

빅데이터 특강

신지은 (서울시립대학교 통계학과)



목차

#01 데이터의 이해

#02 정형 데이터를 위한 분석 방법 및 실습

#03 비정형 데이터를 위한 분석방법 및 실습

- 실습자료 다운로드: <https://github.com/jiieunshin/high-univ>
- 문의메일: jiieunstat@gmail.com

1. 데이터의 이해

데이터의 분류

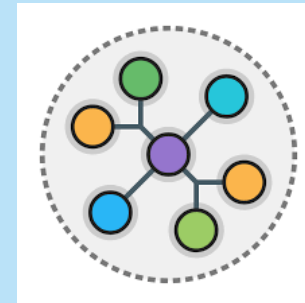
데이터

정형 데이터

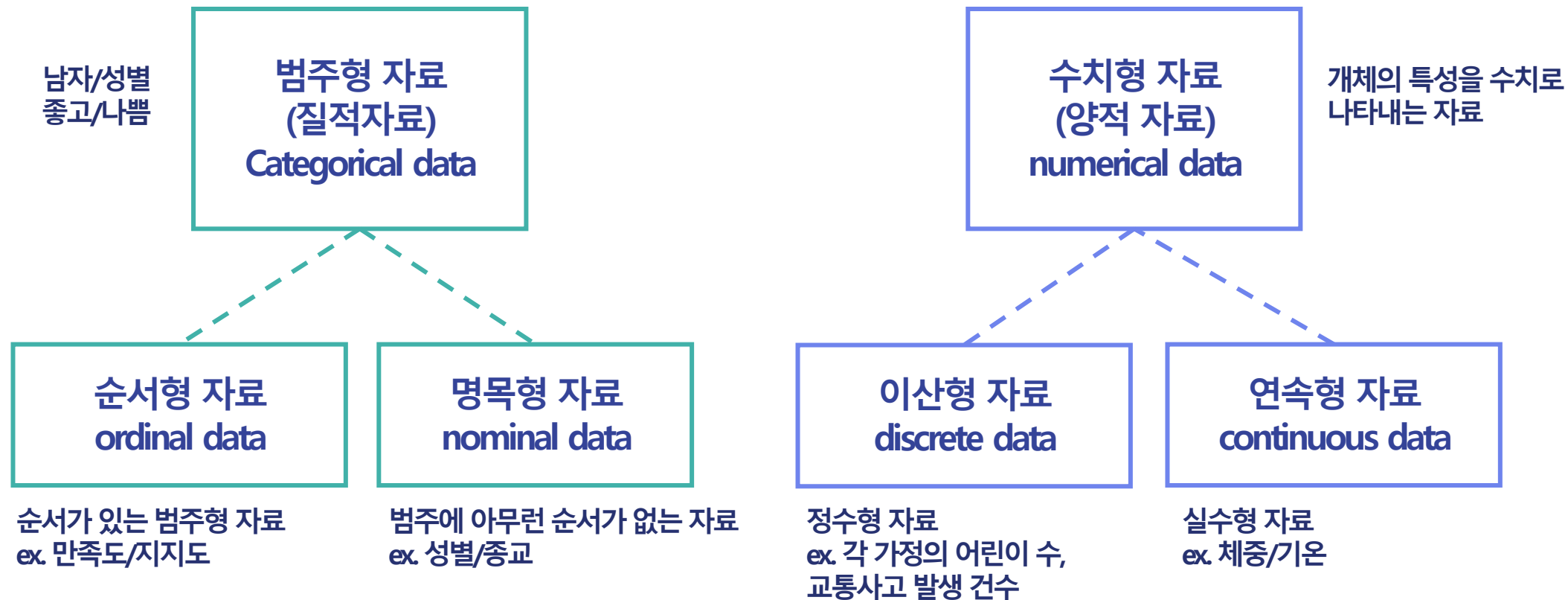
id	이름	나이	성별
01	Kim	32	M
02	Lee	26	F
03	Park	72	F
04	Choi	15	M



비정형 데이터



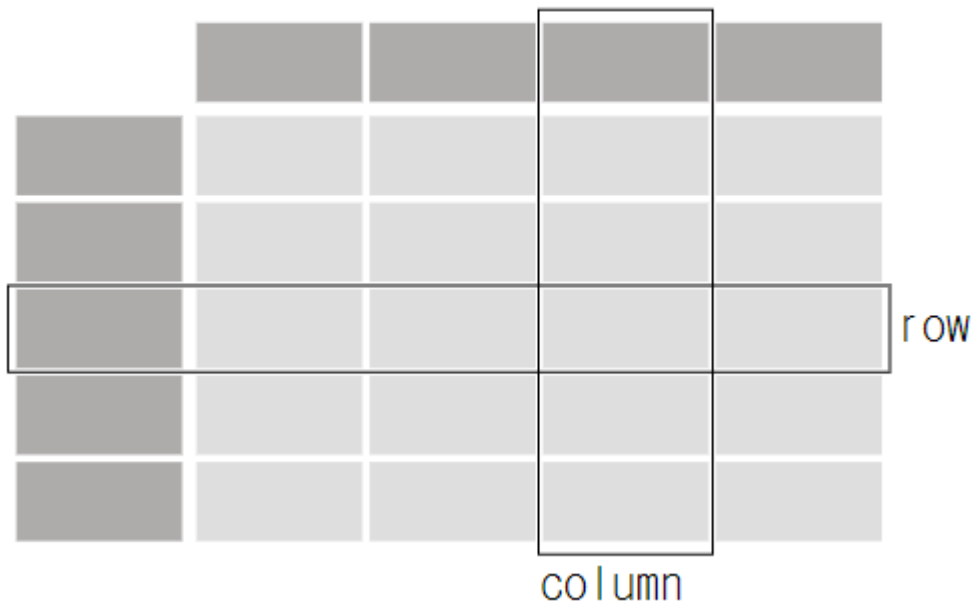
정형 데이터의 분류



1. 데이터의 이해

파이썬 데이터프레임

DataFrame



화재발생 데이터.csv

	화재발생연도	시군구	사망자수	부상자수	재산피해금액	출동횟수	출동횟수_겨울	출동횟수_여름
0	2017	은평구	0.0	3	218200	159	51	32
1	2017	종로구	1.0	3	1077665	234	55	69
2	2017	중구	5.0	14	485392	198	48	47
3	2017	중랑구	2.0	5	332366	196	53	38
4	2018	은평구	5.0	10	419503	214	58	47
5	2018	종로구	14.0	22	574300	254	71	70
6	2018	중구	0.0	23	1257005	275	76	74
7	2018	중랑구	2.0	8	201421	254	72	55
8	2019	은평구	3.0	20	2412769	196	62	34
9	2019	종로구	4.0	16	801094	232	60	63
10	2019	중구	3.0	17	74077097	213	51	39
11	2019	중랑구	1.0	9	322650	210	54	49
12	2020	은평구	2.0	6	504788	192	48	46
13	2020	종로구	2.0	5	639751	217	50	49
14	2020	중구	0.0	10	1284422	185	41	54
15	2020	중랑구	2.0	12	229566	225	54	57
16	2021	은평구	3.0	8	875722	160	57	42
17	2021	종로구	0.0	12	465499	192	48	54

1. 데이터의 이해

정형 데이터의 집계 방법: 기초 통계량

	화재발생연도	시군구	사망자수	부상자수	재산피해금액	출동횟수	출동횟수_겨울	출동횟수_여름
0	2017	은평구	0.0	3	218200	159	51	32
1	2017	종로구	1.0	3	1077665	234	55	69
2	2017	중구	5.0	14	485392	198	48	47
3	2017	중랑구	2.0	5	332366	196	53	38
4	2018	은평구	5.0	10	419503	214	58	47
5	2018	종로구	14.0	22	574300	254	71	70
6	2018	중구	0.0	23	1257005	275	76	74
7	2018	중랑구	2.0	8	201421	254	72	55
8	2019	은평구	3.0	20	2412769	196	62	34
9	2019	종로구	4.0	16	801094	232	60	63
10	2019	중구	3.0	17	74077097	213	51	39
11	2019	중랑구	1.0	9	322650	210	54	49
12	2020	은평구	2.0	6	504788	192	48	46
13	2020	종로구	2.0	5	639751	217	50	49
14	2020	중구	0.0	10	1284422	185	41	54
15	2020	중랑구	2.0	12	229566	225	54	57
16	2021	은평구	3.0	8	875722	160	57	42
17	2021	종로구	0.0	12	465499	192	48	54

셈 척도 갯수 (count), 합산 (sum)

중심척도 평균 (mean), 중위수 (median)

산포척도 최댓값 (max), 최솟값 (min), 분산 (variance), 표준편차 (standard deviation), 백분위 (quantile)

범주형 자료

(셈 척도) count, percent

수치형 자료

(셈 척도) sum
중심척도 모두
산포척도 모두

화재발생연도, 시군구

사망자수, 부상자수,
재산피해금액, 출동횟수

1. 데이터의 이해

집계를 위한 문제와 설계

	화재발생연도	시군구	사망자수	부상자수	재산피해금액	출동횟수	출동횟수_겨울	출동횟수_여름
0	2017	은평구	0.0	3	218200	159	51	32
1	2017	종로구	1.0	3	1077665	234	55	69
2	2017	중구	5.0	14	485392	198	48	47
3	2017	중랑구	2.0	5	332366	196	53	38
4	2018	은평구	5.0	10	419503	214	58	47
5	2018	종로구	14.0	22	574300	254	71	70
6	2018	중구	0.0	23	1257005	275	76	74
7	2018	중랑구	2.0	8	201421	254	72	55
8	2019	은평구	3.0	20	2412769	196	62	34
9	2019	종로구	4.0	16	801094	232	60	63
10	2019	중구	3.0	17	74077097	213	51	39
11	2019	중랑구	1.0	9	322650	210	54	49
12	2020	은평구	2.0	6	504788	192	48	46
13	2020	종로구	2.0	5	639751	217	50	49
14	2020	중구	0.0	10	1284422	185	41	54
15	2020	중랑구	2.0	12	229566	225	54	57
16	2021	은평구	3.0	8	875722	160	57	42
17	2021	종로구	0.0	12	465499	192	48	54

문제

시군구별 평균 재산피해금액과 총 출동횟수

설계

➤ 그룹화: 시군구

➤ 계산하고 싶은 열: 재산피해금액, 출동횟수

➤ 집계함수: sum, mean

파이썬 구현 예시

```
df5.groupby(['시군구']).agg({"재산피해금액" : "mean", "출동횟수" : "sum"})
```

시군구	재산피해금액	출동횟수
-----	--------	------

은평구	886196.4	921
-----	----------	-----

종로구	711661.8	1129
-----	----------	------

중구	15976858.0	1042
----	------------	------

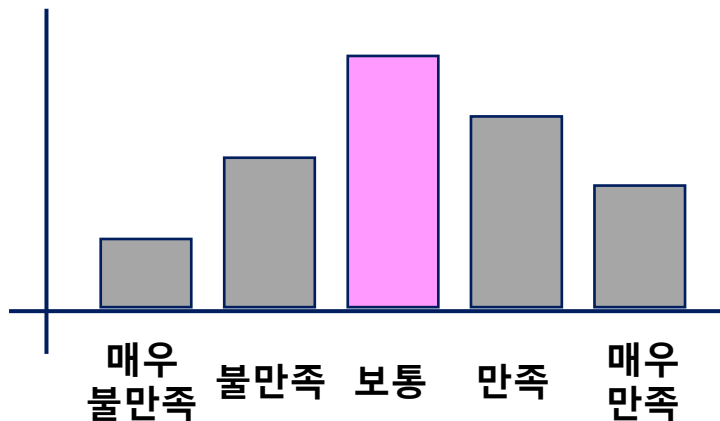
중랑구	286252.0	1098
-----	----------	------

