인공지능

Classification

Byunghwan Jeon, Ph D

pandas 라이브러리

Pandas 라이브러리

데이터 오브젝트는 Series, DataFrame 이 있다.

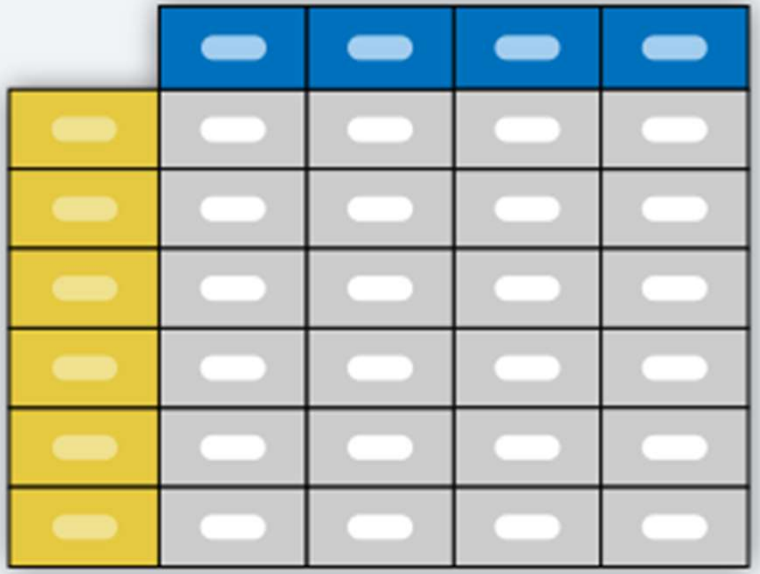
Series는 1차원 배열로 데이터를 담고 있음.

DataFrame은 2차원 배열로 데이터를 담고 있음.

시리즈 (Series)

Series 는 1차원 배열의 형태를 갖는다.
인덱스(노란색)라는 한 가지 기준에
의하여 데이터가 저장된다.

데이터프레임 (DataFrame)



DataFrame 은 2차원 배열의 형태를 갖는다.
인덱스(노란색)와 컬럼(파란색)이라는 두 가지
기준에 의하여 표 형태처럼 데이터가 저장된다.

Pandas 라이브러리: Series

한국외국어대학교 HANKUK UNIVERSITY OF FOREIGN STUDIES

시리즈(Series) 정의

```
s = pd.Series([1, 3, 5, 6, 8])
```

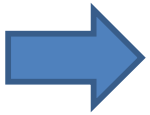
s 출력결과

```
0    1
1    3
2    5
3    6
4    8
dtype: int64
```

Pandas 라이브러리: Series

시리즈(Series) 데이터 값 확인해보기

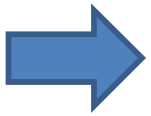
```
s = pd.Series([1, 3, 5, 6, 8])
print(s.values)
```



```
[1 3 5 6 8]
```

type()함수를 사용해서 데이터 타입을 확인해보자.

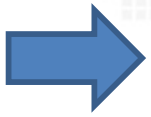
```
print(type(s.values))
```



```
<class 'numpy.ndarray'>
```

pandas 변수로 데이터 타입을 확인해보자.

```
print(s.dtypes)
```



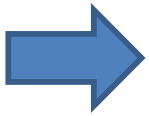
```
<class 'numpy.ndarray'>
int64
```

Pandas 라이브러리: Series

한국외국어대학교 HANKUK UNIVERSITY OF FOREIGN STUDIES

시리즈(Series) 데이터 값 확인해보기

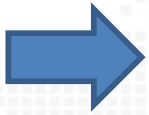
```
s = pd.Series([1, 3, 5, 6, 8])
print(s.values)
```



```
[1 3 5 6 8]
```

- python의 dictionary 자료형을 Series data로 만들 수 있다.
- dictionary의 key가 Series의 index가 된다

```
sdata = {'A': 35000, 'B': 67000, 'C': 12000, 'D': 4000}
obj3 = pd.Series(sdata)
```



인덱스	값
Kim	35000
Beomwoo	67000
Joan	12000
Choi	4000

dtype: int64

Pandas 라이브러리: Data Frame

한국외국어대학교 HANKUK UNIVERSITY OF FOREIGN STUDIES

데이터프레임 (Dataframe)은 엑셀의 테이블처럼 행과 열로 구성된 데이터를 다루는 데이터 구조임.
각 행의 번호를 인덱스(index), 각열의 이름을 속성(attribute)이라고 부름.

```
data = {'name': ['brian','brian','brian', 'andy'],
        'year': [2017, 2018, 2019, 2020],
        'points': [1.7, 3.6, 2.4, 2.9]}

df = pd.DataFrame(data)
```

```
<class 'numpy.ndarray'>
int64
```

	name	year	points
0	brian	2017	1.7
1	brian	2018	3.6
2	brian	2019	2.4
3	andy	2020	2.9

Pandas 라이브러리: Data Frame

한국외국어대학교 HANKUK UNIVERSITY OF FOREIGN STUDIES

데이터프레임 (Dataframe)은 엑셀의 테이블처럼 행과 열로 구성된 데이터를 다루는 데이터 구조임.
각 행의 번호를 인덱스(index), 각열의 이름을 속성(attribute)이라고 부름.

```
data = {'name': ['brian','brian','brian', 'andy'],
        'year': [2017, 2018, 2019, 2020],
        'points': [1.7, 3.6, 2.4, 2.9]}

df = pd.DataFrame(data)
```

```
<class 'numpy.ndarray'>
int64
```

	name	year	points
0	brian	2017	1.7
1	brian	2018	3.6
2	brian	2019	2.4
3	andy	2020	2.9

Pandas 라이브러리: Data Frame

한국외국어대학교 HANKUK UNIVERSITY OF FOREIGN STUDIES

Series 자료를 DataFrame에 추가해보기

```
val = pd.Series([4, 4.5, 3.0, 3.5])

df['grade'] = val
```

	name	year	points	grade
0	brian	2017	1.7	4.0
1	brian	2018	3.6	4.5
2	brian	2019	2.4	3.0
3	andy	2020	2.9	3.5

한국외국어대학교
HANKUK UNIVERSITY OF FOREIGN STUDIES

와인분류문제

Pandas: CSV파일 불러오기

한국외국어대학교 HANKUK UNIVERSITY OF FOREIGN STUDIES

와인 데이터 셋 알아보기

변수 = pd.read_csv(**파일 경로**, sep=':')

red 와인 데이터

```
red = pd.read_csv('http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv', sep=';')
```

white 와인 데이터

```
white = pd.read_csv('http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-white.csv', sep=';')
```

레드와인 N개에 대하여 샘플 추출 후 성분 분석 -> 레드와인은 0이라고 약속
화이트와인 N개에 대하여 샘플 추출 후 성분 분석 -> 화이트와인은 1이라고 약속

정상인 N명의 사람으로부터 혈액 추출 후 성분 분석 -> 정상인이라면 0이라고 약속
병이있는 N명의 환자로부터 혈액 추출 후 성분 분석 -> 환자라면 1이라고 약속

Pandas: 와인 데이터 살펴보기

한국외국어대학교 HANKUK UNIVERSITY OF FOREIGN STUDIES

와인 데이터 셋 알아보기

상위 10개의 데이터 샘플

Attribute(속성), 특징

Tuple, 투플, 데이터 샘플

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	8.8	6.0
1	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.5	6.0
2	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.1	6.0
3	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6.0
4	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6.0
5	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.1	6.0
6	6.2	0.32	0.16	7.0	0.045	30.0	136.0	0.9949	3.18	0.47	9.6	6.0
7	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	8.8	6.0
8	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.5	6.0
9	8.1	0.22	0.43	1.5	0.044	28.0	129.0	0.9938	3.22	0.45	11.0	6.0

Pandas: 와인 데이터 특징

와인 데이터 셋 알아보기

레드와인/화이트와인 둘다 아래와 같은 동일한 12개의 속성을 가지고 있음.

