〈모두의 딥러닝〉

CH12. 다중 분류 문제 해결하기

3조 길태호, 심재성, 양지윤, 이용빈

TABLE OF CONTENTS

01

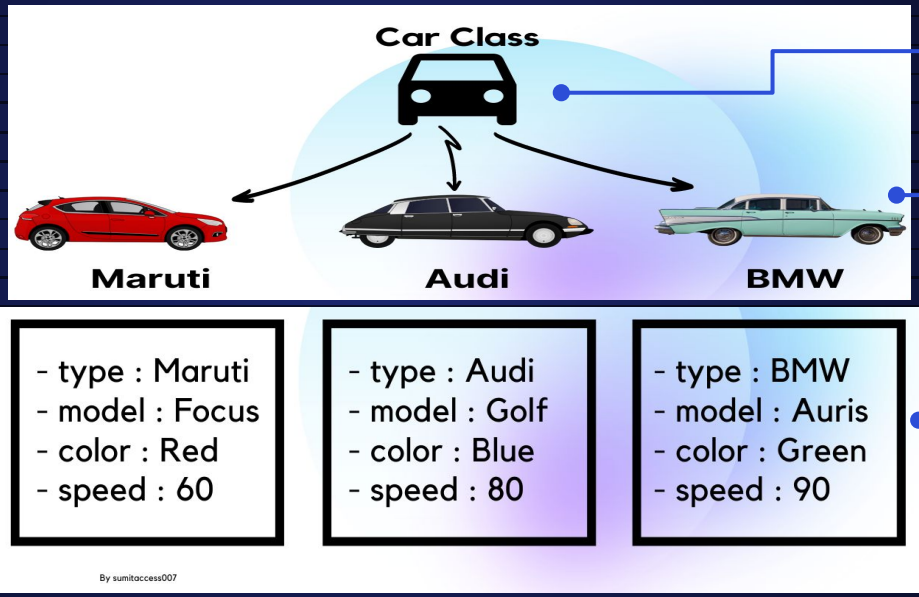
Multi Classification

다중 분류에 대한 개념 설명

Sample / Class / Attribute

〈속성 Attribute〉
1. 클래스 속성
-〉 모든 클래스에 동일하게 영향 미침

2. 인스턴스 속성
-〉 self를 이용해 생성된 인스턴스 내에서만 영향 미침
ex) `def __init__(self)`



샘플

클래스

(클래스) 속성

이항 분류 VS. 다중 분류

이항 분류

클래스가 2개인 분류
-둘 중에 하나를 고름

다중 분류

클래스가 3개 이상인 분류
-여러 개 중에 어떤 것이 답인지 예측함

Kaggle 실습 따라해보기
<날씨 다중 분류 문제>

Kaggle 실습 순서

Stage 1

Kaggle에서 데이터 보기
-5개의 클래스 분류
(흐림, 안개, 비, 해, 일출)

Stage 2

API를 이용하여 Kaggle 데이터를 Colab으로 다운로드
*API(Application Programming Interface):
두 소프트웨어 구성 요소가 서로 통신할 수 있게 하는 메커니즘

<Pre-class No.4 문제 해설>

Q4. 다음 중 옳지 않은 것을 모두 고르시오. *

- 샘플이 '어떤 것'인지 예측할 때, '어떤 것'에 해당하는 것이 클래스이다.
- 다중 분류는 클래스가 2개 이상일 때 이용된다.
- 상관도 그래프를 통해 속성들이 샘플에 따라 다른 것을 알 수 있다.
- 하나의 클래스에는 하나의 속성만 있다.
- 다중 분류와 이항 분류를 접근하는 방식은 다르다.

02

상관도 그래프

상관도 그래프를 통해 프로젝트 감을 잡고 전략 세우기

Seborn dataset

Michael Waskom

mwaskom

mwaskom Merge pull request #27 from koenv... ... 2b29313 2 weeks ago **46** commits