1. 데이터의 이해

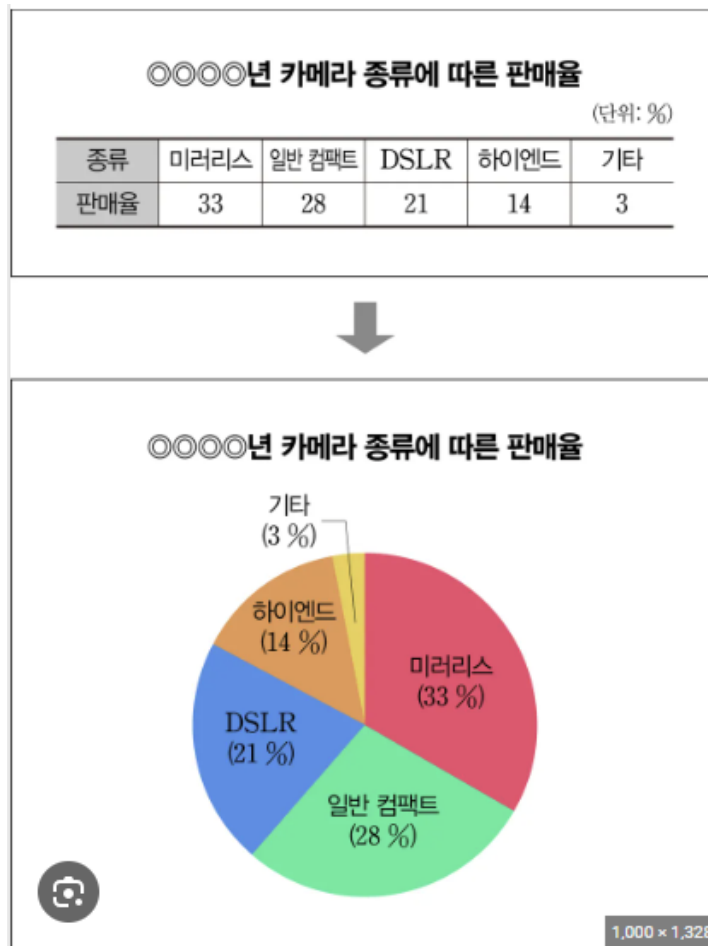
정형 데이터의 시각화

범주형 자료

막대그래프



원 그래프



분할표

교육수준	결혼생활		
	빈약	원만	대단히 양호
대학	72	112	245
고등학교	65	90	120
중학교	95	103	98

[표] 교육수준과 결혼생활

1. 데이터의 이해

정형 데이터의 시각화

범주형 자료



1. 데이터의 이해

정형 데이터의 시각화

수치형 자료

➤ 도수분포표와 히스토그램

아래의 수학 점수를 도수분포표로 나타내보자

Female	Male
7, 59, 78, 79, 60, 65, 68, 71, 75, 48, 51, 55, 56, 41, 43, 44, 75, 78, 80, 81, 83, 83, 85	48, 49, 49, 30, 30, 31, 32, 35, 37, 41, 86, 42, 51, 53, 56, 42, 44, 50, 51, 65, 67, 51, 56, 58, 64, 64, 75



1. 관측치의 최댓값과 최솟값의 차이, 즉 범위를 구한다.
 $\Rightarrow 85 - 7 = 78$
2. 구간을 몇 개로 나눌 것인가?
 $\Rightarrow 10$ 개
3. 구간 폭을 정하자
 $\Rightarrow \text{구간 폭} = (\text{최댓값} - \text{최솟값}) / \text{구간수} = 78 / 10 = 7.8$
4. 도수와 상대도수, 누적도수, 누적상대도수 등을 산출한다.

1. 데이터의 이해

정형 데이터의 시각화

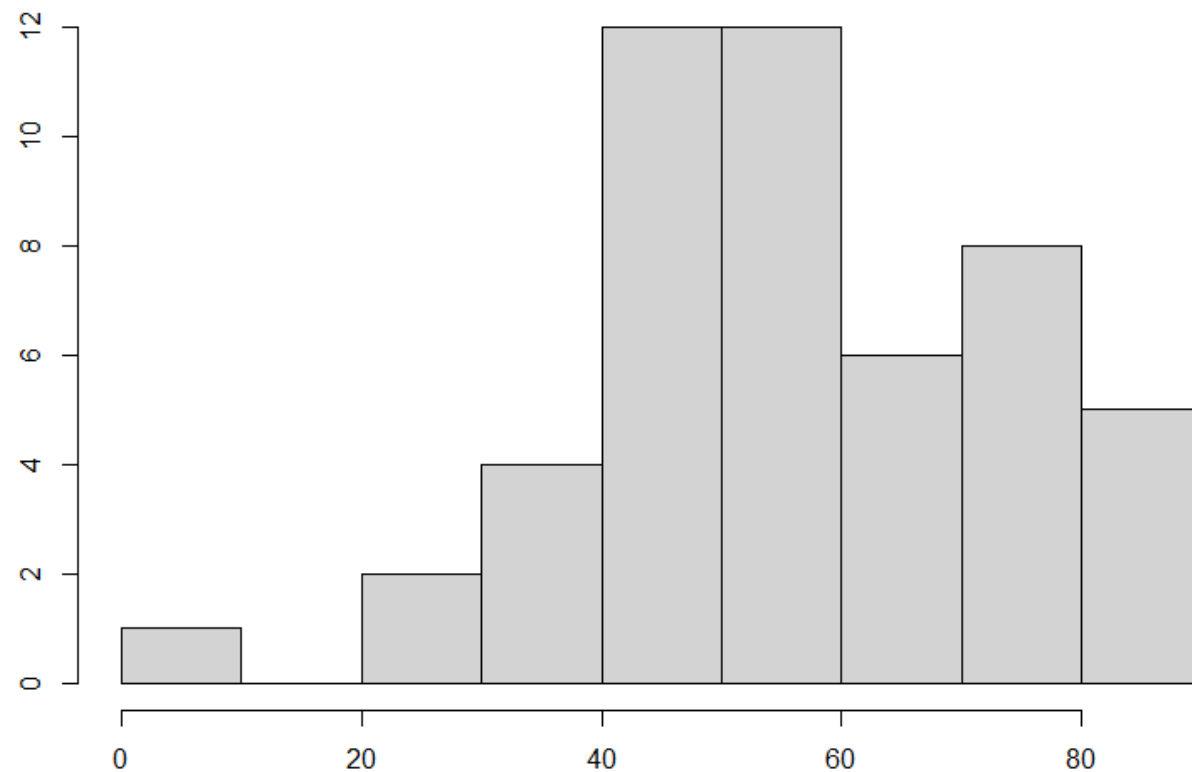
수치형 자료

➤ 도수분포표와 히스토그램

점수	학생 수 (명)
(0, 10]	1
(10, 20]	0
(20, 30]	2
(30, 40]	4
(40, 50]	12
(50, 60]	12
(60, 70]	7
(70, 80]	9
(80, 90]	5
(90, 100]	0
계	50

도수분포표

구간: 10개,
구간 폭: 10



히스토그램

1. 데이터의 이해

정형 데이터의 시각화

수치형 자료

➤ 평균과 중앙값

두 가지 자료 (0, 1, 2, 2, 2, 3, 4)와 (70, 1, 2, 2, 2, 3, 4)의 평균, 중앙값을 비교해보자

0, 1, 2, 2, 2, 3, 4 → 평균 2, 중앙값 2
70, 1, 2, 2, 2, 3, 4 → 평균 12, 중앙값 2

➤ 백분위수와 상자그림

	first quantile	second quantile	third quantile			
first quarter		second quarter		fourth quarter		
24, 25, 26,		27, 30, 32,		40, 44, 50,		52, 55, 57
	Q1	Q2	Q3			
	26/2	36	51			

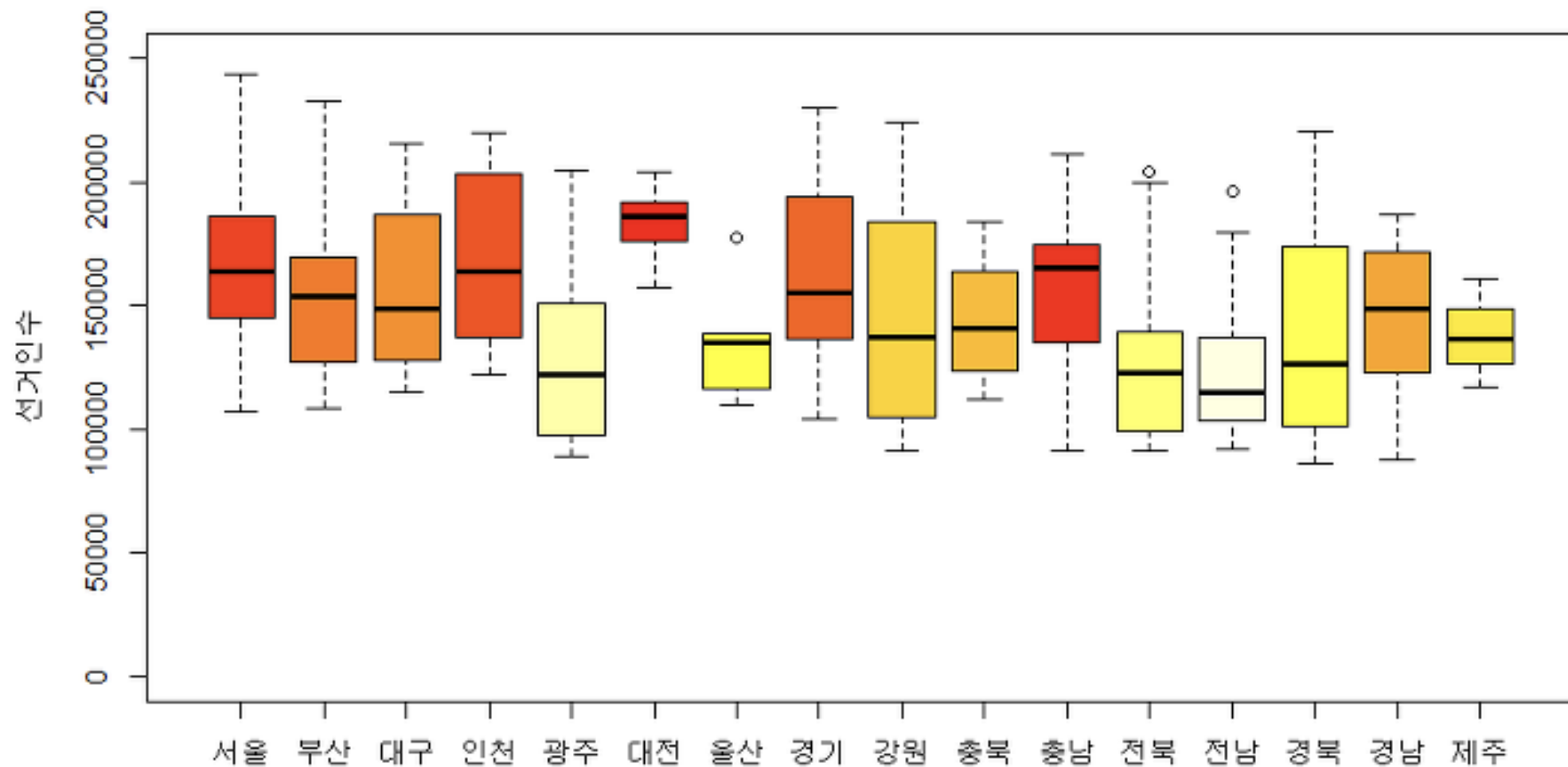


1. 데이터의 이해

정형 데이터의 시각화

수치형 자료

우리나라 18대 국회의원 선거구의 선거인수 분포



정형 데이터의 시각화

수치형 자료

➤ 분산과 표준편차

- 분산 (variance)

: 각 자료값들과 평균과의 차이 $x_i - \bar{x}$ 로 산포를 나타낸다. 즉, 평균으로부터 멀리 떨어져 있을수록 $x_i - \bar{x}$ 의 절댓값이 커짐.

표본분산 s^2 은 다음과 같은 식으로 구한다.

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

- 표준편차 (s.d., standard deviation)

: 분산의 제곱근. 분산을 구할 때 제곱을 취함으로써 원래 자료값의 단위가 달라진 것을 복구한 것이다.