와인의 특징 12가지

속성	설명
Fixed acidity	주석산
Volatile acidity	초산
Citric acid	구연산
Residual sugar	당도
Chlorides	염화물
Free sulfer dioxide	자유 이산화황
Total sulfur dioxide	총 이산화황
Density	밀도
pH	산도
sulphates	황산칼륨
Alcohol	알코올 도수
quality	품질

Pandas: 와인 데이터 특징

AI가 12개의 속성을 기반으로 레드와인인지? 화이트와인인지? 구분할 수 있을까?

구분하기 위해 먼저 12개의 속성 데이터샘플들에 해당하는 정답을 알려줘야 함.

어떻게 알려줘야 할까?

- 이미 데이터셋 자체가 Red와인 데이터셋과 White와인 데이터셋으로 분리되어 있음.
- 데이터셋별로 0또는 1로 구분하능 하도록 속성 하나를 더 추가하여 구분지어줄 수 있음.

Pandas: 와인 데이터 특징

한국외국어대학교
HANKUK UNIVERSITY OF FOREIGN STUDIES

red와인데이터

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	8.8	6.0
1	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.5	6.0
2	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.1	6.0
3	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6.0
4	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6.0
5	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.1	6.0
6	6.2	0.32	0.16	7.0	0.045	30.0	136.0	0.9949	3.18	0.47	9.6	6.0
7	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	8.8	6.0
8	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.5	6.0
9	8.1	0.22	0.43	1.5	0.044	28.0	129.0	0.9938	3.22	0.45	11.0	6.0

white와인데이터

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	8.8	6.0
1	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.5	6.0
2	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.1	6.0
3	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6.0
4	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6.0
5	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.1	6.0
6	6.2	0.32	0.16	7.0	0.045	30.0	136.0	0.9949	3.18	0.47	9.6	6.0
7	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	8.8	6.0
8	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.5	6.0
9	8.1	0.22	0.43	1.5	0.044	28.0	129.0	0.9938	3.22	0.45	11.0	6.0

Pandas: 와인 데이터 정답 부여하기

한국외국어대학교 HANKUK UNIVERSITY OF FOREIGN STUDIES

레드/화이트 -> 0/1 으로 레이블링하기위해 타입을 추가해보자.

레드와인데이터

red

type 속성 추가하여 0으로 레이블링

```
red['type'] = 0
```

화이트와인데이터

wthie

type 속성 추가하여 1로 레이블링

```
white['type'] = 1
```

Pandas: 와인 데이터 정답 부여하기

한국외국어대학교 HANKUK UNIVERSITY OF FOREIGN STUDIES

데이터 프레임에 새로운 속성을 추가하는 방법은 아래와 같이 단순히 속성명과 값을 직접 지정해주면됨.

```
red['type'] = 0
white['type'] = 1
```

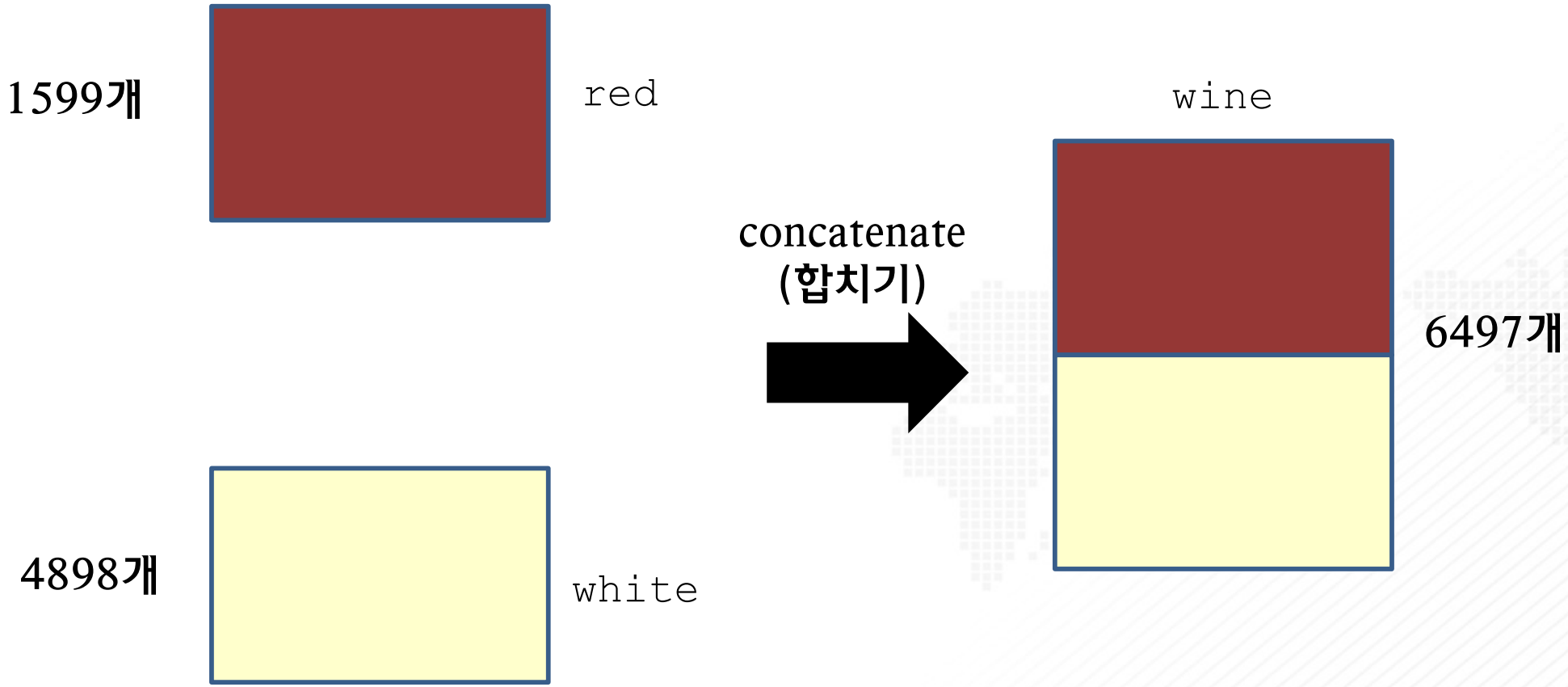
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	type
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.99780	3.51	0.56	9.4	5	0
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.99680	3.20	0.68	9.8	5	0
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.99700	3.26	0.65	9.8	5	0
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.99800	3.16	0.58	9.8	6	0
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.99780	3.51	0.56	9.4	5	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
4893	6.2	0.21	0.29	1.6	0.039	24.0	92.0	0.99114	3.27	0.50	11.2	6	1
4894	6.6	0.32	0.36	8.0	0.047	57.0	168.0	0.99490	3.15	0.46	9.6	5	1
4895	6.5	0.24	0.19	1.2	0.041	30.0	111.0	0.99254	2.99	0.46	9.4	6	1
4896	5.5	0.29	0.30	1.1	0.022	20.0	110.0	0.98869	3.34	0.38	12.8	7	1
4897	6.0	0.21	0.38	0.8	0.020	22.0	98.0	0.98941	3.26	0.32	11.8	6	1