png	More updates	last year
process	Remove one-off 2021 datapoint from healthexp ...	5 months ago
raw	Add dowjones dataset	5 months ago
README.md	Add dowjones dataset	5 months ago
anagrams.csv	Rename messy anagrams dataset	2 years ago
anscombe.csv	Add anscombe dataset	9 years ago
attention.csv	Add attention dataset	9 years ago
brain_networks.csv	Add brain networks dataset	8 years ago
car_crashes.csv	Add 538 car crash dataset	8 years ago
dataset_names.txt	Add a file containing all available dataset names	2 weeks ago
diamonds.csv	Add diamonds dataset	4 years ago
dots.csv	Add dots dataset	5 years ago
dowjones.csv	Add dowjones dataset	5 months ago
exercise.csv	Add exercise dataset	9 years ago
flights.csv	Add flights dataset	8 years ago
fmri.csv	Change sorting of events in fmri data	5 years ago
geyser.csv	Add geyser dataset	2 years ago
glue.csv	Add several new datasets	5 months ago
healthexp.csv	Remove one-off 2021 datapoint from healthexp ...	5 months ago
iris.csv	Add iris dataset	8 years ago
mpg.csv	Add mpg dataset	4 years ago
penguins.csv	Change culmen to bill in penguins dataset	2 years ago
planets.csv	Add planets dataset	9 years ago
seaice.csv	Add several new datasets	5 months ago
taxis.csv	Add green taxis to the taxis dataset	last year
tips.csv	Add tips dataset	9 years ago
titanic.csv	Update titanic datset to remove index variable	9 years ago

실제 발표는 코랩으로.

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

df = sns.load_dataset('penguins')
df
```

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0	Male
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0	Female
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0	Female
3	Adelie	Torgersen	NaN	NaN	NaN	NaN	NaN
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0	Female
...	...	...	...	...	...	...	...
339	Gentoo	Biscoe	NaN	NaN	NaN	NaN	NaN
340	Gentoo	Biscoe	46.8	14.3	215.0	4850.0	Female
341	Gentoo	Biscoe	50.4	15.7	222.0	5750.0	Male
342	Gentoo	Biscoe	45.2	14.8	212.0	5200.0	Female
343	Gentoo	Biscoe	49.9	16.1	213.0	5400.0	Male

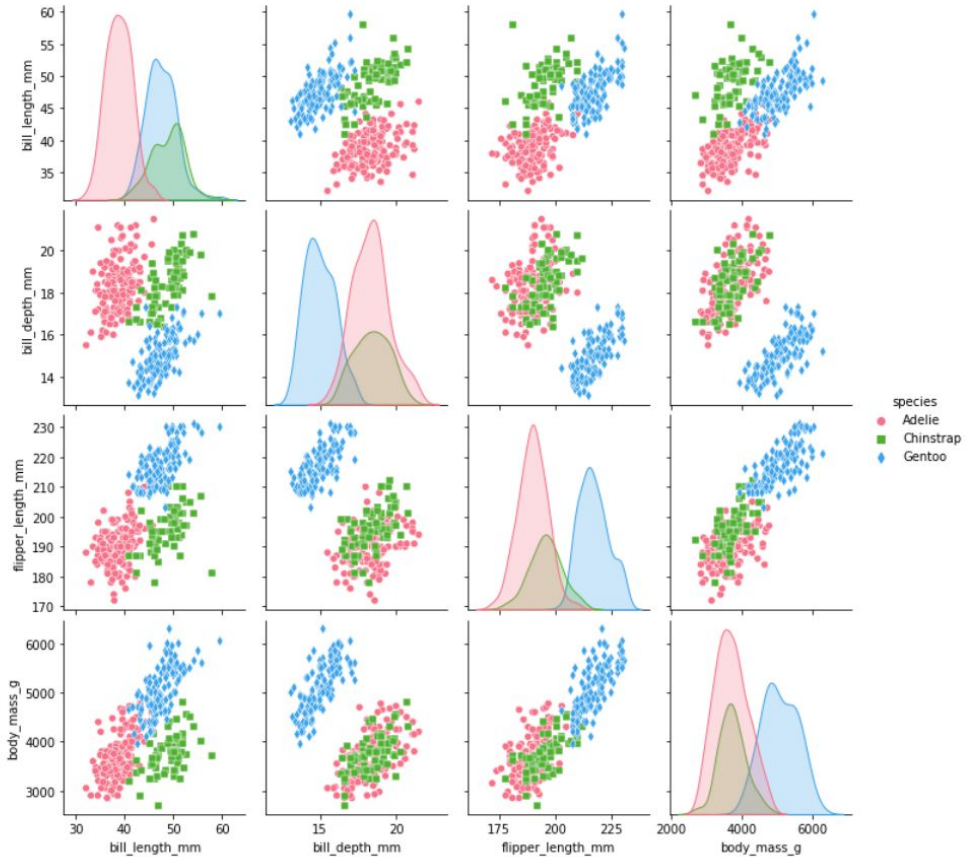
344 rows × 7 columns