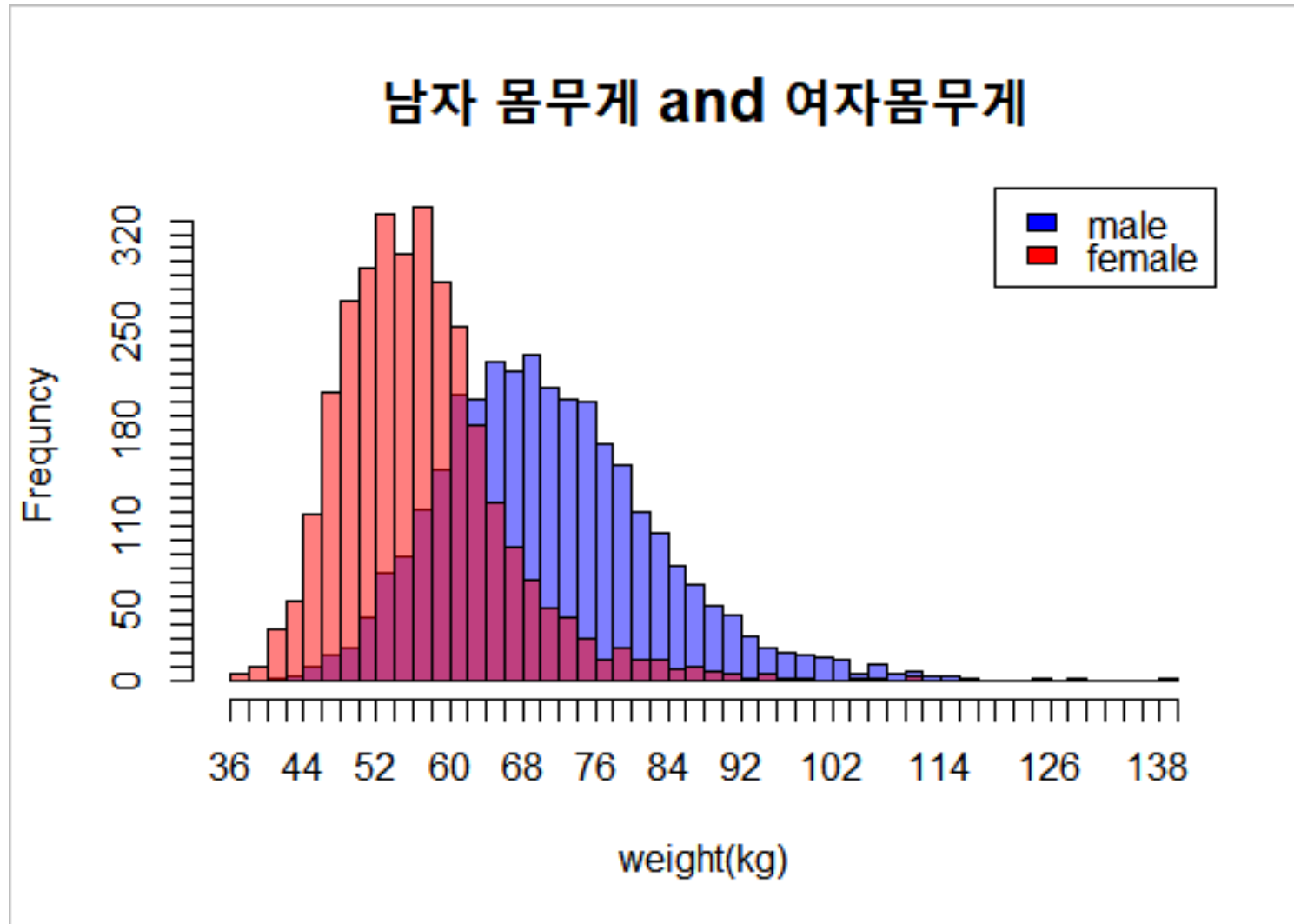
표본표준편차 s 은 다음과 같은 식으로 구한다.

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

1. 데이터의 이해

정형 데이터의 시각화

수치형 자료



비정형 데이터의 분류

이미지 데이터

CIFAR-10

- 32x32픽셀의 60,000개의 컬러 이미지
- 각10개의 클래스

비행기

자동차

새

고양이

사슴

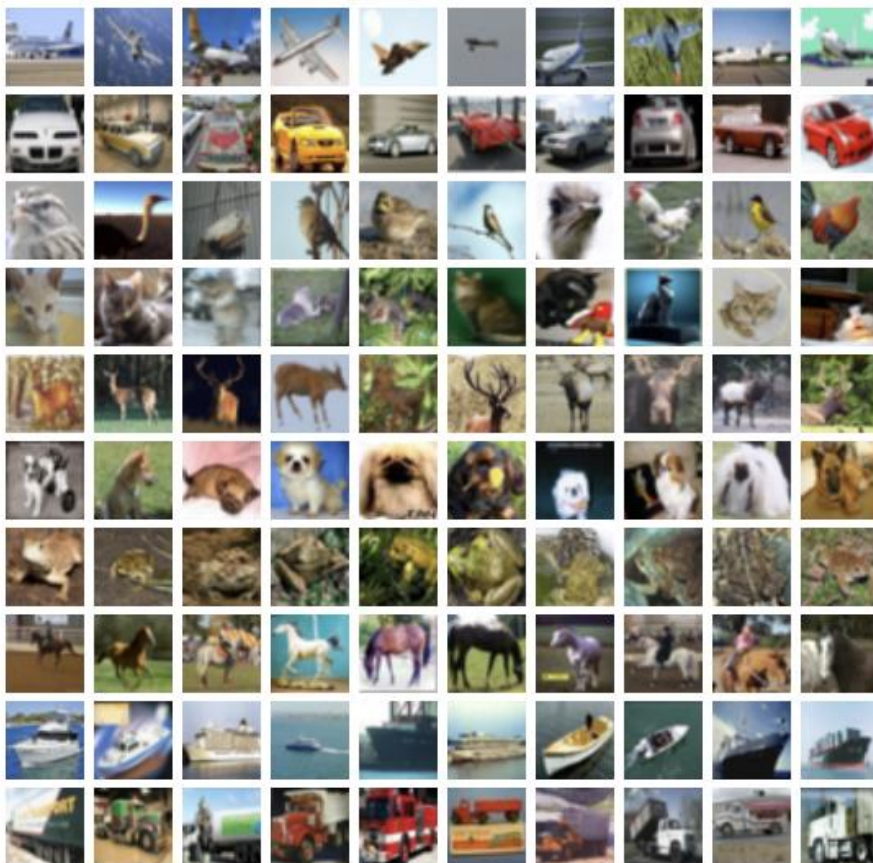
개

개구리

말

배

트럭



MNIST

- 28x28픽셀의 60,000개의 손글씨 데이터
- 각10개의 클래스



텍스트 데이터

IMDB 리뷰 데이터

- 감성분류를 위한 영화 사이트 IMDB의 50,000개의 리뷰 데이터
- 해당 리뷰가 긍정인 경우 1(50%), 부정인 경우 0(50%)으로 라벨링

<sos> this film was just brilliant casting location scenery story direction everyone's really suited the part they played and you could just imagine being there robert <unk> is an amazing actor and now the same being director <unk> father came from the same scottish island as myself so i loved the fact there was a real connection with this film the witty remarks throughout the film were great it was just brilliant so much that i bought the film as soon as it was released for <unk> and would recommend it to everyone to watch and the fly fishing was amazing really cried at the end it was so sad and you know what they say if you cry at a film it must have been good and this definitely was also <unk> to the two little boy's that played the <unk> of norman and paul they were just brilliant children are often left out of the <unk> list i think because the stars that play them all grown up are such a big profile for the whole film but these children are amazing and should be praised for what they have done don't you think the whole story was so lovely because it was true and was someone's life after all that was shared with us all

2. 정형 데이터를 위한 분석 방법 및 실습

이변량 데이터

Female과Male을 동시에 분석할 수는 없을까?

Female	Male
7, 59, 78, 79, 60, 65, 68, 71, 75, 48, 51, 55, 56, 41, 43, 44, 75, 78, 80, 81, 83, 83, 85	48, 49, 49, 30, 30, 31, 32, 35, 37, 41, 86, 42, 51, 53, 56, 42, 44, 50, 51, 65, 67, 51, 56, 58, 64, 64, 75



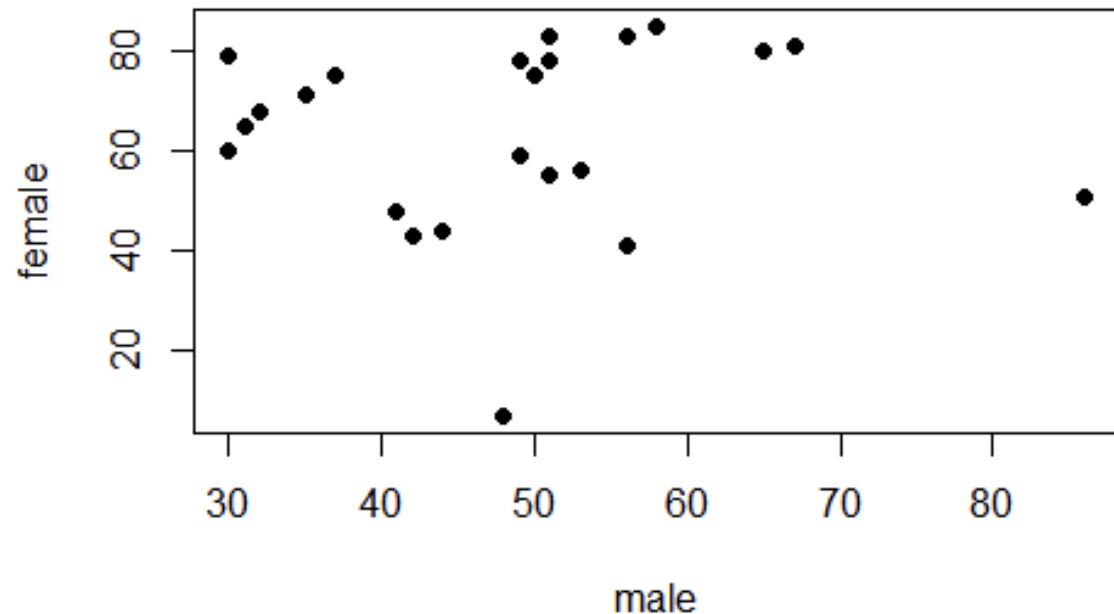
테이블을 재구성하자

obs	Female	Male
1	7	48
2	59	49
3	78	49
4	79	30
5	60	30
6	65	31
...

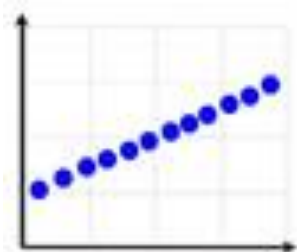
2. 정형 데이터를 위한 분석 방법 및 실습

이변량 데이터의 시각화: 산점도

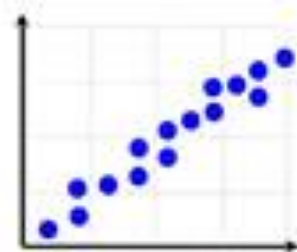
```
female = c(7, 59, 78, 79, 60, 65, 68, 71, 75, 48, 51,  
           55, 56, 41, 43, 44, 75, 78, 80, 81, 83, 83, 85)  
male = c(48, 49, 49, 30, 30, 31, 32, 35, 37, 41, 86,  
         51, 53, 56, 42, 44, 50, 51, 65, 67, 51, 56, 58)  
plot(male, female, pch = 16)
```



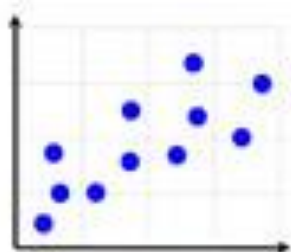
두 변수간 관련성이 있는가?