6497 rows × 13 columns

Pandas: 데이터 합치기

type까지 총 13개의 속성이 동일하게 존재한다.

```
wine = pd.concat([red, white])
```



Pandas: 데이터 통계정보 확인

한국외국어대학교 HANKUK UNIVERSITY OF FOREIGN STUDIES

기본적인 데이터의 통계량과 최대, 최소 값 들을 계산하여 보여줌.

```
wine.describe()
```

```
       fixed acidity  volatile acidity  ...      quality         type
count    6497.000000       6497.000000  ...  6497.000000  6497.000000
mean        7.215307          0.339666  ...     5.818378     0.753886
std         1.296434          0.164636  ...     0.873255     0.430779
min         3.800000          0.080000  ...     3.000000     0.000000
25%         6.400000          0.230000  ...     5.000000     1.000000
50%         7.000000          0.290000  ...     6.000000     1.000000
75%         7.700000          0.400000  ...     6.000000     1.000000
max        15.900000          1.580000  ...     9.000000     1.000000

[8 rows x 13 columns]
```

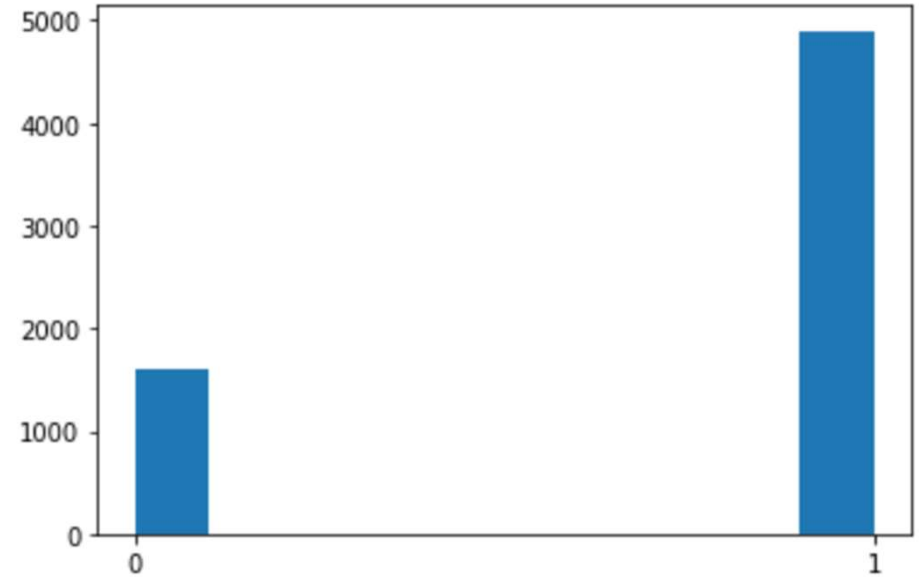
Panda: 데이터 통계정보 확인

한국외국어대학교 HANKUK UNIVERSITY OF FOREIGN STUDIES

기본적인 데이터의 통계량과 최대, 최소 값 들을 계산하여 보여줌.

```
#데이터 히스토그램 그려보기
plt.hist(wine['type'])
plt.xticks([0, 1])
plt.show()

print(wine['type'].value_counts())
```

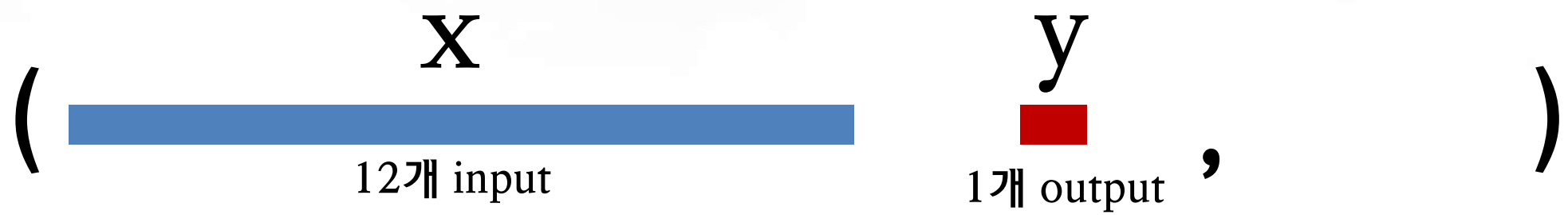



```
1    4898
0    1599
Name: type, dtype: int64
```

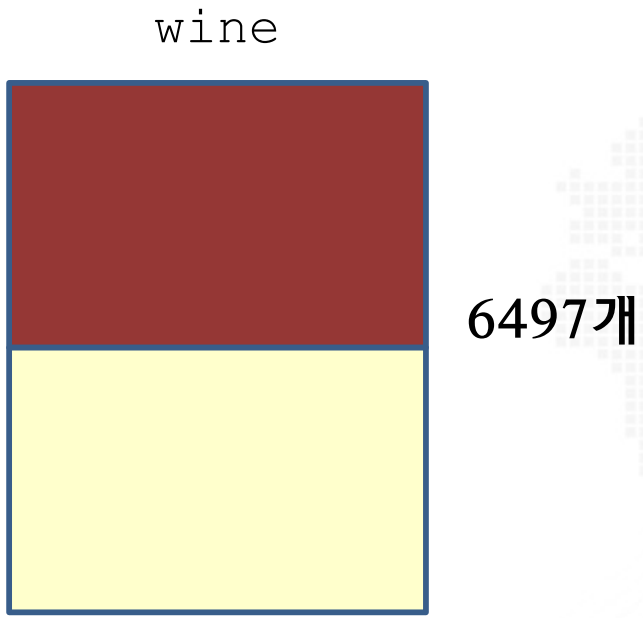
데이터셋 상태

한국외국어대학교 HANKUK UNIVERSITY OF FOREIGN STUDIES

하나의 데이터의 표현



현재 총 6497개의 데이터 셋이 존재 함.



학습효율을 위한 데이터 정규화

한국외국어대학교 HANKUK UNIVERSITY OF FOREIGN STUDIES

데이터 정규화를 통하여 학습이 더 잘되기 만들 수 있음.

```
print(wine.info())
```

```
Int64Index: 6497 entries, 0 to 4897
Data columns (total 13 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   fixed acidity         6497 non-null   float64
 1   volatile acidity      6497 non-null   float64
 2   citric acid           6497 non-null   float64
 3   residual sugar        6497 non-null   float64
 4   chlorides             6497 non-null   float64
 5   free sulfur dioxide   6497 non-null   float64
 6   total sulfur dioxide  6497 non-null   float64
 7   density               6497 non-null   float64
 8   pH                    6497 non-null   float64
 9   sulphates             6497 non-null   float64
 10  alcohol               6497 non-null   float64
 11  quality               6497 non-null   int64
 12  type                  6497 non-null   int64
dtypes: float64(11), int64(2)
memory usage: 710.6 KB
None
```

float64

int64

우리가 알 수 있는 것

- **비어 있는 요소가 없음.**
- **11개는 float64 실수 타입**
- **2개는 int64 정수 타입**

학습효율을 위한 데이터 정규화

한국외국어대학교 HANKUK UNIVERSITY OF FOREIGN STUDIES

데이터 최대-최소 정규화를 통하여 학습이 더 잘되기 만들 수 있음.
모든 값들을 attribute별로 최대, 최소값을 구하여 정규화 시킴. 0~1사이의 값으로 만듦.

