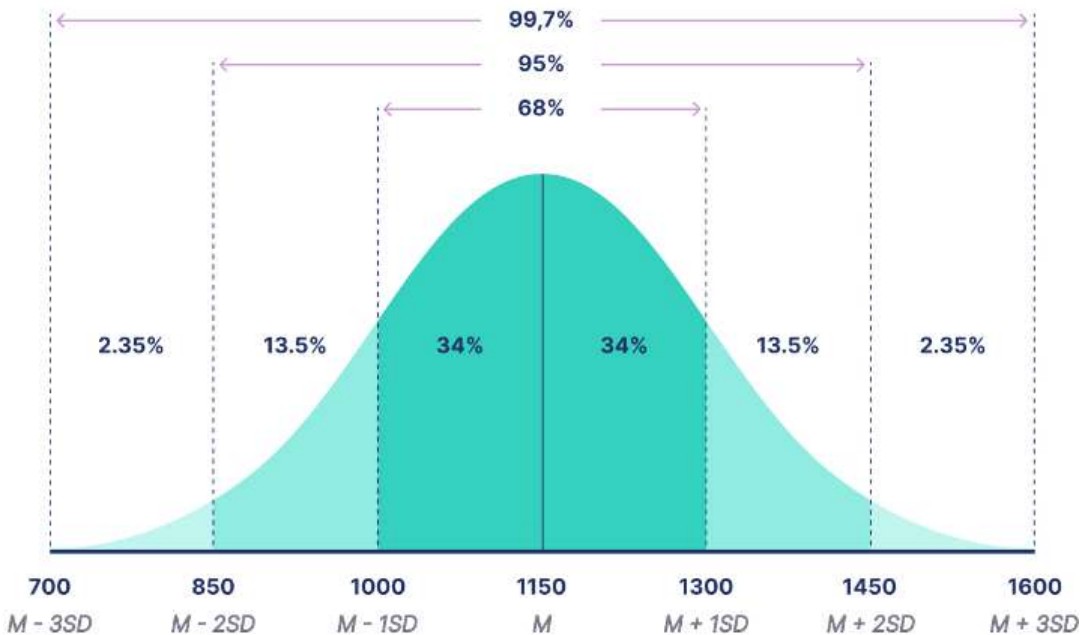
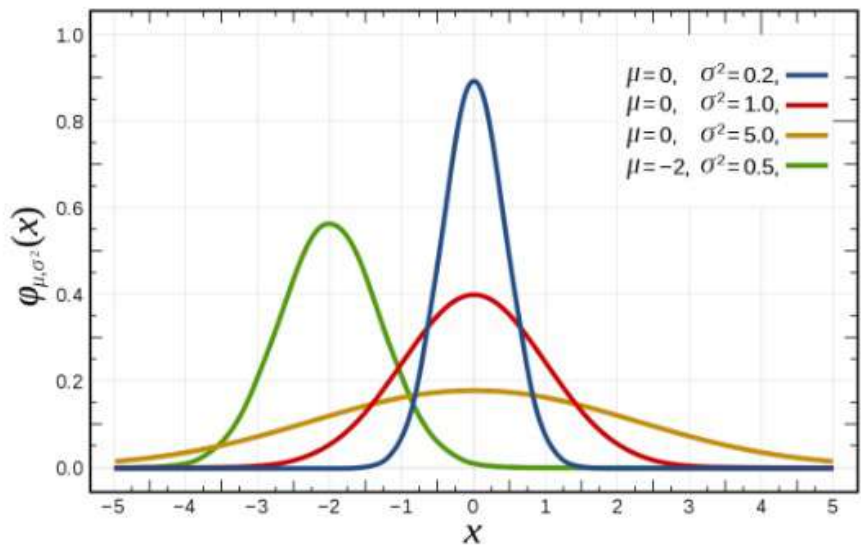