```
y1 = df['species']
print(y1.unique())
```

```
['Adelie' 'Chinstrap' 'Gentoo']
```

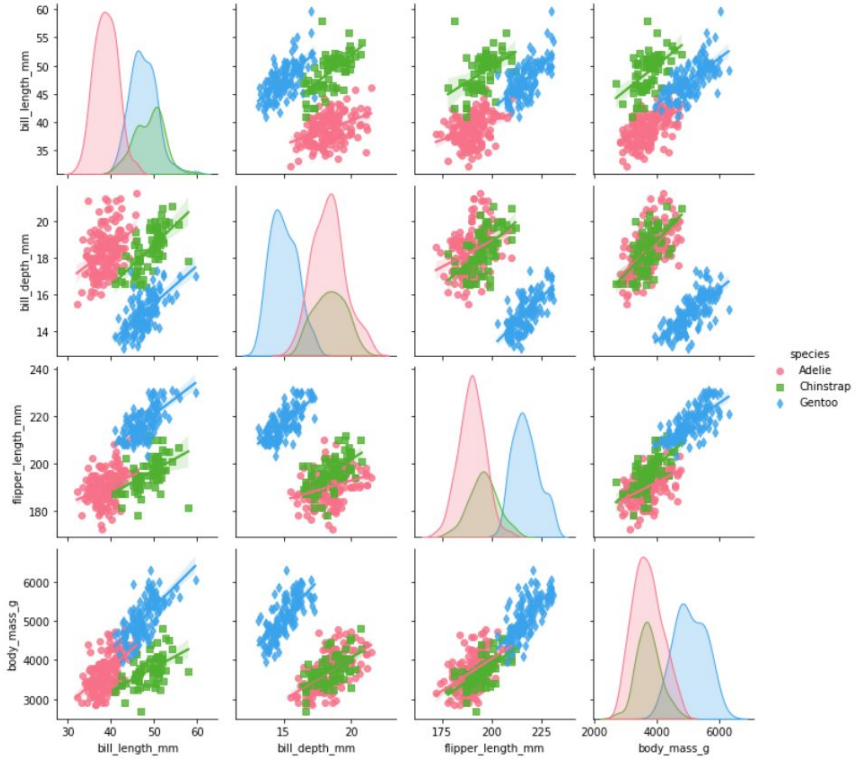
```
sns.pairplot(df,hue ='species', palette ='husl', markers=['o','s','d'])
```

```
<seaborn.axisgrid.PairGrid at 0x7f32bbe18df0>
```



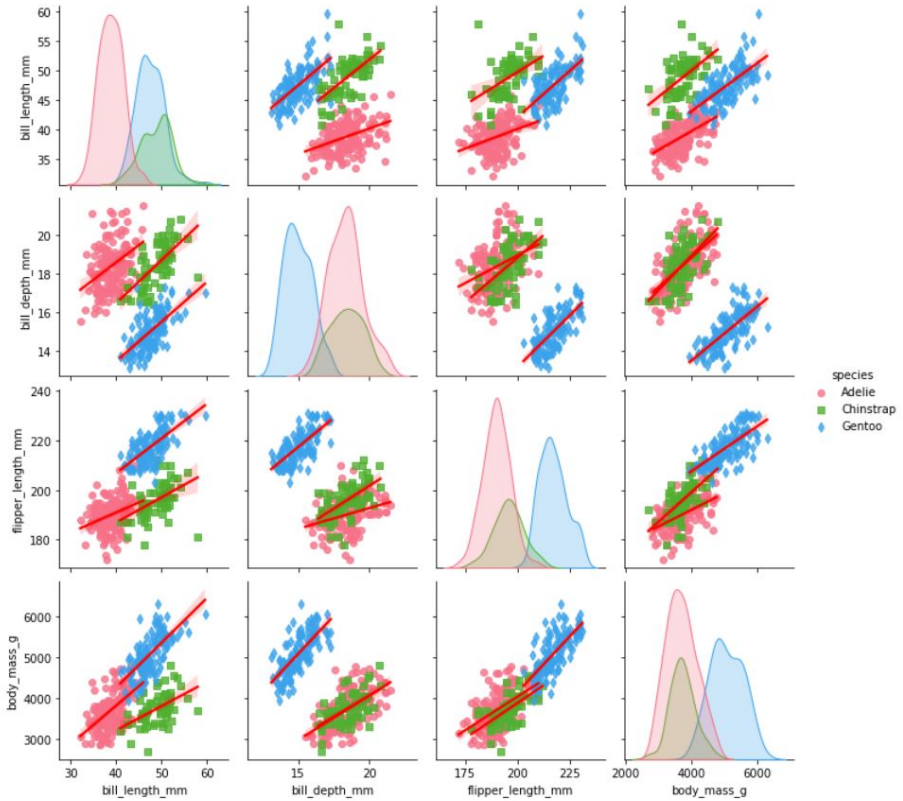
```
sns.pairplot(df,hue ='species', palette ='husl', markers=['o','s','d'], kind ='reg')
```