$$z = \frac{\mathrm{x} - min(x)}{[\max(x) - min(x)]}$$

```
wine_norm = (wine-wine.min())/(wine.max()-wine.min())
```

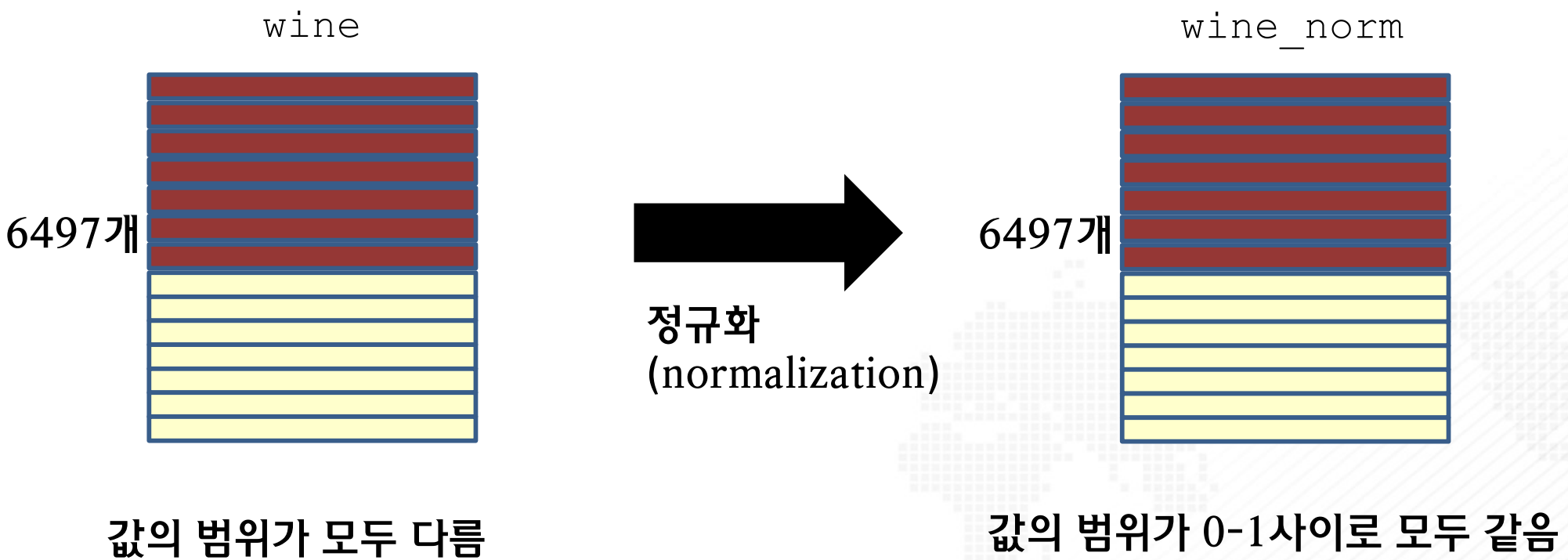
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	type
0	0.297521	0.413333	0.000000	0.019939	0.111296	0.034722	0.064516	0.206092	0.612403	0.191011	0.202899	0.333333	0.0
1	0.330579	0.533333	0.000000	0.030675	0.147841	0.083333	0.140553	0.186813	0.372093	0.258427	0.260870	0.333333	0.0
2	0.330579	0.453333	0.024096	0.026074	0.137874	0.048611	0.110599	0.190669	0.418605	0.241573	0.260870	0.333333	0.0
3	0.611570	0.133333	0.337349	0.019939	0.109635	0.055556	0.124424	0.209948	0.341085	0.202247	0.260870	0.500000	0.0
4	0.297521	0.413333	0.000000	0.019939	0.111296	0.034722	0.064516	0.206092	0.612403	0.191011	0.202899	0.333333	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
4893	0.198347	0.086667	0.174699	0.015337	0.049834	0.079861	0.198157	0.077694	0.426357	0.157303	0.463768	0.500000	1.0
4894	0.231405	0.160000	0.216867	0.113497	0.063123	0.194444	0.373272	0.150183	0.333333	0.134831	0.231884	0.333333	1.0
4895	0.223140	0.106667	0.114458	0.009202	0.053156	0.100694	0.241935	0.104685	0.209302	0.134831	0.202899	0.500000	1.0
4896	0.140496	0.140000	0.180723	0.007669	0.021595	0.065972	0.239631	0.030461	0.480620	0.089888	0.695652	0.666667	1.0
4897	0.181818	0.086667	0.228916	0.003067	0.018272	0.072917	0.211982	0.044342	0.418605	0.056180	0.550725	0.500000	1.0

6497 rows × 13 columns

학습효율을 위한 데이터 정규화

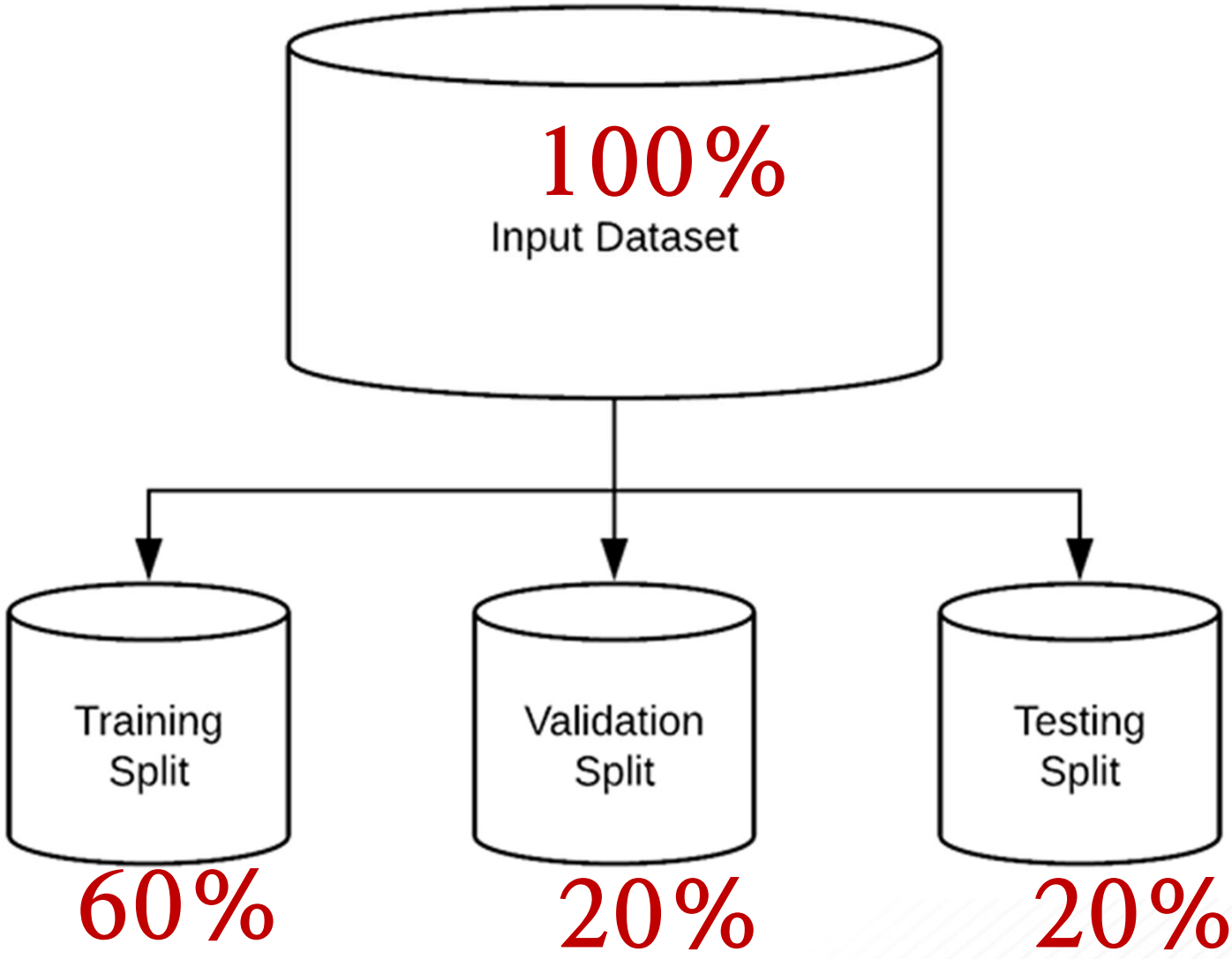
데이터 정규화를 통하여 학습이 더 잘되기 만들 수 있음.

```
wine_norm = (wine-wine.min())/(wine.max()-wine.min())
```