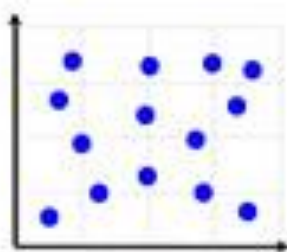
Perfect
Positive
Correlation



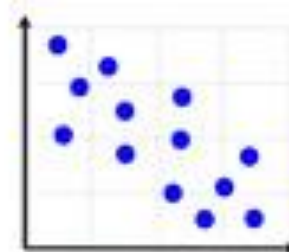
Strong
Positive
Correlation



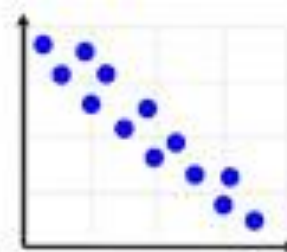
Weak
Positive
Correlation



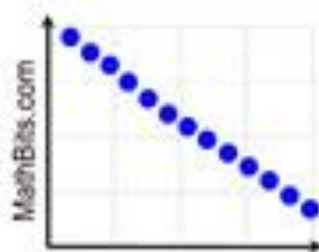
No
Correlation



Weak
Negative
Correlation



Strong
Negative
Correlation



Perfect
Negative
Correlation

➤ 상관계수

- 변수 간의 관계의 강함을 보는 척도
- $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ 을 얻어진 표본 (2변량 자료)이라 하자. \bar{x} 와 \bar{y} 를 각각 x 와 y 의 표본평균으로 하였을 때

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

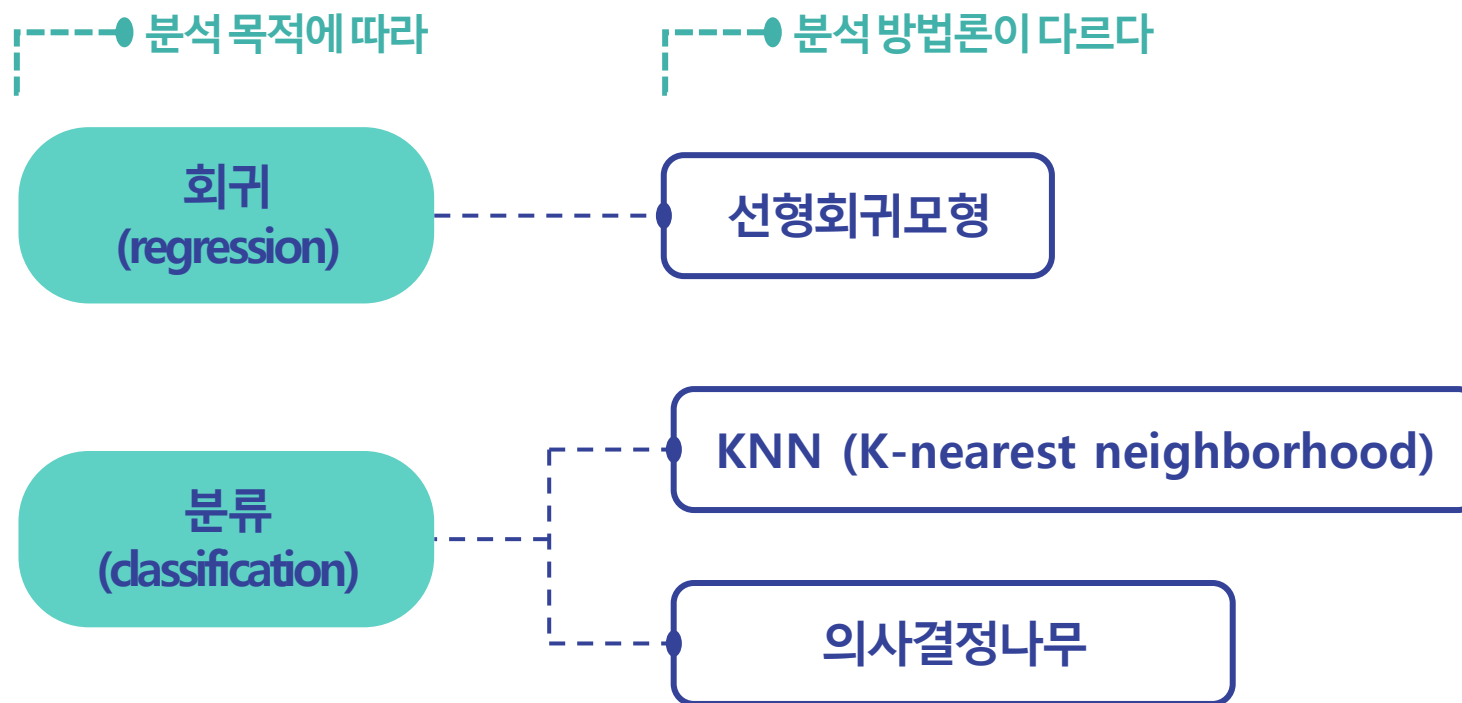
x 와 y 표본 공분산(sample covariance)이라고 한다.

- 또 s_x^2 와 s_y^2 을 각각 x 와 y 의 표본분산이라고 하면

$$\gamma = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}}$$

를 표본상관계수 (sample correlation coefficient)라고 한다.