```
<seaborn.axisgrid.PairGrid at 0x7f32b9bb6c40>
```



```
sns.pairplot(df,hue ='species', palette ='husl', markers=['o','s','d'], kind ='reg',plot_kws={'line_kws' :{'color':'red'}})
```

```
<seaborn.axisgrid.PairGrid at 0x7f32b9bad730>
```



03

원-핫 인코딩

원-핫 인코딩의 개념과 한계점 & 관련 함수

1. 원-핫 인코딩이란?

COLOR
RED
BLUE
GREEN

RED	BLUE	GREEN
1	0	0
0	1	0
0	0	1

레이블 인코딩

LabelEncoder( ) 함수

COLOR
RED
BLUE
GREEN

COLOR	NUMBER
RED	1
BLUE	2
GREEN	3

원-핫 인코딩

Categorical ( ) 함수

RED	BLUE	GREEN
1	0	0
0	1	0
0	0	1

2. 원-핫 인코딩의 한계

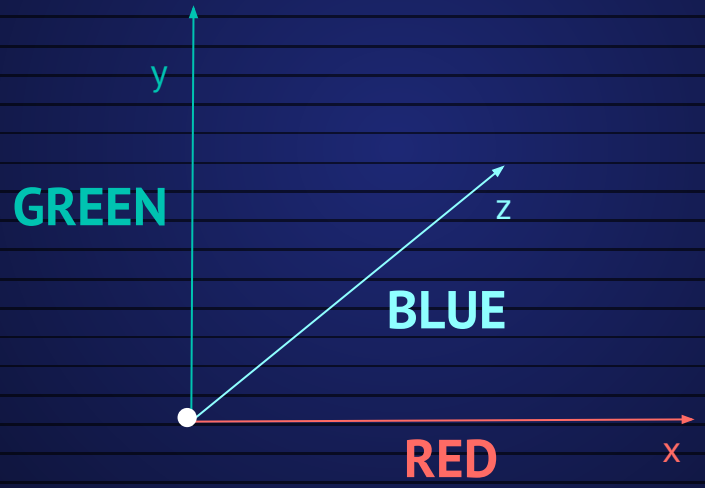
· 저장 공간 문제

COLOR
RED
BLUE
GREEN
⋮
YELLOW

RED	BLUE	GREEN	• • •	YELLOW
1	0	0	• • •	0
0	1	0	• • •	0