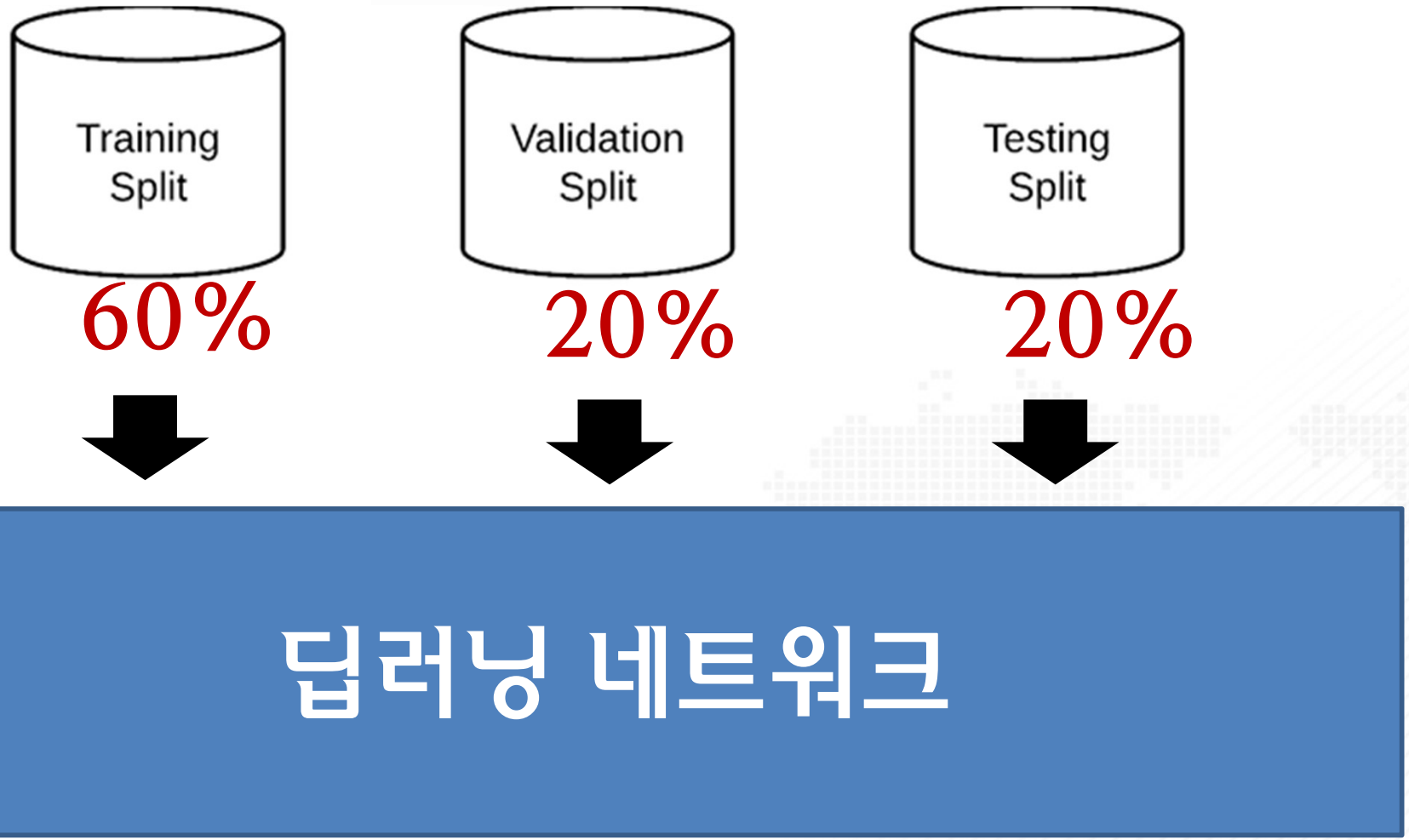
학습을 위한 훈련, 검증, 테스트 데이터셋

전체 데이터로부터 3가지 즉, 학습용, 검증용, 테스트용으로 나누어 학습이 진행됨.



학습을 위한 훈련, 검증, 테스트 데이터셋

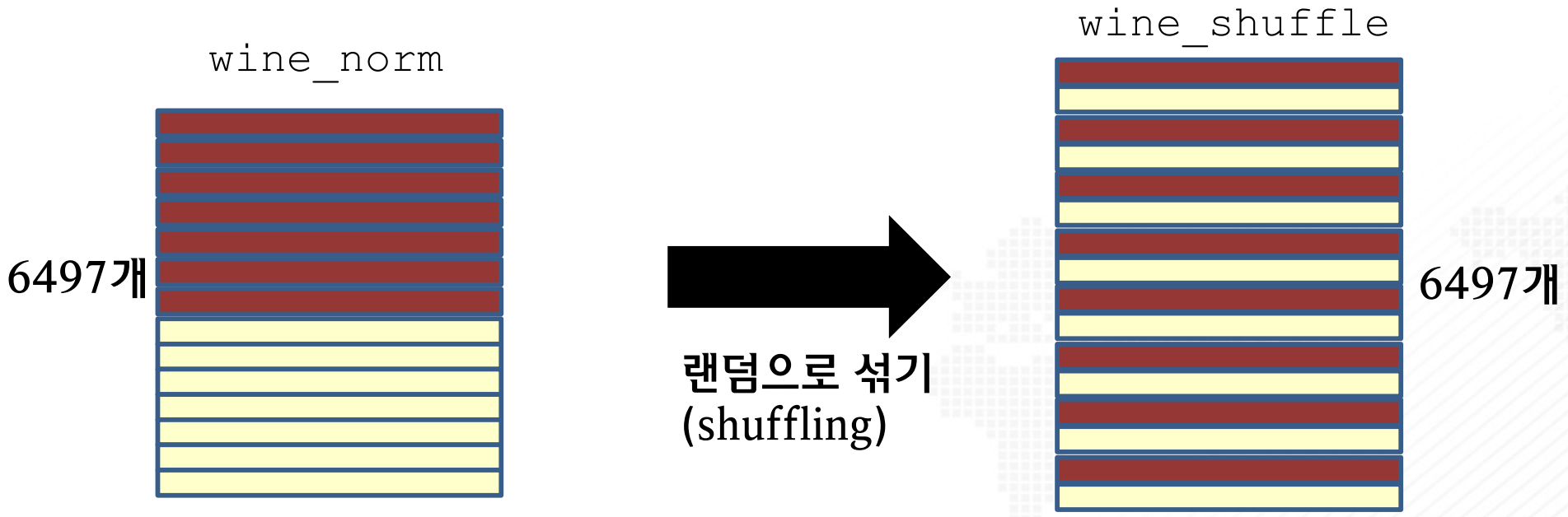
전체 데이터로부터 3가지 즉, 학습용, 검증용, 테스트용으로 나누어 학습이 진행됨.