사분위수 범위(Interquartile range) = $Q3 - Q1 = 9 - 4 = 5$

통계 기초

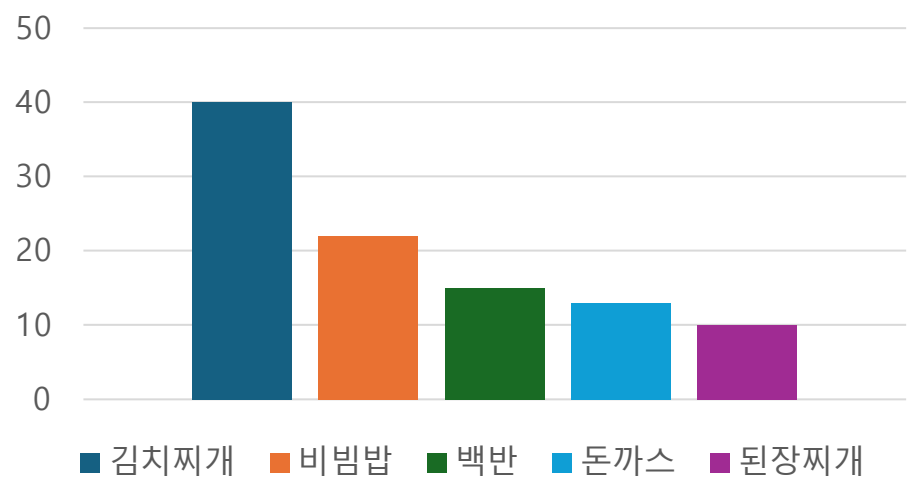
실험군 (Group=1)		
환자	그룹	점수
1	1	10
2	1	16
3	1	27
4	1	15
5	1	21
6	1	14
7	1	16
8	1	21
9	1	22
10	1	23
11	1	25
12	1	28
13	1	27
14	1	13
15	1	15
16	1	16
17	1	21
18	1	22
19	1	25
20	1	28

대조군 (Group=2)		
환자	그룹	점수
21	2	23
22	2	26
23	2	27
24	2	23
25	2	16
26	2	18
27	2	31
28	2	33
29	2	28
30	2	36
31	2	18
32	2	21
33	2	26
34	2	28
35	2	29
36	2	33
37	2	32
38	2	16
39	2	18
40	2	23



통계 기초

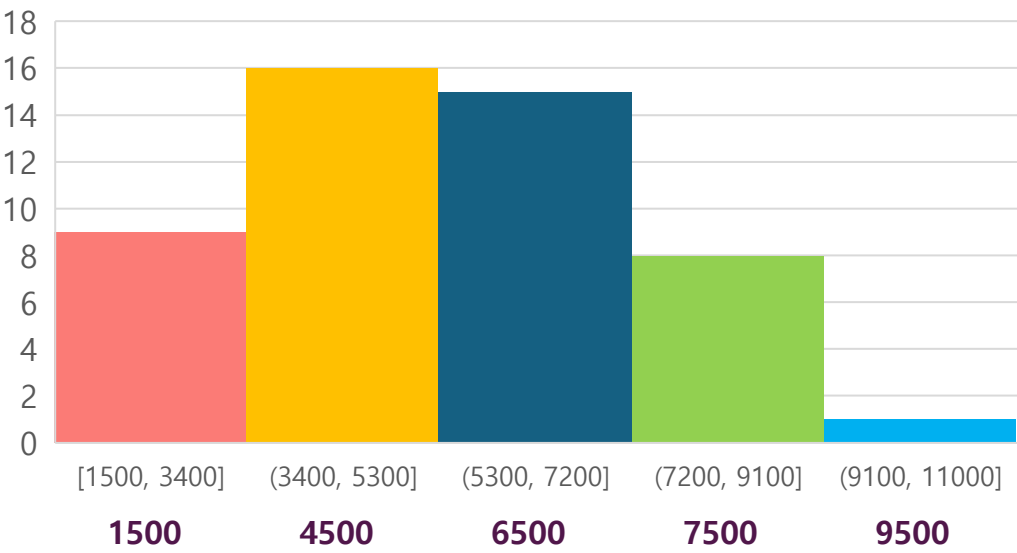
막대그래프



[막대그래프]

