넷째마당
딥러닝 기본기 다지기

12장 다중 분류 문제 해결하기

1 다중 분류 문제

2 상관도 그래프

3 원-핫 인코딩

4 소프트맥스

5 아이리스 품종 예측의 실행

1 다중 분류 문제

1 다중 분류 문제

- **다중 분류 문제**
  - 아이리스는 그 꽃봉오리가 마치 먹물을 머금은 붓과 같다 하여 우리나라에서는 '붓꽃'이라고 부르는 아름다운 꽃
  - 아이리스는 꽃잎의 모양과 길이에 따라 여러 가지 품종으로 나뉨
  - 사진을 보면 품종마다 비슷해 보이는데요. 과연 딥러닝을 사용해서 이들을 구별해 낼 수 있을까?

▼ **그림 12-1 | 아이리스의 품종**

Iris-virginica

Iris-setosa

Iris-versicolor

1 다중 분류 문제

- **다중 분류 문제**
  - 아이리스 품종 예측 데이터는 예제 파일의 data 폴더에서 찾을 수 있음(data/iris3.csv)
  - 데이터의 구조는 다음과 같음

▼ **그림 12-2 | 아이리스 데이터의 샘플, 속성, 클래스 구분**

속성 / 클래스

	정보 1	정보 2	정보 3	정보 4	품종
샘플	1번째 아이리스	5.1	3.5	4.0	0.2	Iris-setosa
	2번째 아이리스	4.9	3.0	1.4	0.2	Iris-setosa
	3번째 아이리스	4.7	3.2	1.3	0.3	Iris-setosa
	...	...	...	...	...	...
	150번째 아이리스	5.9	3.0	5.1	1.8	Iris-virginica

1 다중 분류 문제

- 샘플 수: 150
- 속성 수: 4
  - 정보 1: 꽃받침 길이(sepal length, 단위: cm)
  - 정보 2: 꽃받침 너비(sepal width, 단위: cm)
  - 정보 3: 꽃잎 길이(petal length, 단위: cm)
  - 정보 4: 꽃잎 너비(petal width, 단위: cm)
- 클래스: Iris-setosa, Iris-versicolor, Iris-virginica

1 다중 분류 문제

- **다중 분류 문제**
  - 속성을 보니 우리가 앞서 다루었던 것과 중요한 차이가 있음
  - 바로 클래스가 두 개가 아니라 세 개
  - 즉, 참(1)과 거짓(0)으로 해결하는 것이 아니라, 여러 개 중에 어떤 것이 답인지 예측하는 문제
  - 이렇게 여러 개의 답 중 하나를 고르는 분류 문제를 **다중 분류**(multi classification)라고 함
  - 다중 분류 문제는 둘 중에 하나를 고르는 이항 분류(binary classification)와는 접근 방식이 조금 다름
  - 지금부터 아이리스 품종을 예측하는 실습을 통해 다중 분류 문제를 해결해 보자

2 상관도 그래프

2 상관도 그래프

- **상관도 그래프**
  - 먼저 데이터의 일부를 불러와 내용을 보자

```
import pandas as pd

# 아이리스 데이터를 불러옵니다.
df = pd.read_csv('./data/iris3.csv')

df.head() # 첫 다섯 줄을 봅니다.
```

2 상관도 그래프

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

- 샘플 수: 150
- 속성 수: 4
  - 정보 1: 꽃받침 길이(sepal length, 단위: cm)
  - 정보 2: 꽃받침 너비(sepal width, 단위: cm)
  - 정보 3: 꽃잎 길이(petal length, 단위: cm)
  - 정보 4: 꽃잎 너비(petal width, 단위: cm)
- 클래스: Iris-setosa, Iris-versicolor, Iris-virginica

Iris-virginica

Iris-setosa

Iris-versicolor

2 상관도 그래프

- **상관도 그래프**
  - 이번에는 시본(seaborn) 라이브러리에 있는 pairplot() 함수를 써서 전체 상관도를 볼 수 있는 그래프를 출력해 보자