학습을 위한 훈련, 검증, 테스트 데이터셋

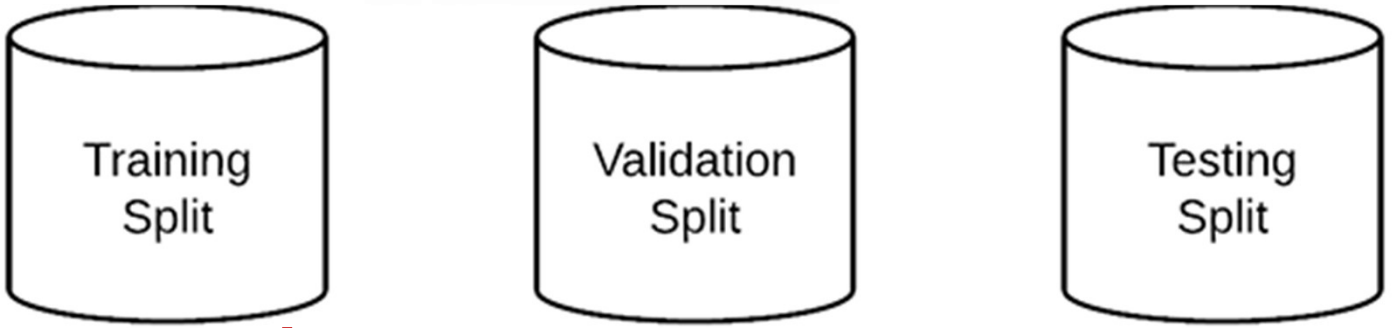
데이터셋을 균등하게 나누기 위하여 먼저 데이터를 섞어주어야 함.
frac=1 은 100% 모두에 대해 랜덤 샘플링해서 재배열한다는 의미

```
wine_shuffle = wine_norm.sample(frac=1)
```



학습을 위한 훈련, 검증, 테스트 데이터셋

아래와 같은 비율로 나누기 위해 먼저 numpy로 자료형 변환



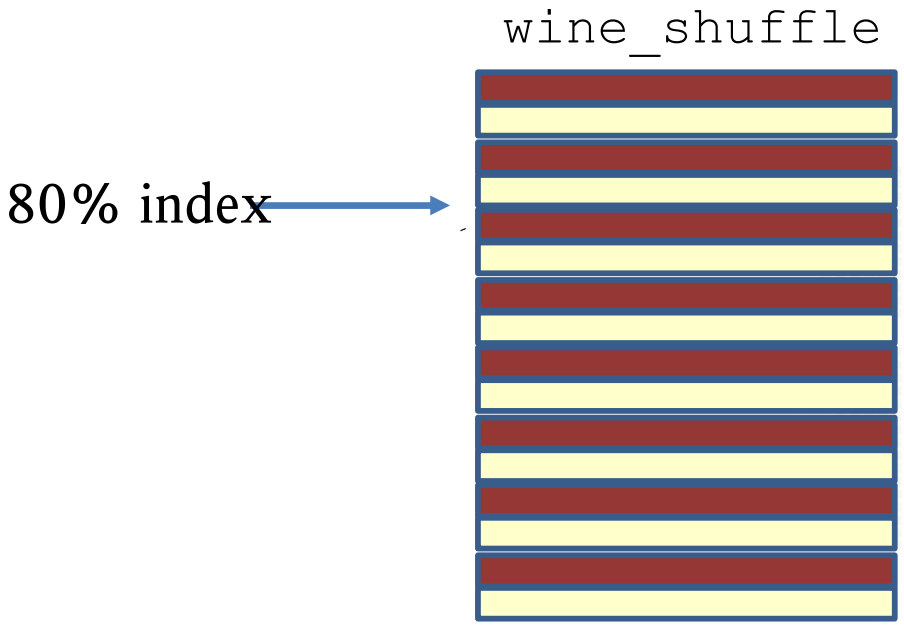

```
wine_np = wine_shuffle.to_numpy()
```