범주(category)로 구분되는 데이터
막대로 표현하려는 범부의 순서는 의도에 따라 바뀔수 있으며
막대간 일정한 간격을 유지

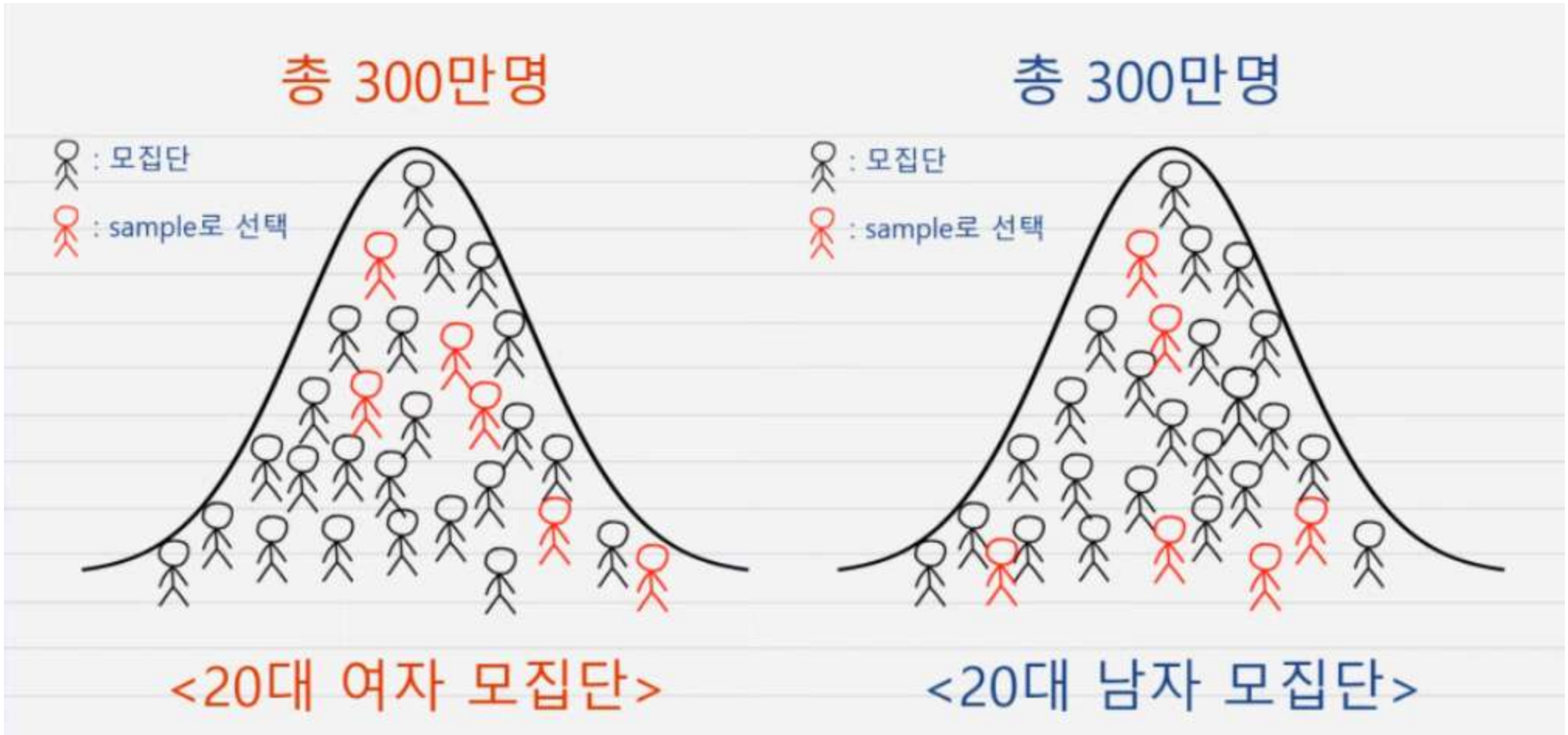
히스토그램



[히스토그램]

범측정된 연속적인 값(몸무게, 성적등) 으로 표시되는 데이터를 표현
막대의 순서를 임의적으로 바꿀수 없으며 막대간의 간격없이 표현

통계 기초





감사합니다.