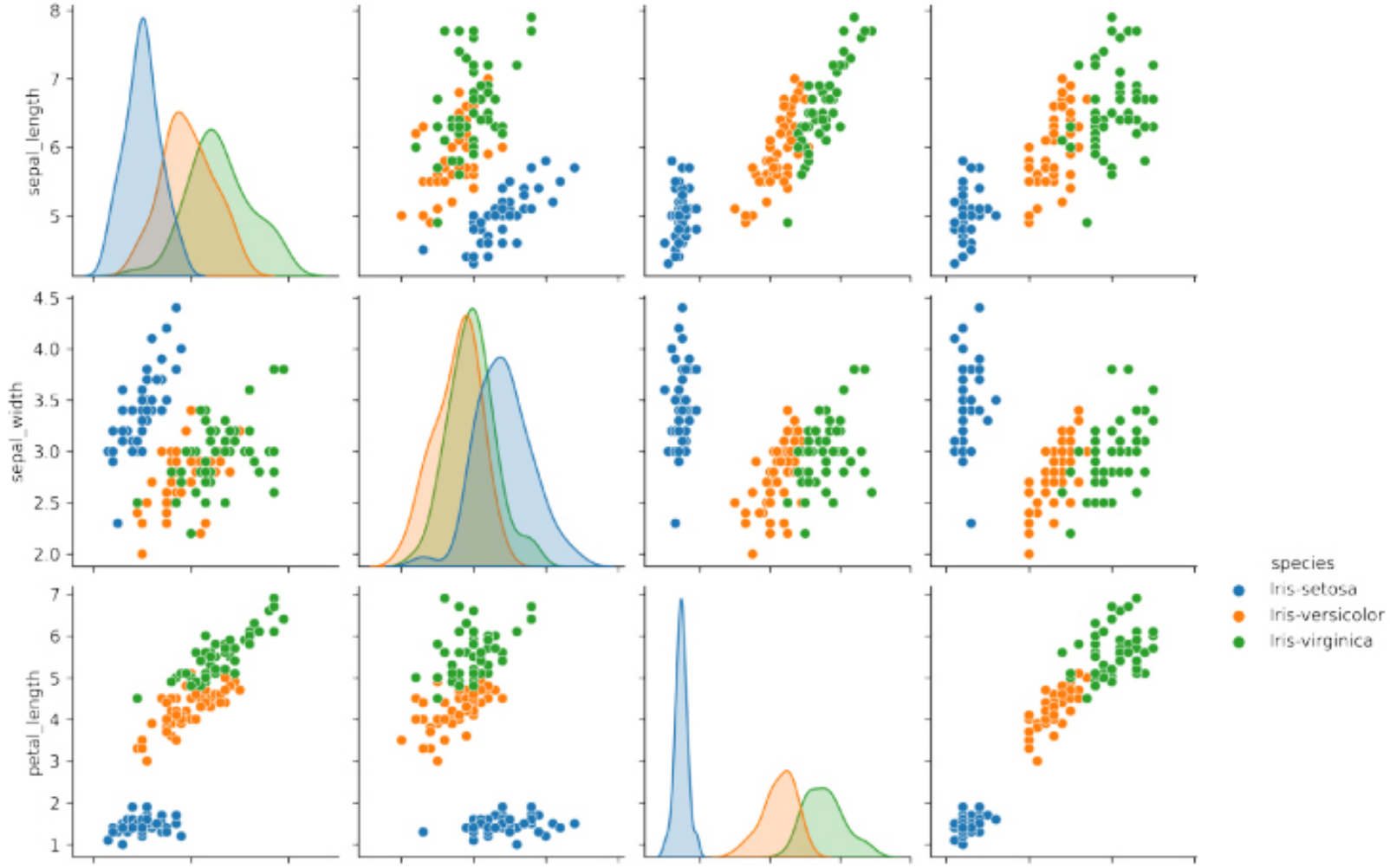
```
import seaborn as sns
import matplotlib.pyplot as plt

sns.pairplot(df, hue='species');
plt.show()
```