학습을 위한 훈련, 검증, 테스트 데이터셋

아래와 같은 비율로 나누기 위해 먼저 numpy로 자료형 변환

```
train_idx = int(len(wine_np) * 0.8)
```

전체 개수 * 0.8 = 80%의 위치 index를 가져올 수 있음

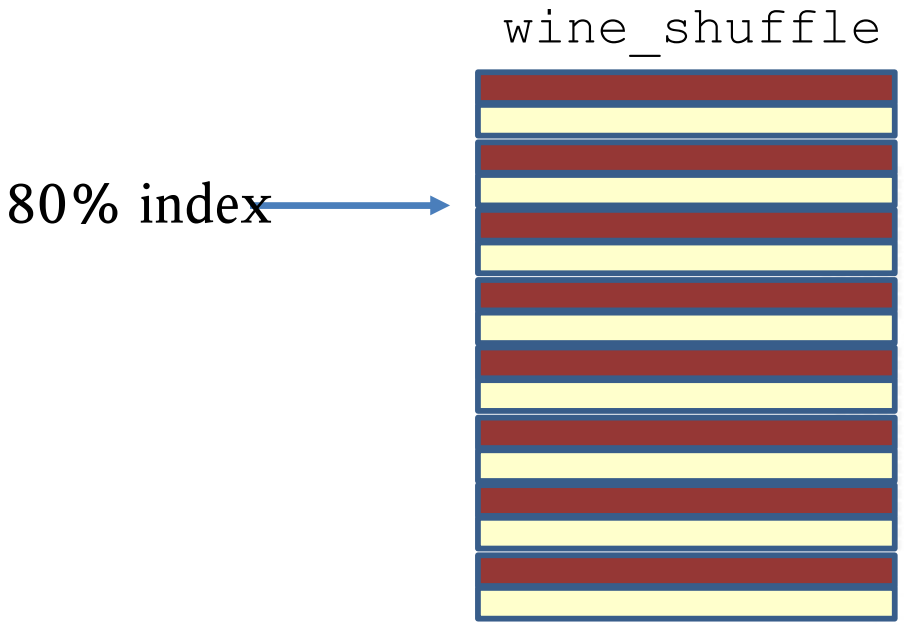


학습을 위한 훈련, 검증, 테스트 데이터셋

아래와 같은 비율로 나누기 위해 먼저 numpy로 자료형 변환

```
train_idx = int(len(wine_np) * 0.8)
```

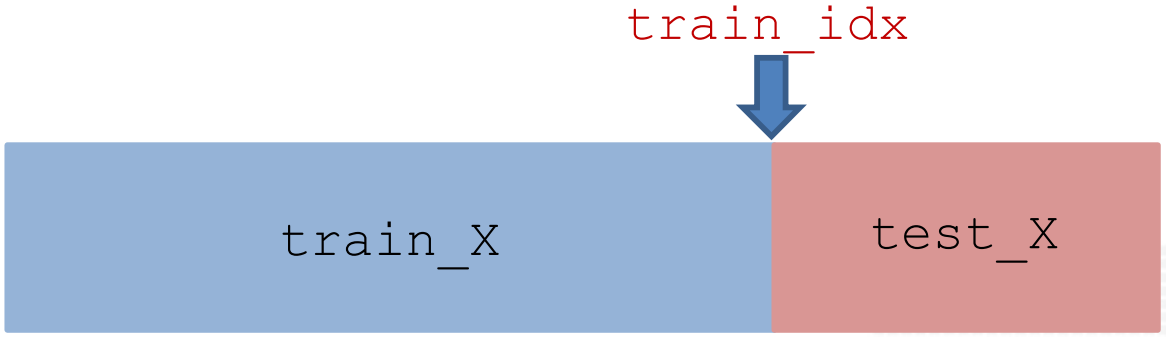
전체 개수 * 0.8 = 80%의 위치 index를 가져올 수 있음



학습을 위한 훈련, 검증, 테스트 데이터셋

아래와 같은 비율로 나누기 위해 먼저 numpy로 자료형 변환

```
train_X, train_Y = wine_np[:train_idx, :-1], wine_np[:train_idx, -1]
```



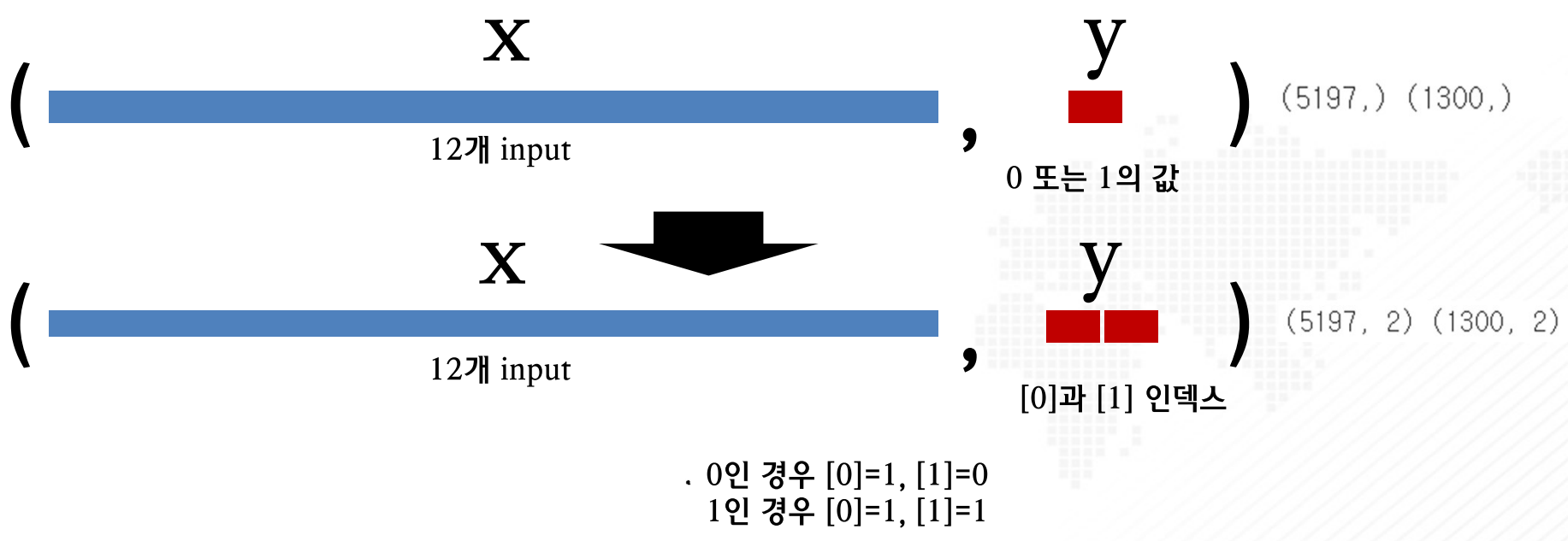

```
test_X, test_Y = wine_np[train_idx:, :-1], wine_np[train_idx:, -1]
```

One-hot 인코딩 벡터로 변환

한국외국어대학교 HANKUK UNIVERSITY OF FOREIGN STUDIES

카테고리가 2개인 문제임 -> 각 카테고리에 해당하는 클래스만 1이 되도록 해줌.

```
train_Y = tf.keras.utils.to_categorical(train_Y, num_classes =2)
test_Y = tf.keras.utils.to_categorical(test_Y, num_classes =2)
```