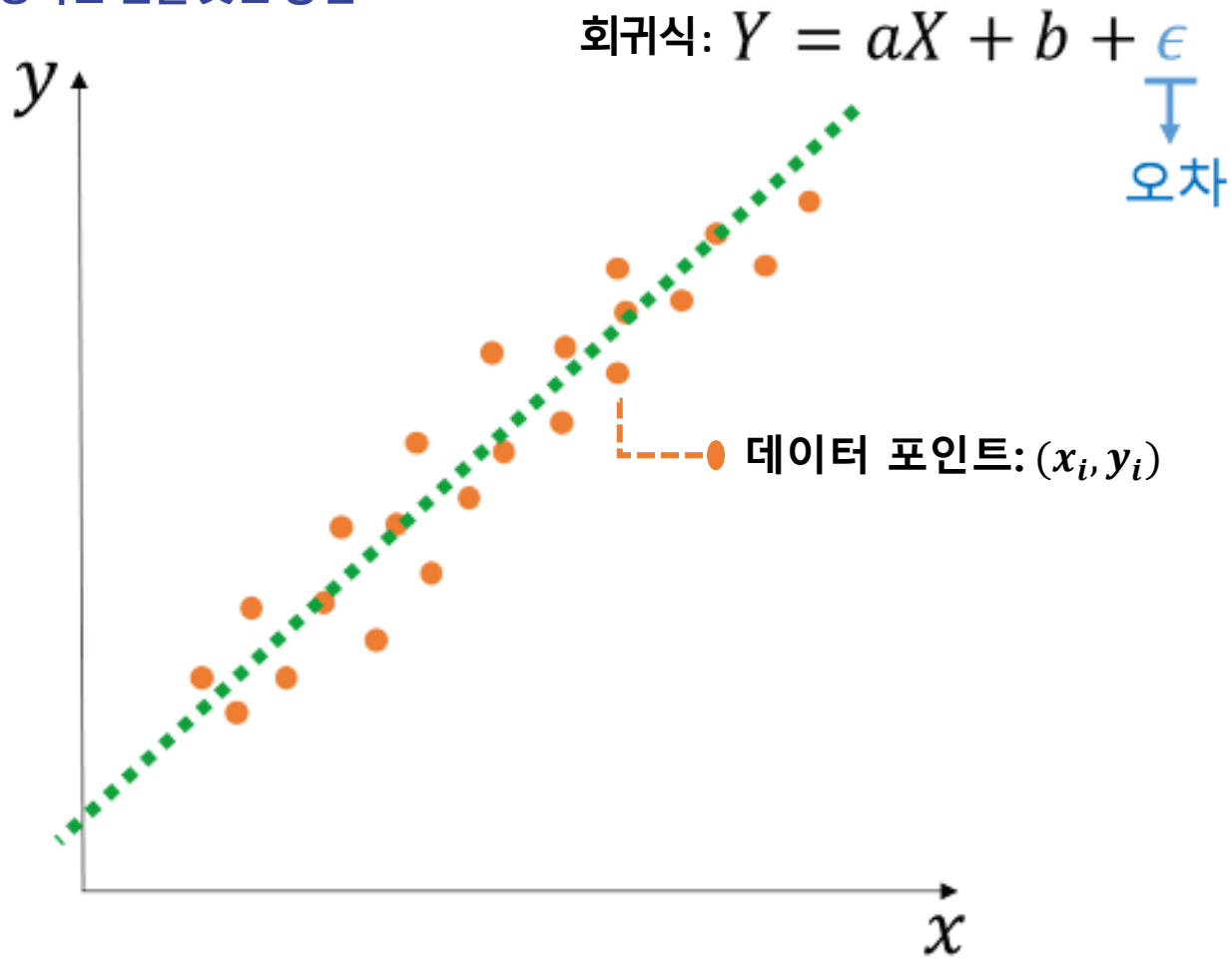
정형 데이터 분석방법



회귀

선형 회귀모형

-데이터를 가장 잘 설명하는 선을 찾는 방법

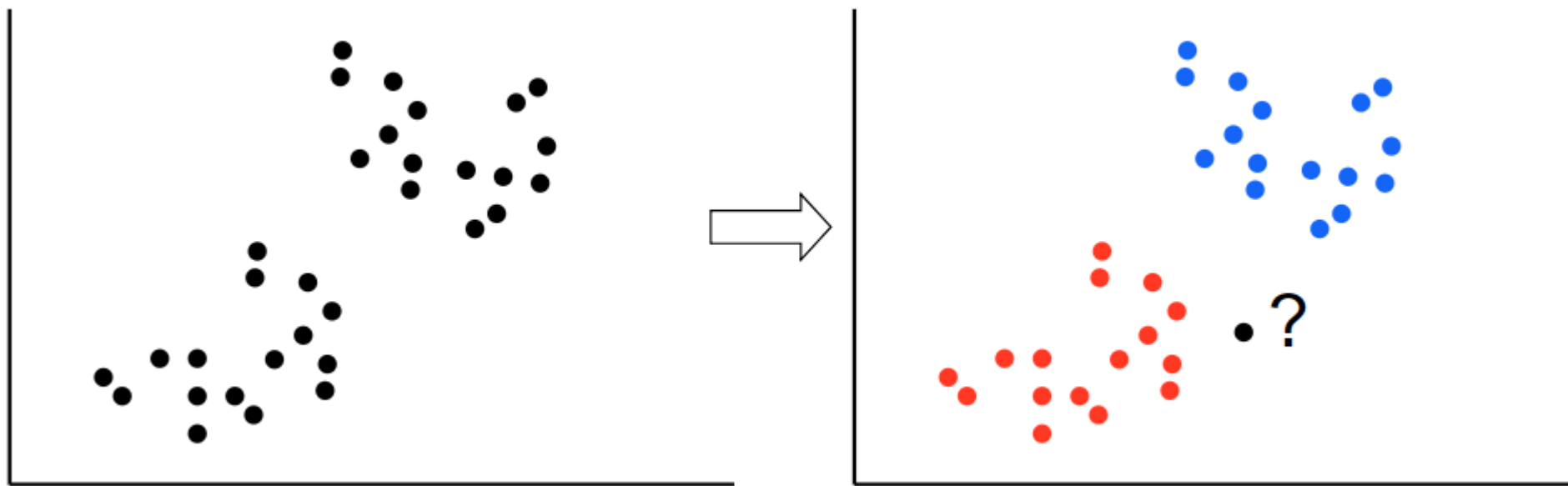


2. 정형 데이터를 위한 분석 방법 및 실습

분류

KNN

-내 이웃의 정보를 사용하여 데이터를 나누는 방법

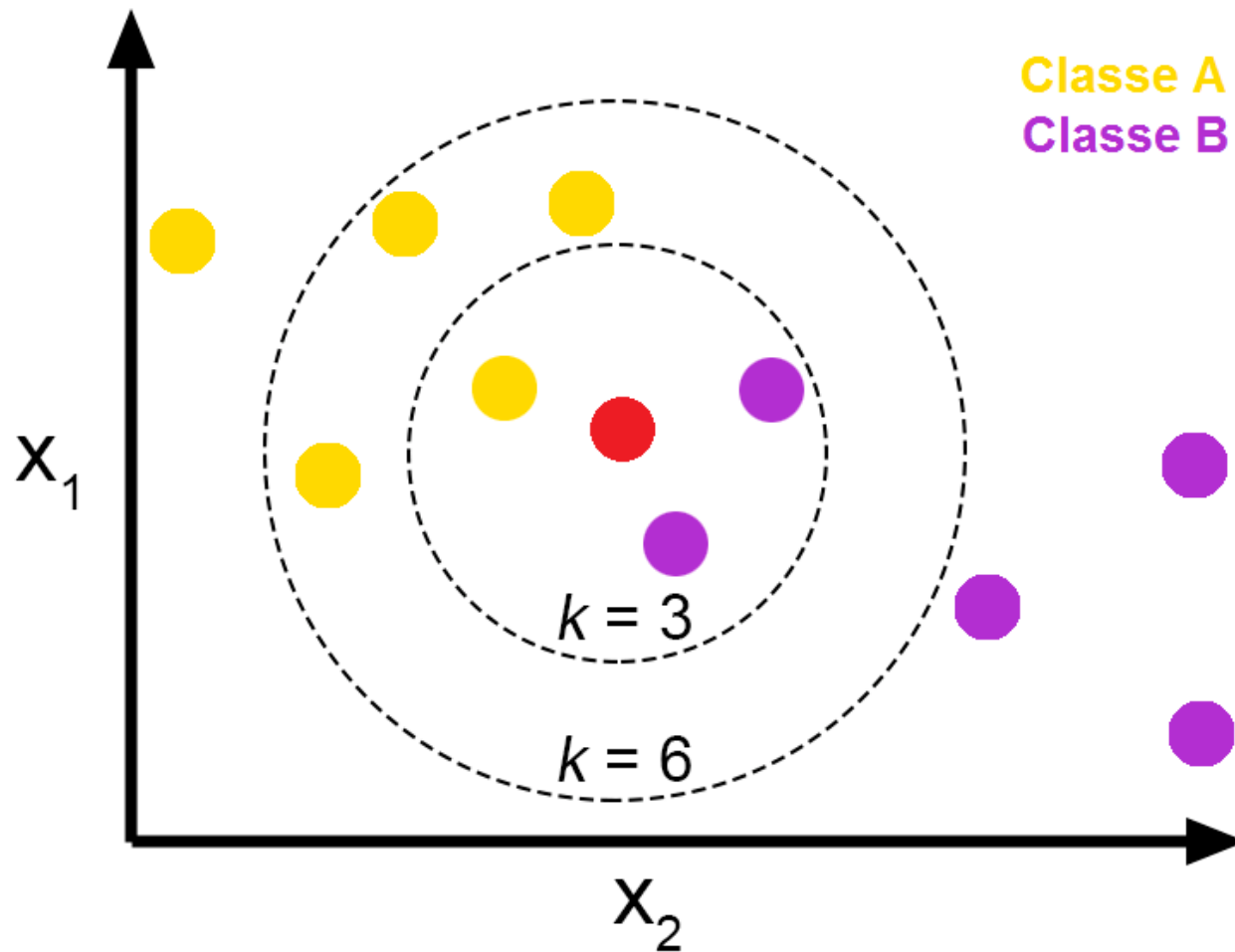


2. 정형 데이터를 위한 분석 방법 및 실습

분류

KNN

-내 이웃의 정보를 사용하여 데이터를 나누는 방법



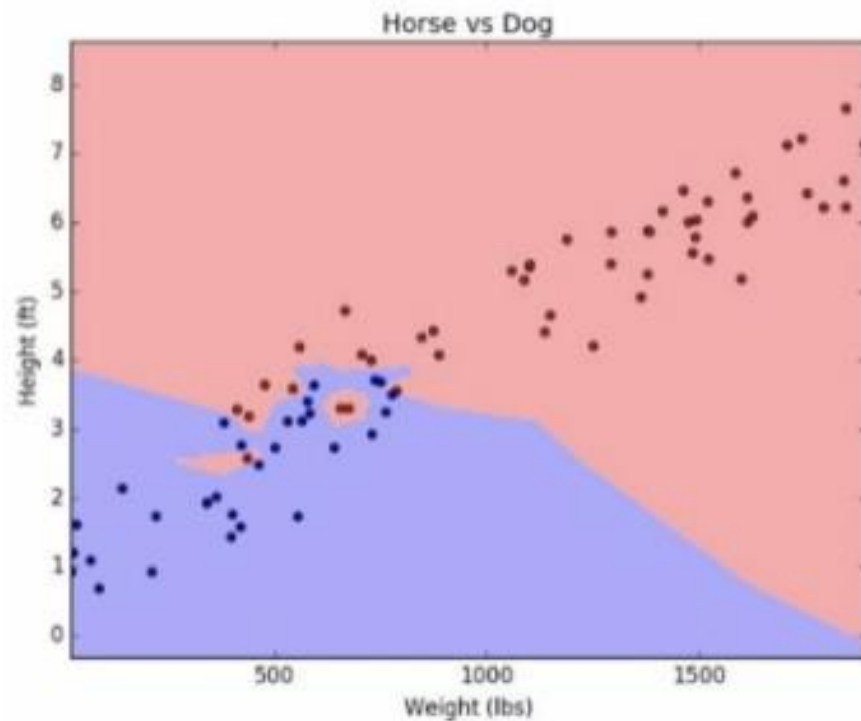
2. 정형 데이터를 위한 분석 방법 및 실습

분류

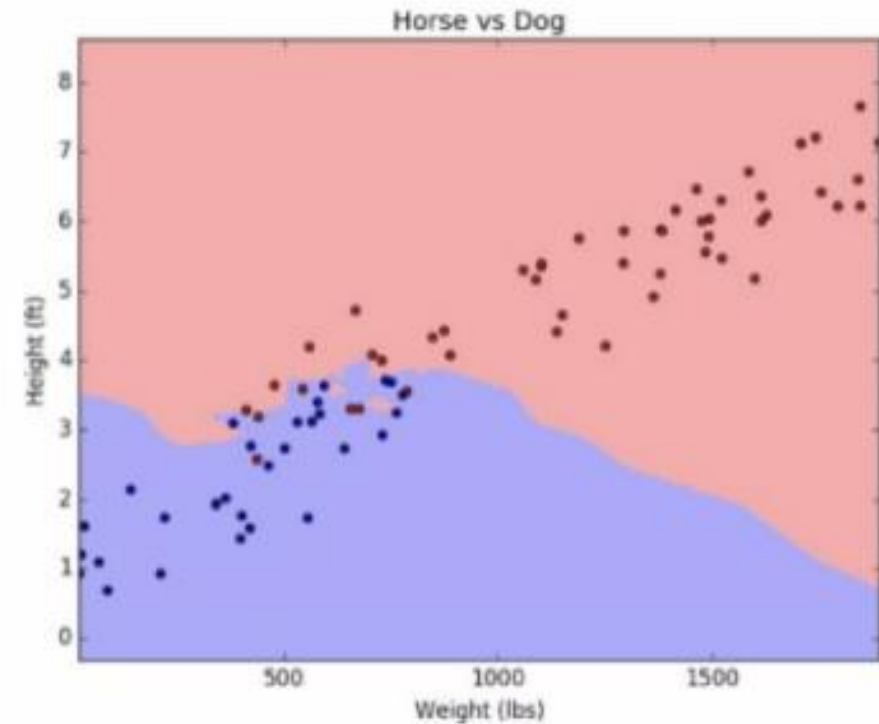
KNN

-내 이웃의 정보를 사용하여 데이터를 나누는 방법

k=1



k=5

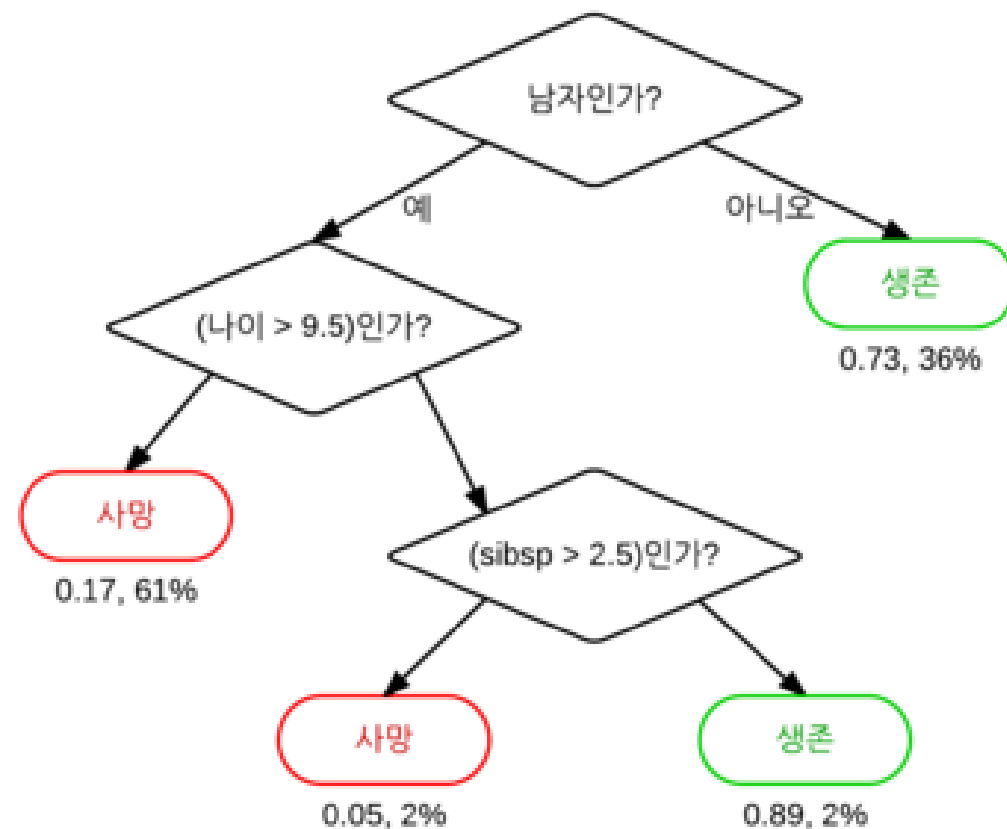
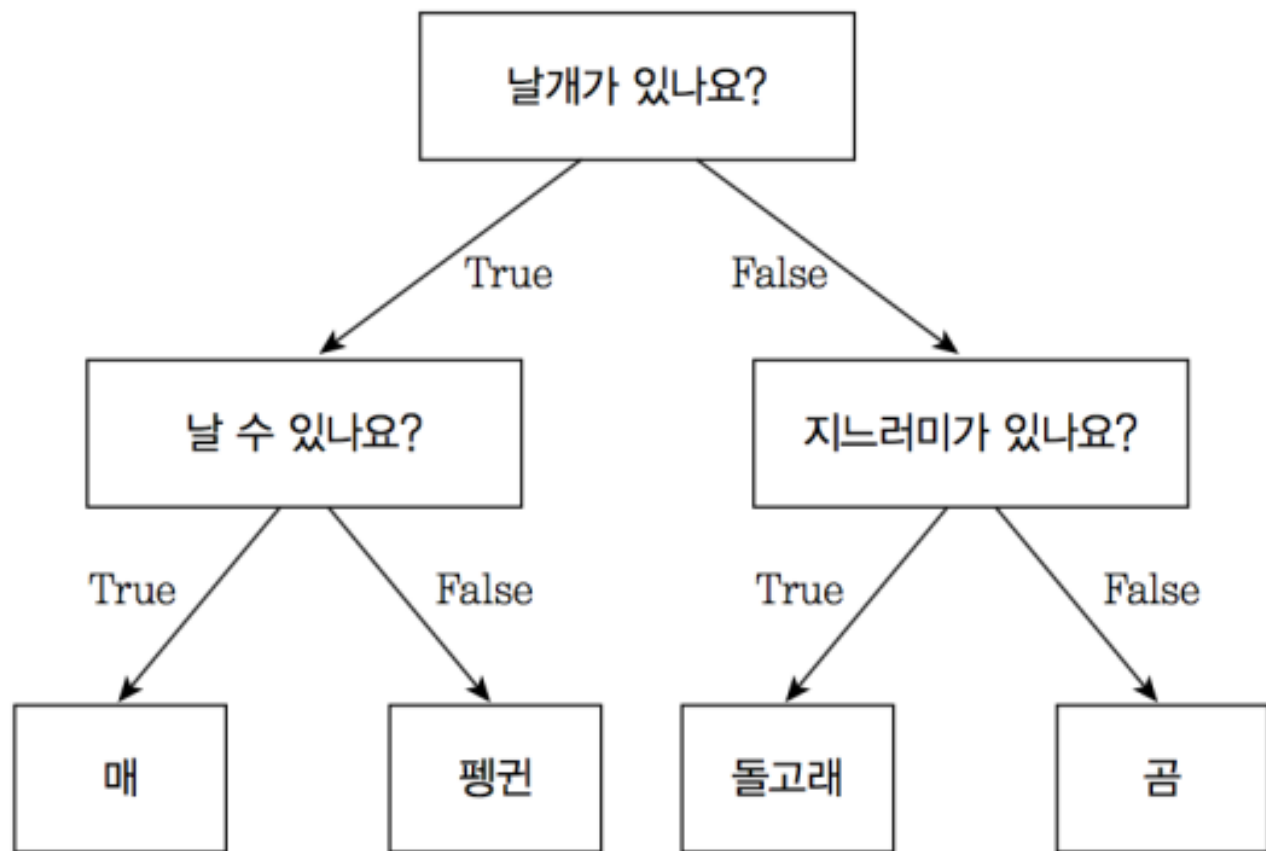