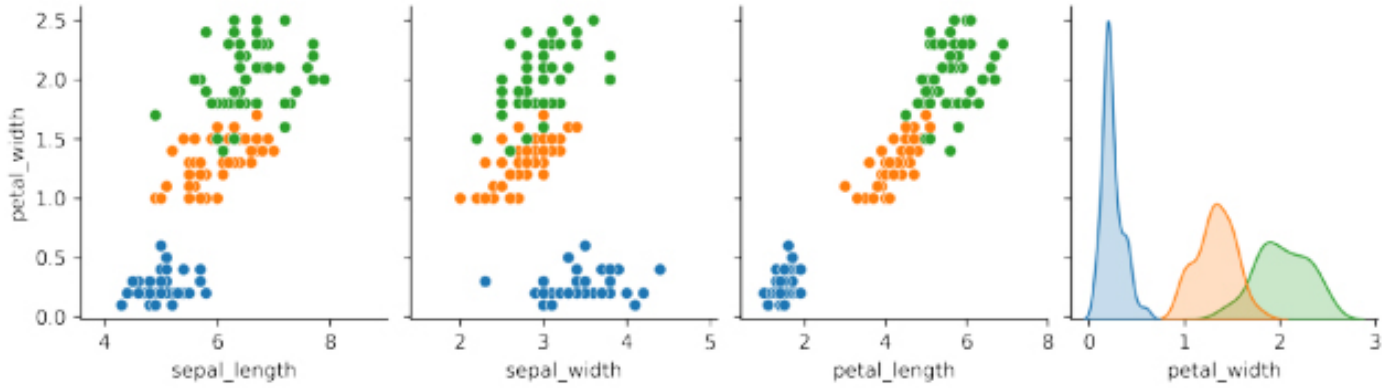
❶

2 상관도 그래프

▼ 그림 12-3 | pairplot 함수로 데이터 한번에 보기



2 상관도 그래프


petal_width
sepal_length
sepal_width
petal_length
petal_width

3 원-핫 인코딩

- **원-핫 인코딩**
  - 이제 케라스를 이용해 아이리스의 품종을 예측해 보자
  - Iris-setosa, Iris-virginica 등 데이터 안에 문자열이 포함되어 있음
  - 먼저 조금 전 불러온 데이터 프레임을 X와 y로 나누겠음

```
X = df.iloc[:,0:4]
y = df.iloc[:,4]
```

3 원-핫 인코딩

- **원-핫 인코딩**
  - 여러 개의 값으로 된 문자열을 0과 1로만 이루어진 형태로 만들어 주는 과정을 **원-핫 인코딩**(one-hot encoding)이라고 함
  - 판다스가 제공하는 get_dummies() 함수를 사용하면 간단하게 해낼 수 있음

```
# 원-핫 인코딩 처리를 합니다.
y = pd.get_dummies(y)

# 원-핫 인코딩 결과를 확인합니다.
print(y[0:5])
```

❷

species	setosa	versicolor	virginica
setosa	1	0	0
versicolor	0	1	0
virginica	0	0	1
versicolor	0	1	0
...	...	...	...

❶ ❸

3 원-핫 인코딩

- 원-핫 인코딩

실행 결과

	Iris-setosa	Iris-versicolor	Iris-virginica
0	1	0	0
1	1	0	0
2	1	0	0
3	1	0	0
4	1	0	0

4 소프트맥스


```
# 모델 설정
model = Sequential()
model.add(Dense(12, input_dim=4, activation='relu'))
model.add(Dense(8, activation='relu'))
model.add(Dense(3, activation='softmax'))
model.summary()

# 모델 컴파일
model.compile(loss='categorical_crossentropy', optimizer='adam',
metrics=['accuracy'])
```

- 세 가지가 달라졌음
- 첫째 출력층의 노드 수가 3으로 바뀜
- 활성화 함수가 softmax로 바뀜
- 마지막으로 컴파일 부분에서 손실 함수 부분이 categorical_crossentropy로 바뀜

5 아이리스 품종 예측의 실행

- 아이리스 품종 예측의 실행

실행 결과

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, 12)                60
_________________________________________________________________
dense_1 (Dense)              (None, 8)                 104
_________________________________________________________________
dense_2 (Dense)              (None, 3)                 27
=================================================================
```

4 소프트맥스

▼ 그림 12-5 | 소프트맥스

샘플		setosa일 확률	versicolor일 확률	virginica일 확률
1번 샘플	예측 실행	0.2	0.7	0.1
2번 샘플		0.8	0.1	0.1
3번 샘플		0.2	0.2	0.6