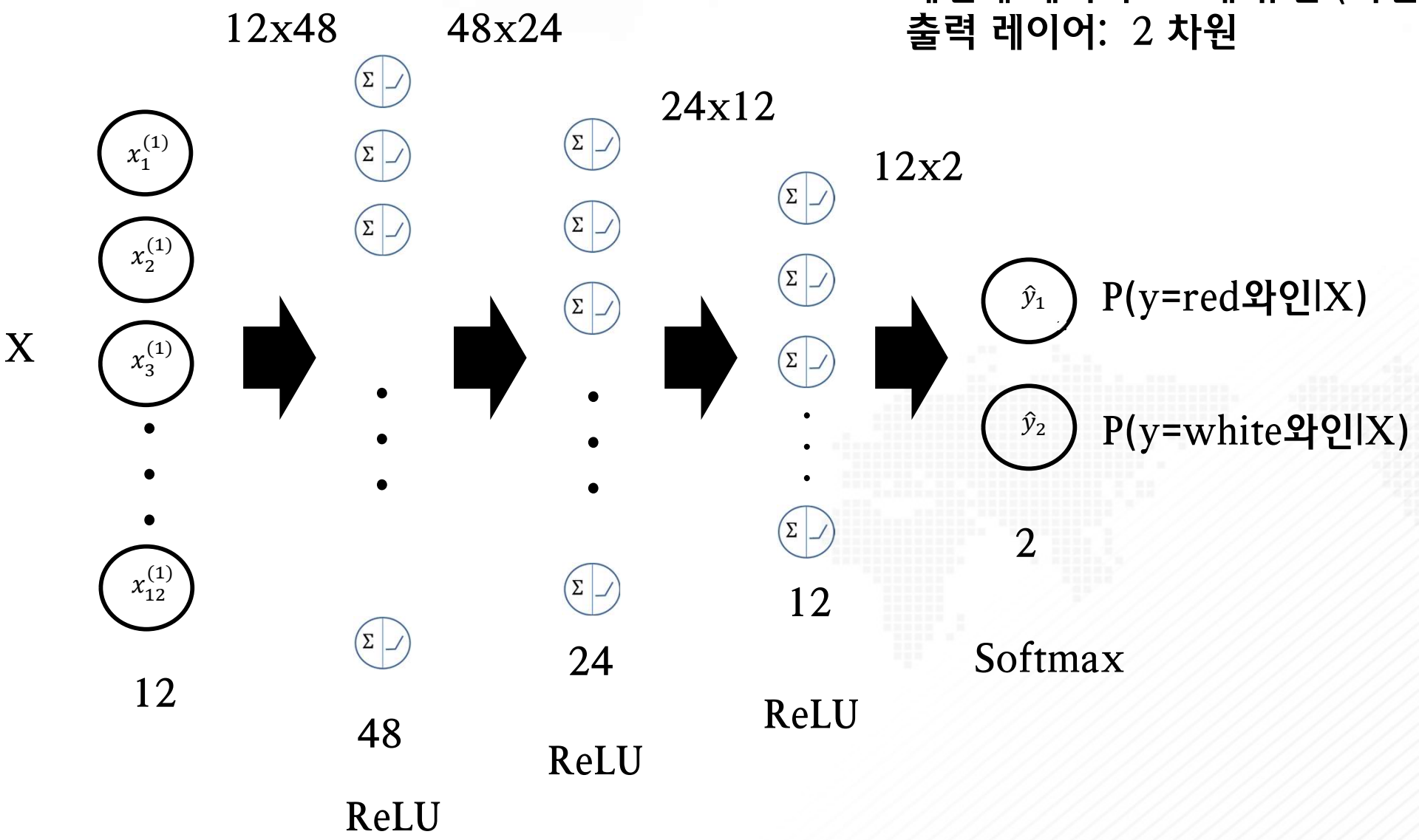


. 0인 경우 [0]=1, [1]=0
1인 경우 [0]=1, [1]=1

사용된 뉴럴 네트워크

한국외국어대학교 HANKUK UNIVERSITY OF FOREIGN STUDIES

입력 레이어: 12차원
첫번째 레이어: 48개 뉴런 (차원)
두번째 레이어: 24개 뉴런 (차원)
세번째 레이어: 12개 뉴런 (차원)
출력 레이어: 2 차원



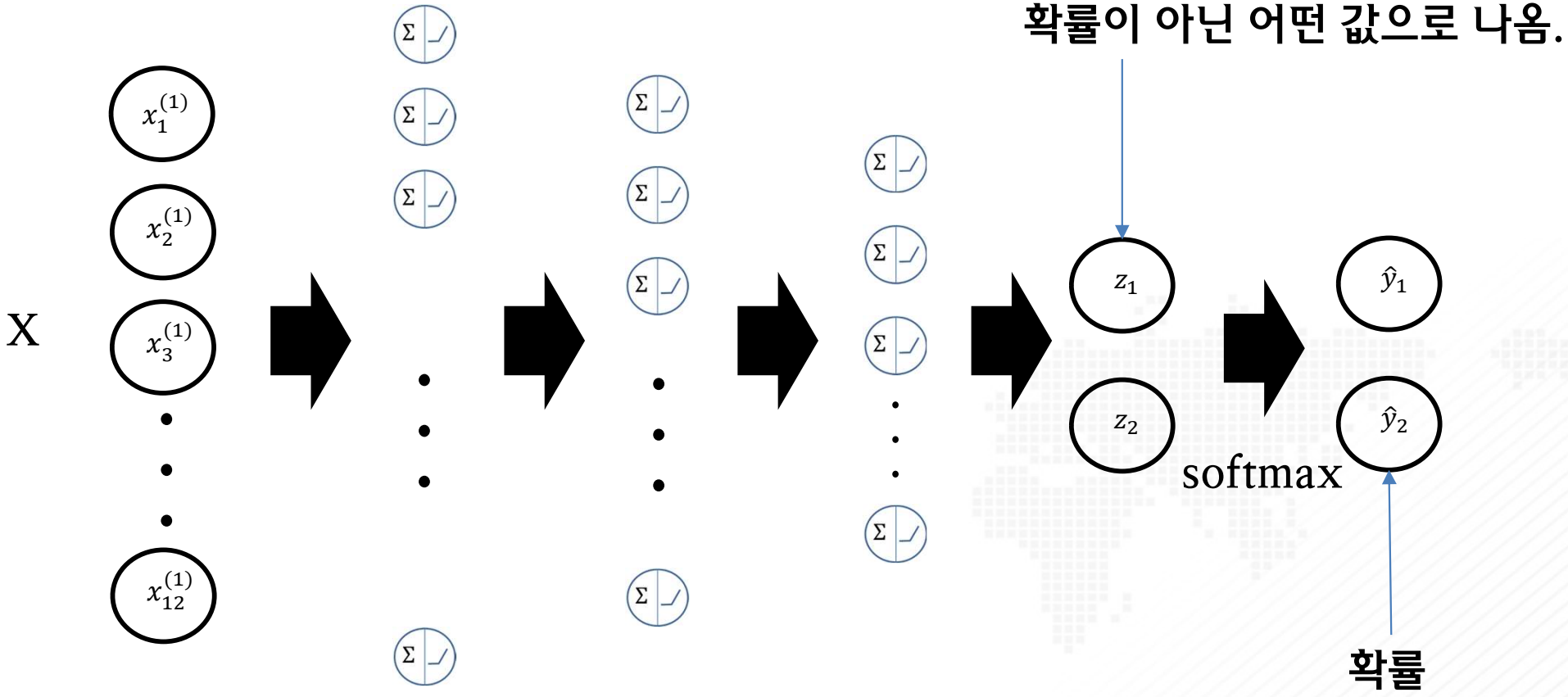
하이퍼파라메터

학습을 위한 전체적인 설정에 필요한 값들을 의미

* 전체 데이터를 몇 번 학습할 것인가? epochs = 25
* 한번 학습할 때 어느만큼의 부분데이터를 학습할 것인가? batch = 32
* 학습율은 어떻게 설정할 것인가? learning rate = 0.07

Softmax

output 클래스(카테고리)별 확률을 구할 때 사용



Softmax

output 클래스(카테고리)별 확률을 구할 때 사용

- 자연상수 e를 밑으로 하는 지수에 logits, z_j 를 넣어주어 값들의 특성을 더욱 극대화 시킨다.
- 큰 값은 더 강조되고, 작은 값은 상대적으로 약화시키는 효과를 가져옴.
- 확률특성상 $\sum_{j=1}^{J} \sigma(z)_j = 1$

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{k} e^{z_k}} \; for\; j = 1, \ldots, k$$

Categorical Cross Entropy

한국외국어대학교 HANKUK UNIVERSITY OF FOREIGN STUDIES

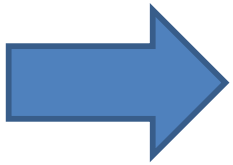
Classification문제에서 error (loss) function 기능으로 사용가능

- Cross Entropy는 정보이론에서 가져온 개념
- Entropy를 줄인다는 의미 -> 우리가 정답클래스만 1로 설정해 놓은 one hot vector로 setup된 확률에 기반하여 entropy를 낮추게 됨. -> 추정된 정답 확률이 높아지도록 학습 가능함.
- 여러 개의 카테고리들을 classification할 때 학습이 더 잘된다고 알려져 있음.

정답 레이블 확률

$$CCE = -\frac{1}{J}\sum p(\mathrm{x})\log q(x)$$

추론 확률



정답 레이블 확률

$$CCE = -\frac{1}{J}\sum \mathrm{y}\log \hat{y}$$

추론 확률