0	0	1	• • •	0
⋮	⋮	⋮	⋱	⋮
0	0	0	• • •	1

2. 원-핫 인코딩의 한계

· 유사도 표현 문제



3. LabelEncoder( ) 함수

레이블 인코딩

LabelEncoder( ) 함수

COLOR
RED
BLUE
GREEN

COLOR	NUMBER
RED	1
BLUE	2
GREEN	3

Method	의미
fit(y)	레이블 인코더를 fit
fit_transform(y)	Fit 시킨 후 인코드 레이블 리턴
get_params([deep])	추정량의 파라미터 구하기
inverse_transform(y)	레이블을 원래 인코딩으로 변환
set_output(*[,transform])	출력 컨테이너 정하기
set_params(**params)	추정량의 파라미터 구하기
transform(y)	레이블을 정상화한 인코딩으로 변환

4. Categorical () 함수

원-핫 인코딩

Categorical ( ) 함수

COLOR	NUMBER
RED	1
BLUE	2
GREEN	3

RED	BLUE	GREEN
1	0	0
0	1	0
0	0	1

· tf.keras.utils.to_categorical(y, num_classes=None, dtype='float32')

함수의 인자	의미
y	Array 형태의 클래스값을 행렬로 변환
num_classes	전체 클래스의 수 (None이면 max(y)+1)
dtype	입력 데이터 형태 (Default: float32)

5. Pre-Class 5번 문제

Q5. 원-핫 인코딩에 대한 다음 내용 중 옳지 않은 것을 고르시오.

1. 원-핫 인코딩은 여러 개의 Y 값을 0과 1로만 이루어진 형태로 바꿔주는 기법이다.

2.Y값이 문자열인 경우, 문자열을 숫자로 바꾸기 위해 LabelEncoder () 함수를 사용한다.

3.Y값을 숫자 0과 1로만 이루어지도록 하기 위해 tf.keras.utils.categorical() 함수를 사용한다.

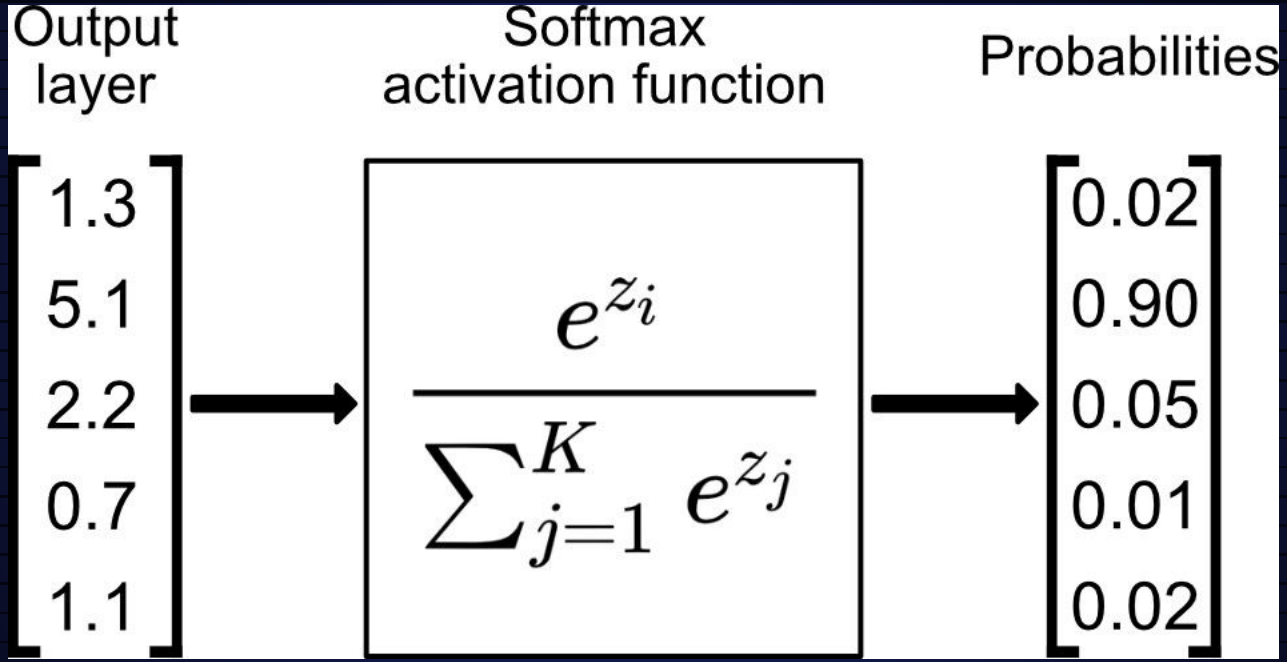
4.tf.keras.utils.categorical() 함수는 array ([[1. , 0.] ,[0. ,1.]]) 를 array([1,2]) 로 바꾸는 과정이다.

04

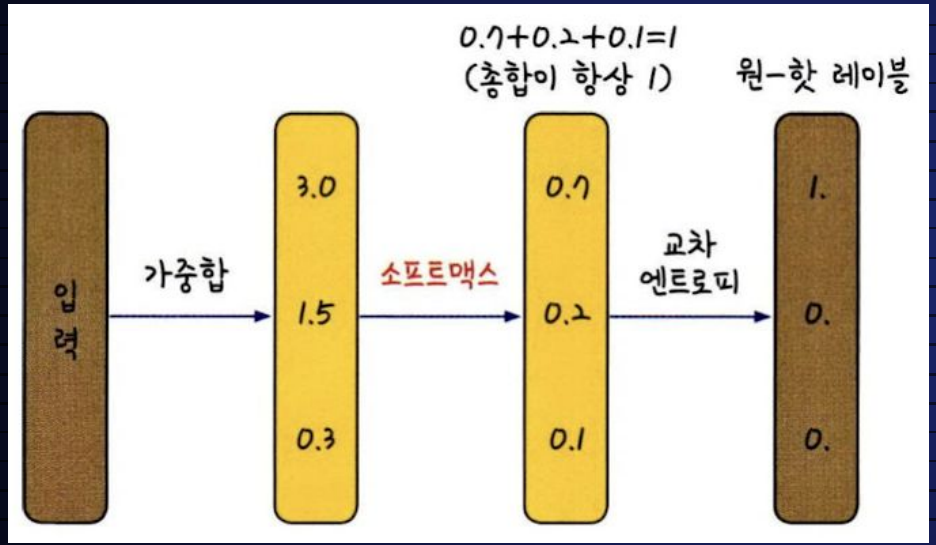
Softmax

출력을 확률로 바꾸어주는 함수

softmax

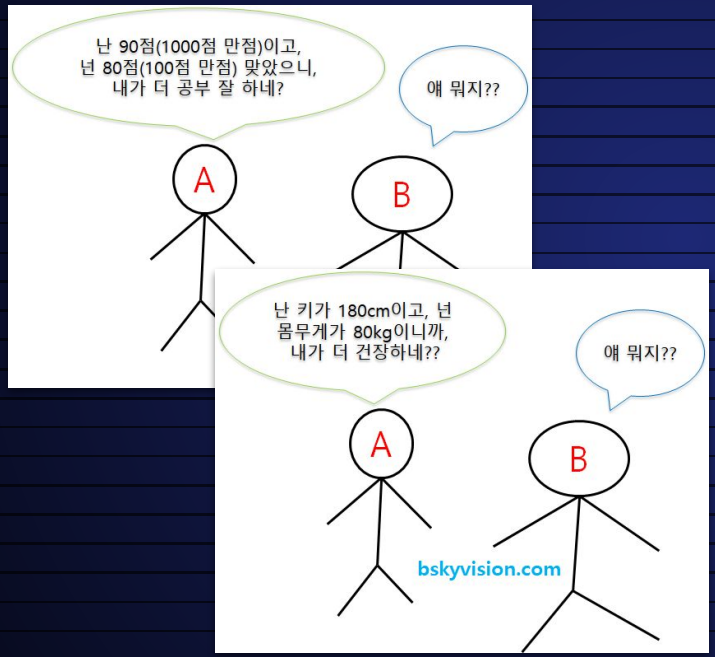

Output layer
Softmax activation function
Probabilities
1.3
5.1
2.2
0.7
1.1