2. 정형 데이터를 위한 분석 방법 및 실습

분류

의사결정나무

-데이터의 조합에 대한 의사결정 규칙에 따라 데이터를 분류하는 방법



2. 정형 데이터를 위한 분석 방법 및 실습

분류

의사결정나무

-데이터의 조합에 대한 의사결정 규칙에 따라 데이터를 분류하는 방법

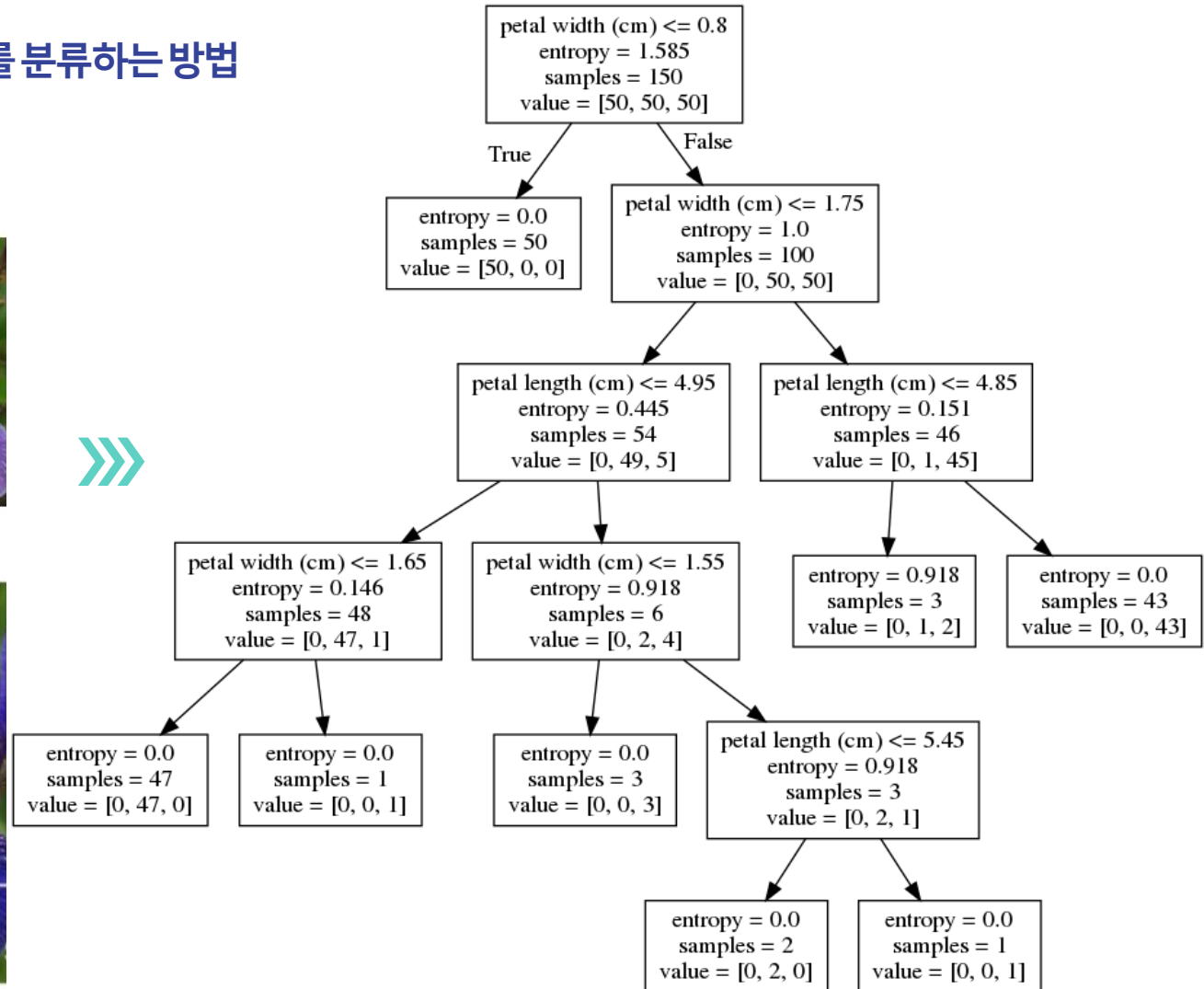
- 데이터 과학에서 Iris DataSet

- 아이리스 품종 중 Setosa, Versicolor, Virginica 분류에 대한 로널드 피셔의 1936년 논문에서 사용된 데이터 셋.



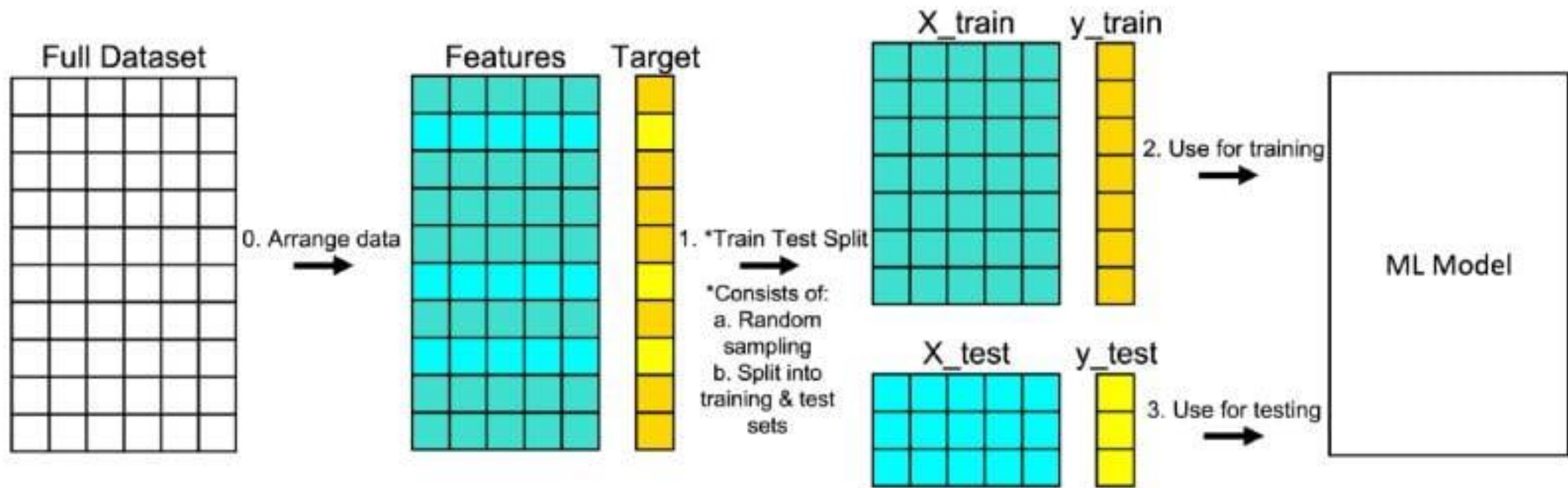
- 꽃받침(Sepal)과 꽃잎(Petal)의 길이 너비로 세개 품종을 분류

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	species
0	4.9	3.0	1.4	0.2	setosa
1	4.7	3.2	1.3	0.2	setosa
2	4.6	3.1	1.5	0.2	setosa
3	6.4	3.2	4.5	1.5	versicolor
4	6.9	3.1	4.9	1.5	versicolor
5	5.5	2.3	4.0	1.3	versicolor
6	7.1	3.0	5.9	2.1	virginica
7	6.3	2.9	5.6	1.8	virginica
8	7.6	3.0	6.6	2.1	virginica



데이터의 분할

-----● 데이터를 분할시켜 학습하는 것이 기본!



3. 비정형 데이터를 위한 분석 방법 및 실습

● 데이터를 왜 분할해야 될까?

	화재발생연도	시군구	사망자수	부상자수	재산피해금액	출동횟수	출동횟수_겨울	출동횟수_여름
0	2017	은평구	0.0	3	218200	159	51	32
1	2017	종로구	1.0	3	1077665	234	55	69
2	2017	중구	5.0	14	485392	198	48	47
3	2017	중랑구	2.0	5	332366	196	53	38
4	2018	은평구	5.0	10	419503	214	58	47
5	2018	종로구	14.0	22	574300	254	71	70
6	2018	중구	0.0	23	1257005	275	76	74
7	2018	중랑구	2.0	8	201421	254	72	55
8	2019	은평구	3.0	20	2412769	196	62	34
9	2019	종로구	4.0	16	801094	232	60	63
10	2019	중구	3.0	17	74077097	213	51	39
11	2019	중랑구	1.0	9	322650	210	54	49
12	2020	은평구	2.0	6	504788	192	48	46
13	2020	종로구	2.0	5	639751	217	50	49
14	2020	중구	0.0	10	1284422	185	41	54
15	2020	중랑구	2.0	12	229566	225	54	57
16	2021	은평구	3.0	8	875722	160	57	42
17	2021	종로구	0.0	12	465499	192	48	54

모델링의 목적

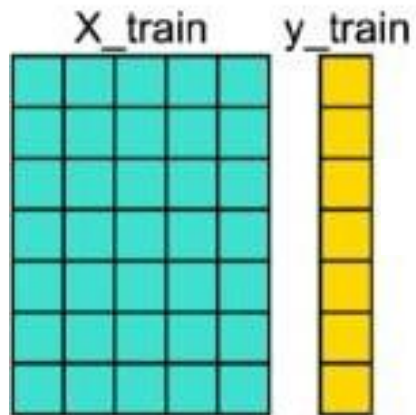
새로운 데이터가 들어왔을 때
이 데이터의 값/라벨을 예측하는 것!

● 2017~2021년도의 화재발생 데이터로 모델링

→ '2022년도의 화재발생 사망자수' 예측할 수 있지 않을까??

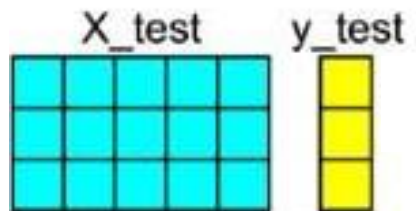
3. 비정형 데이터를 위한 분석 방법 및 실습

데이터분할 설계하기



훈련데이터

독립변수(X_train): 2017~2021년도의 시군구/출동건수/부상자수/...
종속변수(y_train): 2017~2021년도의 사망자수



시험데이터

독립변수(X_test): 2022년도의 시군구/출동건수/부상자수/...
종속변수(y_test)=?

Use for training
→

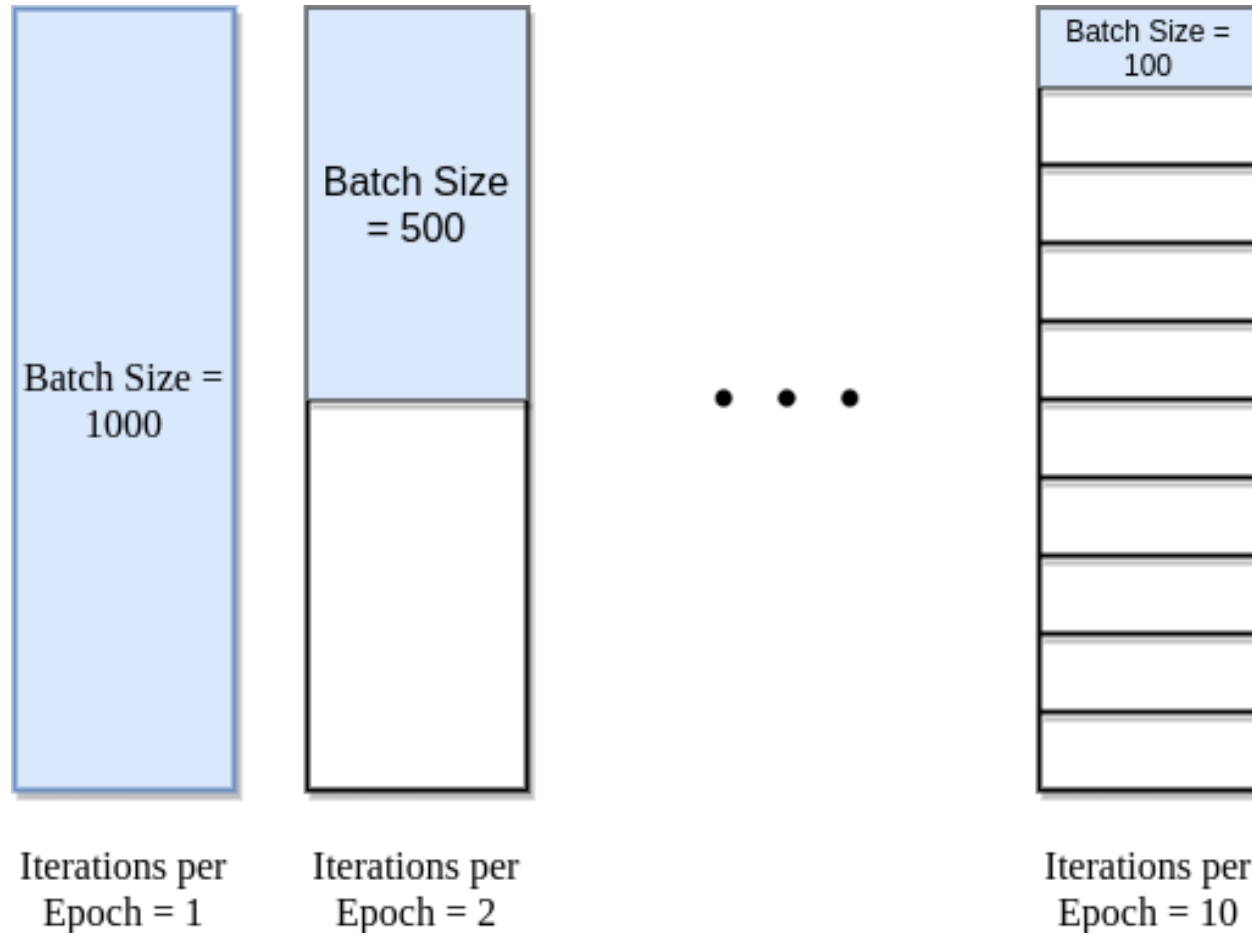
Use for testing
→

ML Model

3. 비정형 데이터를 위한 분석 방법 및 실습

3. 비정형 데이터를 위한 분석 방법 및 실습

-----● 훈련 데이터로 모델을 학습시키자

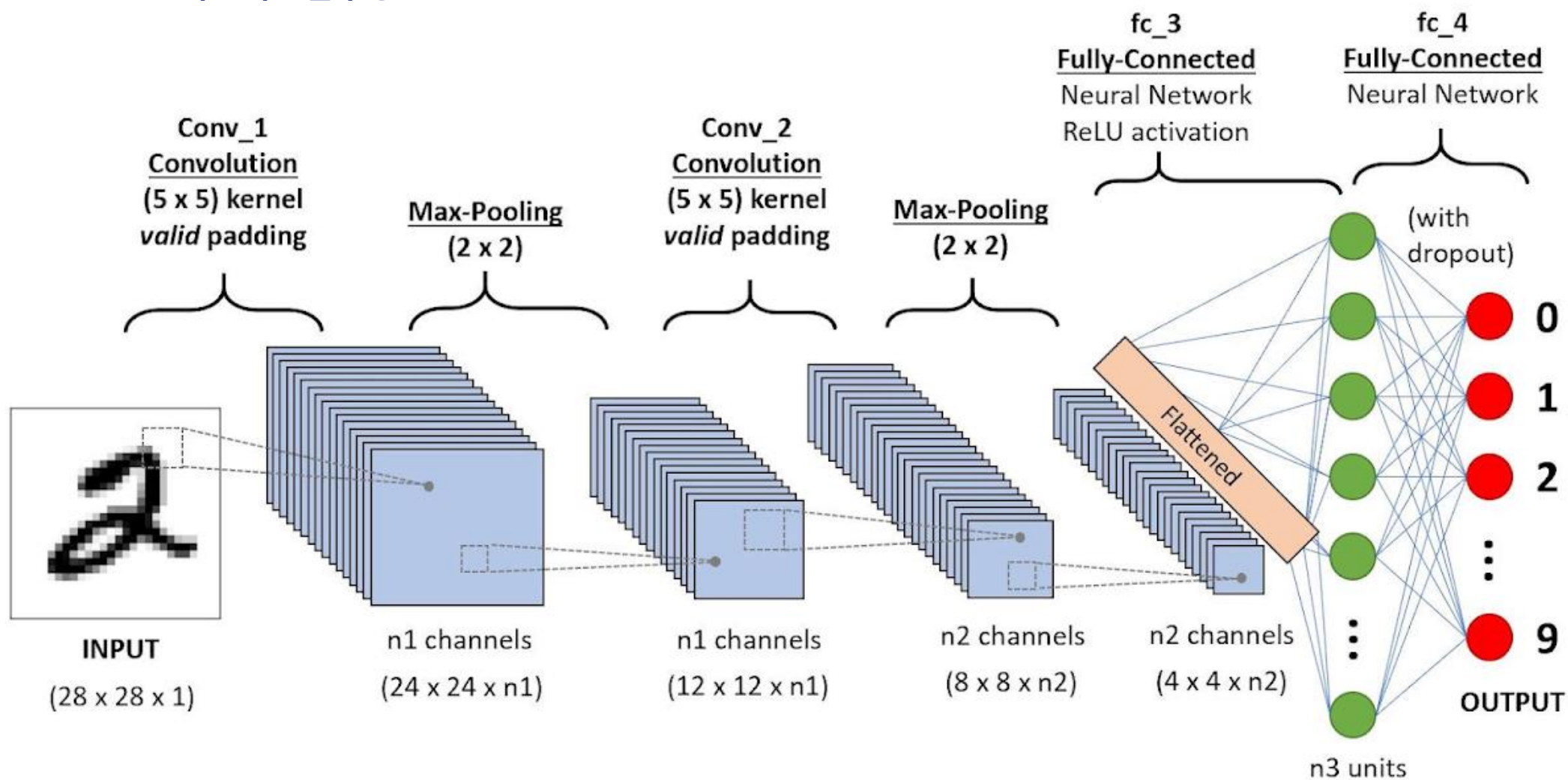


3. 비정형 데이터를 위한 분석 방법 및 실습

이미지 분류

CNN (Convolution Neural Network)

● CNN의 전체 모델 구성

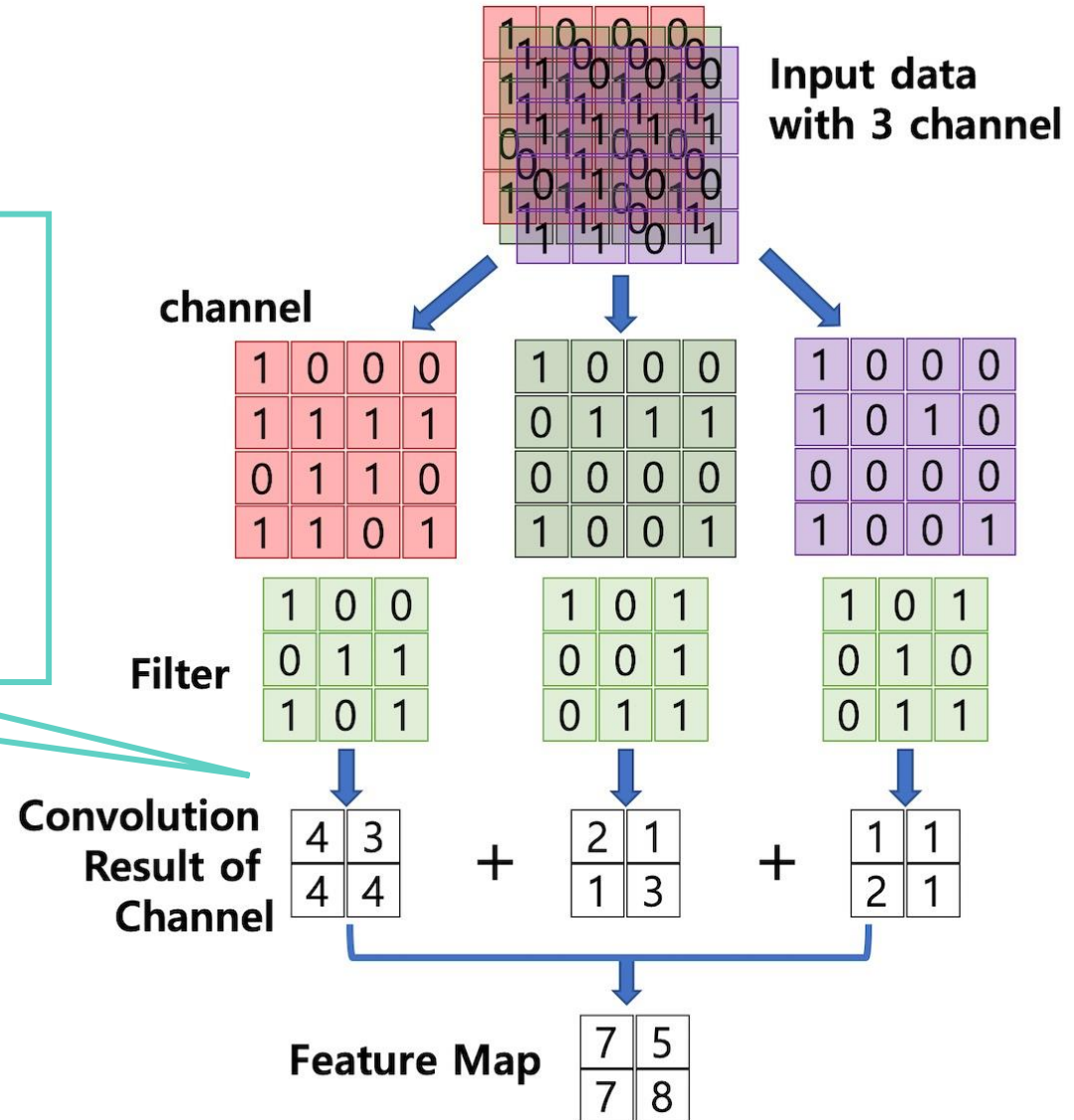
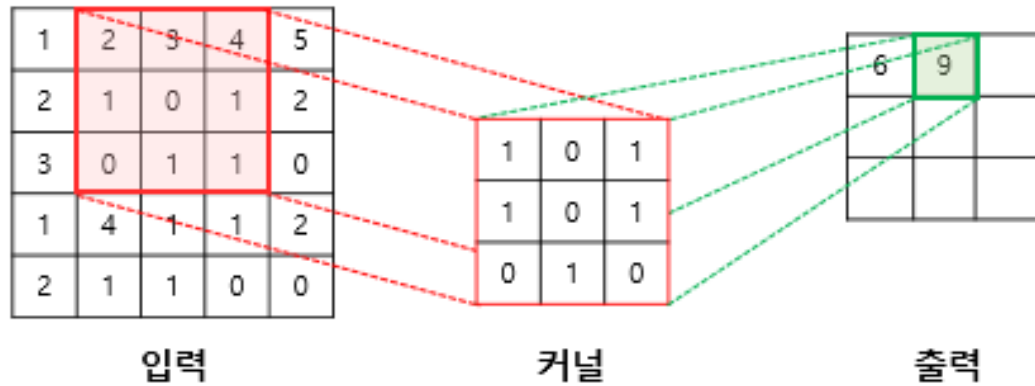


3. 비정형 데이터를 위한 분석 방법 및 실습

이미지 분류

CNN (Convolution Neural Network)