$\frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$
0.02
0.90
0.05
0.01
0.02

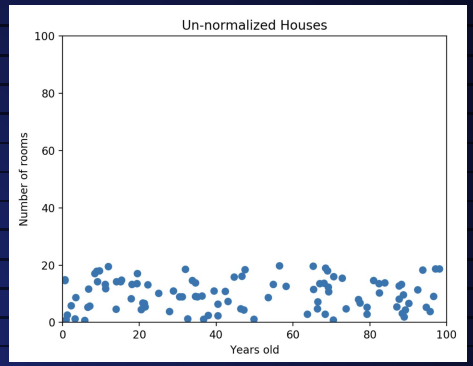
softmax

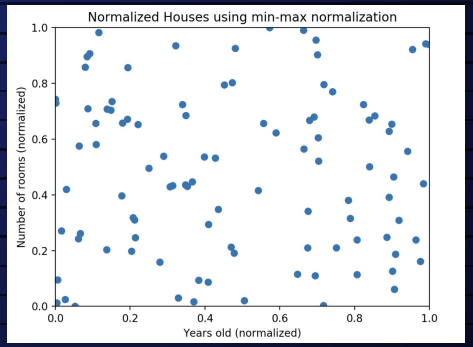


$$\frac{3}{4.8} = 0.625$$

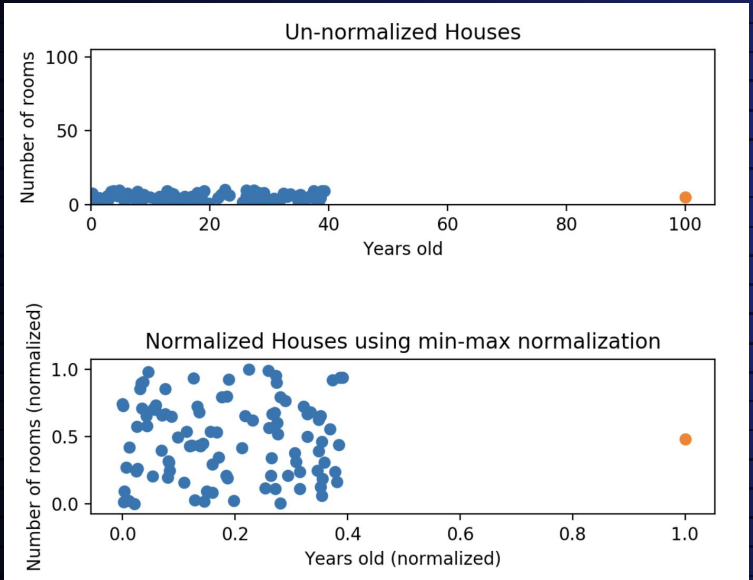
데이터 전처리

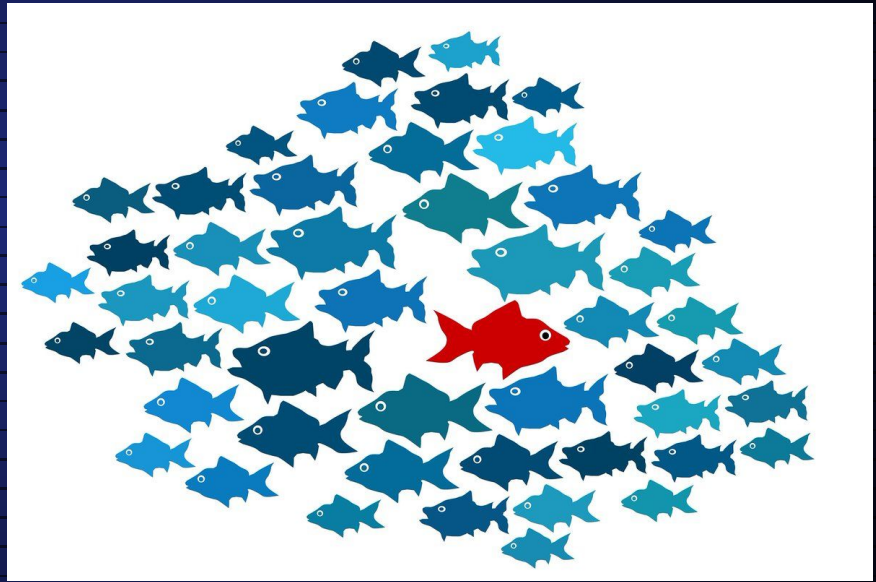

난 90점(1000점 만점)이고,
넌 80점(100점 만점) 맞았으니,
내가 더 공부 잘 하네?
얘 뭐지??
A
B
난 키가 180cm이고, 넌
몸무게가 80kg이니까,
내가 더 건장하네??
얘 뭐지??
A
B
bskyvision.com


Un-normalized Houses
Number of rooms
Years old
0
20
40
60
80
100

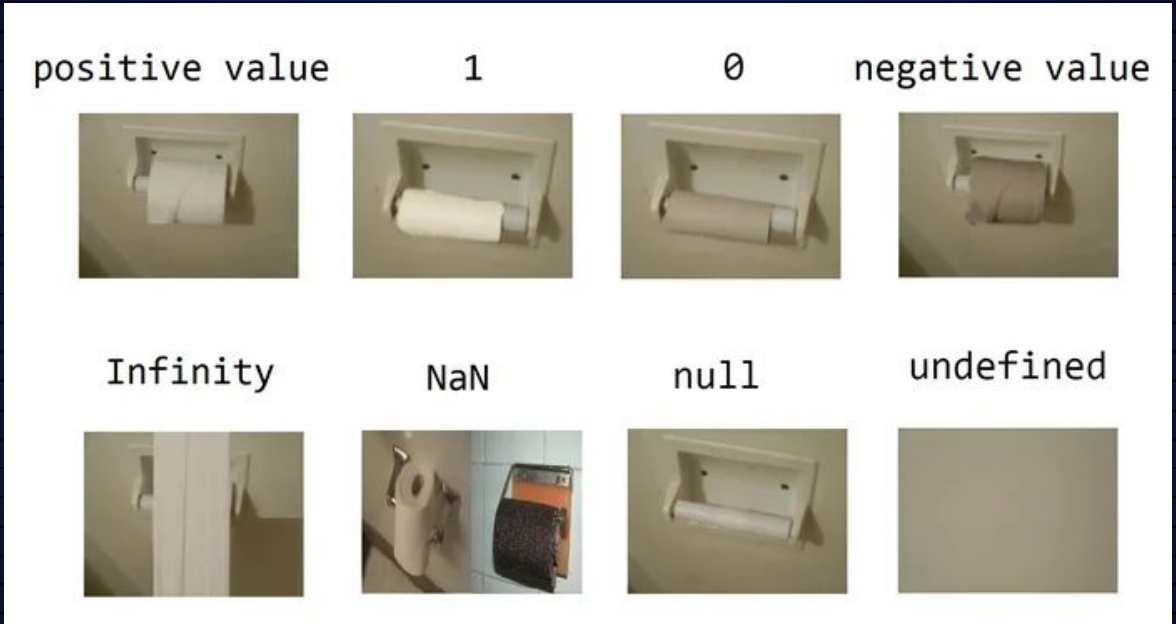

Normalized Houses using min-max normalization
Number of rooms (normalized)
Years old (normalized)
0.0
0.2
0.4
0.6
0.8
1.0

outlier