1. 합성곱 과정

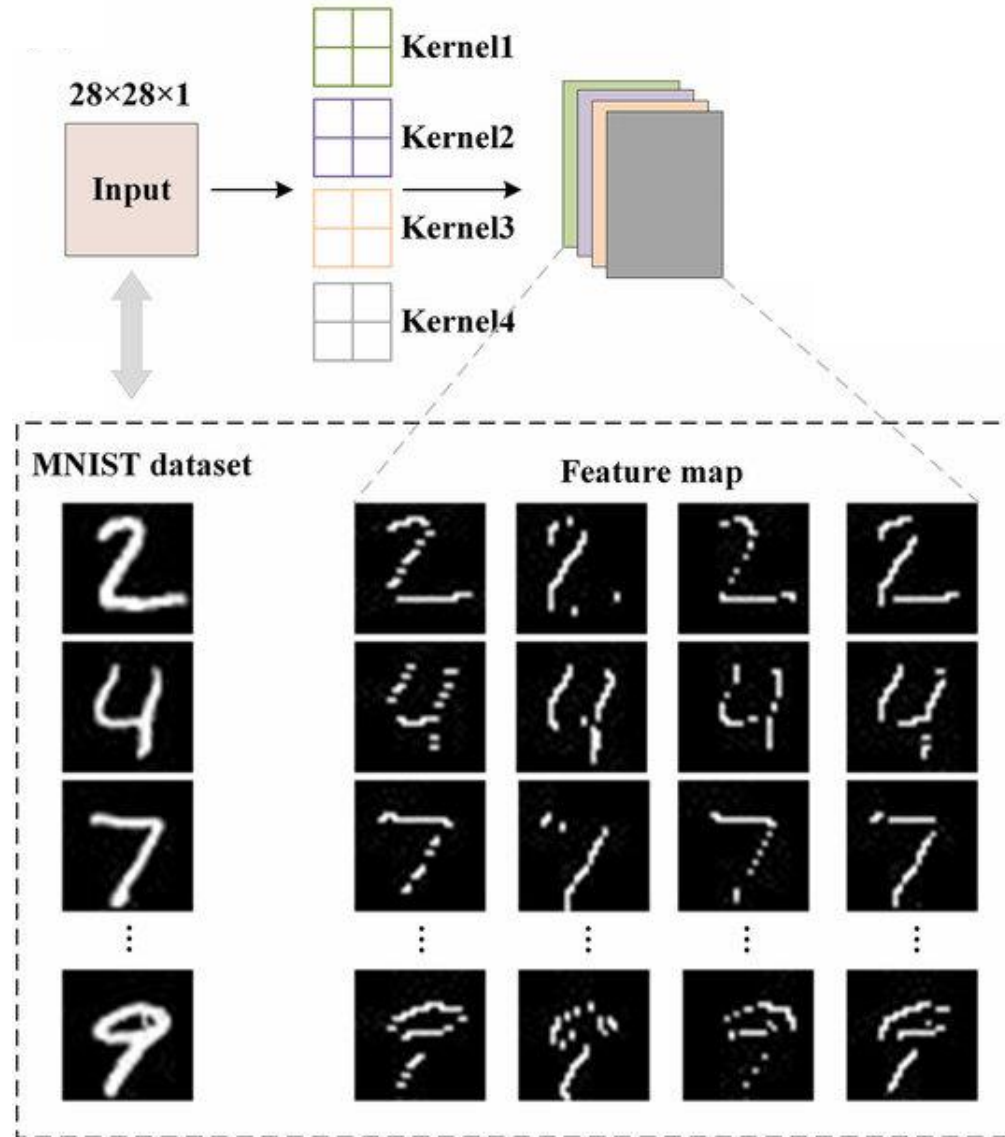


3. 비정형 데이터를 위한 분석 방법 및 실습

이미지 분류

CNN (Convolution Neural Network)

1. 합성곱 과정

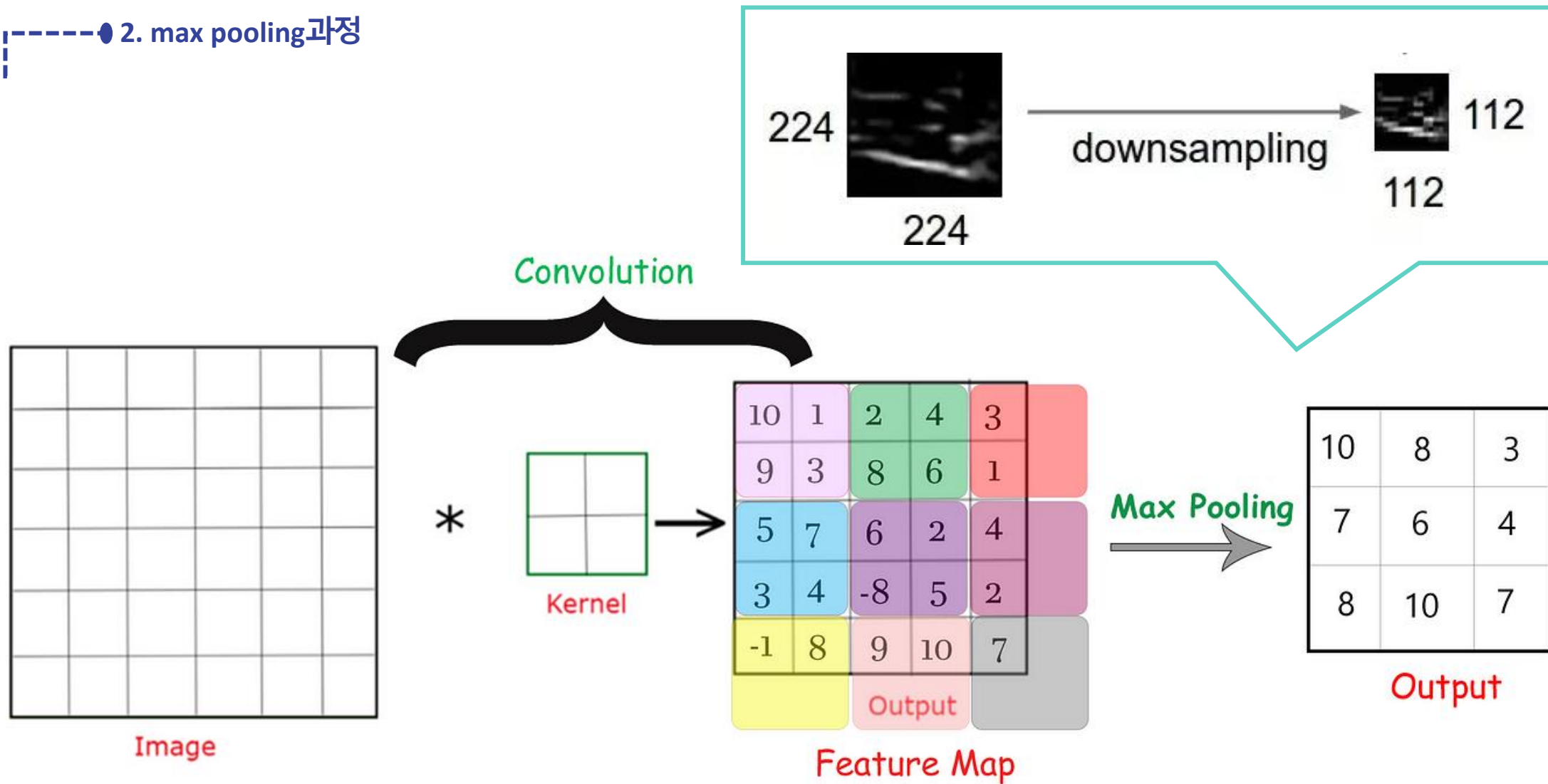


3. 비정형 데이터를 위한 분석 방법 및 실습

이미지 분류

CNN (Convolution Neural Network)

2. max pooling과정

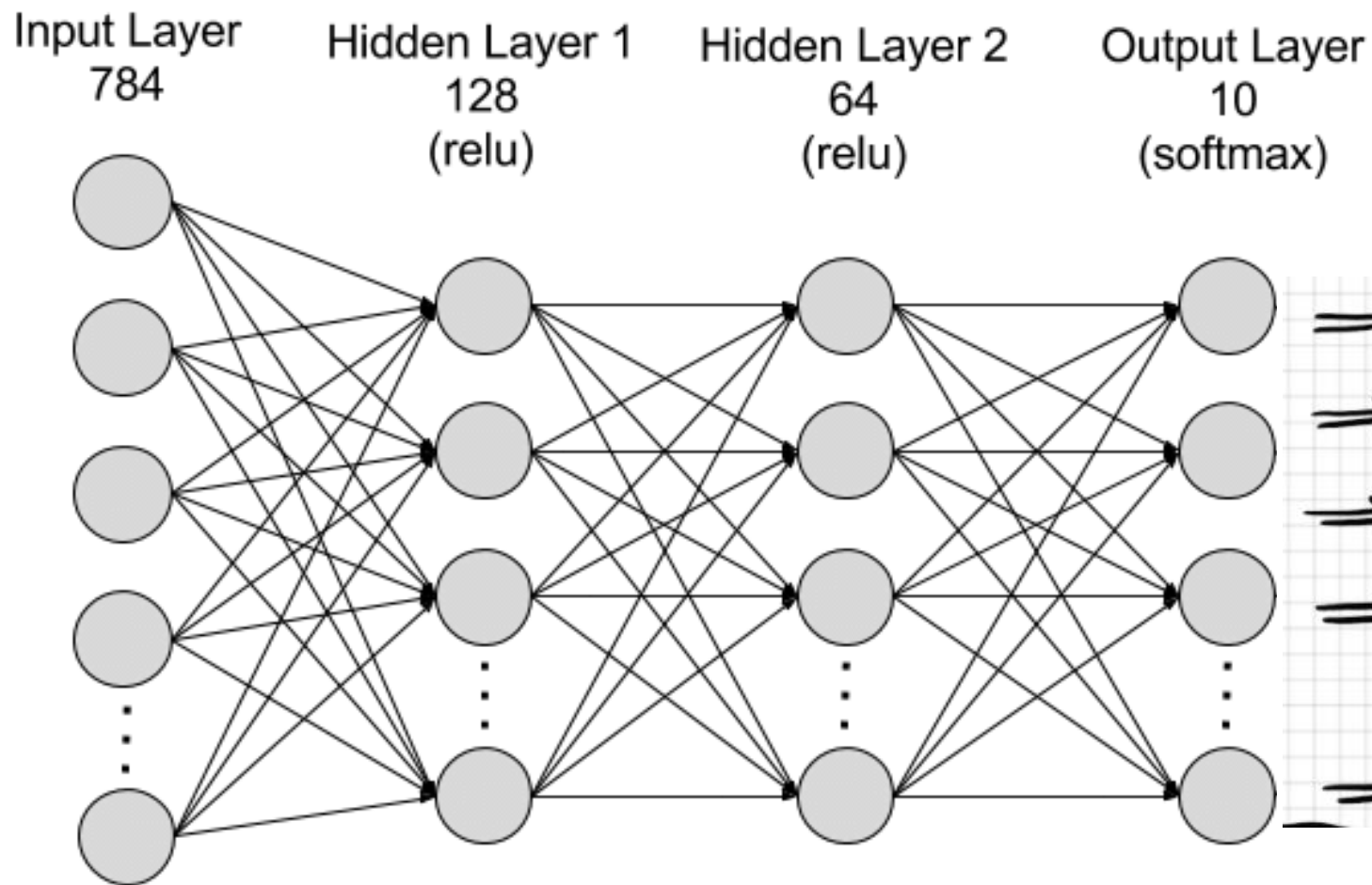


3. 비정형 데이터를 위한 분석 방법 및 실습

이미지 분류

CNN (Convolution Neural Network)

3. fully connected과정



$$\sum_{n=0}^q P(X_n) = 1.$$

$\Rightarrow P(X_0)$
 $\Rightarrow P(X_1)$
 $\Rightarrow P(X_2)$
 $\Rightarrow P(X_3)$
 \vdots
 $\Rightarrow P(X_n)$

3. 비정형 데이터를 위한 분석 방법 및 실습

텍스트 분류

RNN (Recursive Neural Network)