Un-normalized Houses
Number of rooms
100
50
0
0
20
40
60
80
100
Years old
Normalized Houses using min-max normalization
Number of rooms (normalized)
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Years old (normalized)



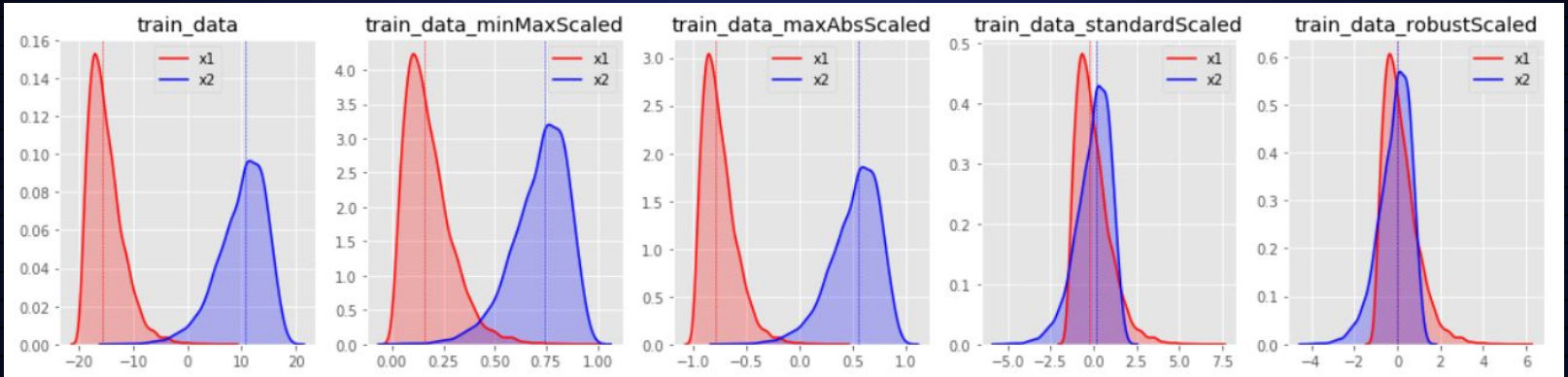
Null, NaN, Na



NA : Not Available

NaN: Not a Number

Scaler (표준화, 정규화)



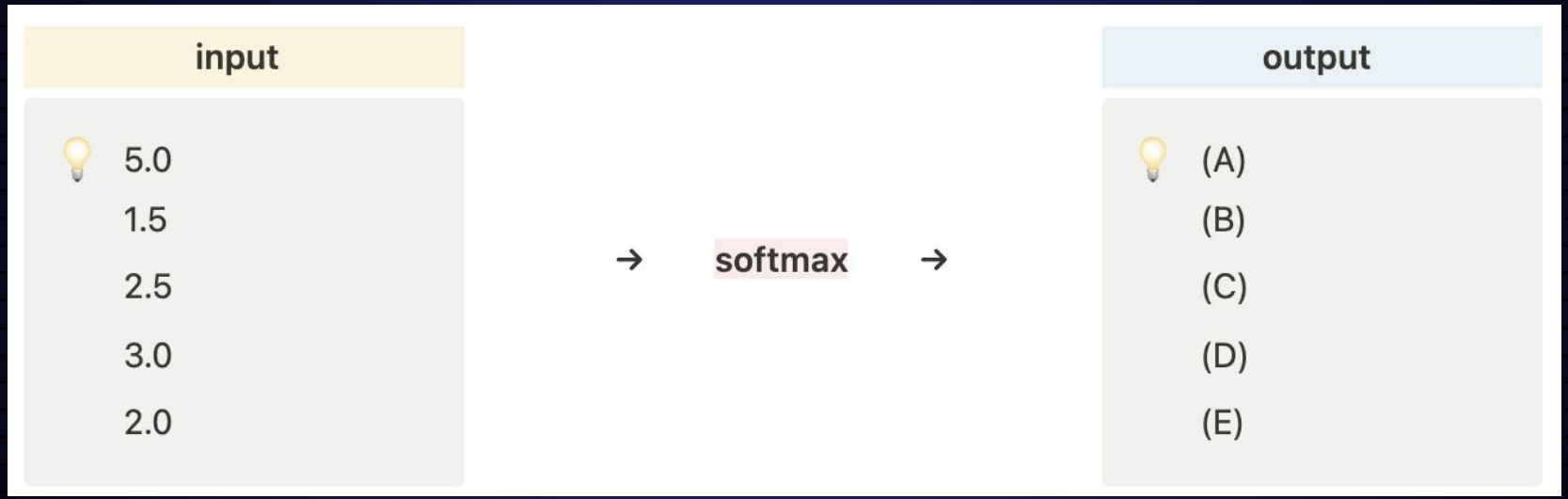
minMaxScaler: 0 ~ 1

maxAbsScaler: -1 ~ 1

standardScaler: 평균이 0 분산이 1로 변환

robustScaler: median 사용 (outlier 최소화)

Quiz 6: 소프트맥스 함수의 출력



정답: (D) = 0.21

05

아이리스 품종 예측 문제

You can enter a subtitle here if you need it

코랩으로 실습하기

06
실습 과제

실습과제에 대한 설명

Thank you!
감사합니다:)