Fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

선수 지식 - 통계
확률 분포의 추정

확률 분포의 추정 | 딥러닝의 기초가 되는 확률 개념 알아보기

강사 나동빈

fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

선수 지식 - 통계

확률 분포의 추정

fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

확률 분포 추정이 필요한 이유

- 우리가 확률 분포를 이미 알고 있다면, 확률 변수를 넣어 확률 값을 구할 수 있다.
- 실제 확률 분포가 다음과 같다고 해보자.
- $P(X = 1) = 35\%$
- $P(X = 2) = 15\%$
- $P(X = 3) = 50\%$
- 확률 변수 X의 값이 3일 확률은 50%이다.

→ 현실에서는 확률 분포 함수에서 나온 데이터만 얻을 수 있는 경우가 많다.

→ **내재된 확률 분포 함수**를 모를 때, 어떻게 추정할 수 있을까?

fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

확률 분포의 추정

- 우리가 가지고 있는 데이터로부터 **확률 분포**를 추정하는 기술을 의미한다.
- 우리는 결과적으로 확률 분포를 알고 싶으며, 내가 가지고 있는 데이터는 확률변수의 분포를 계산하기 위한 도구로 이해할 수 있다.

fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

확률 분포를 추정하는 기본적인 방법

- 기본적으로 우리가 데이터의 형태를 보고, 원하는 분포로 추정할 수 있다.
- **베르누이 분포**: 데이터가 0 혹은 1의 형태
- **정규 분포**: 데이터가 크기 제한이 없는 실수 형태
- **카테고리 분포**: 데이터가 카테고리 값 형태

fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

확률 분포를 추정하는 기본적인 방법

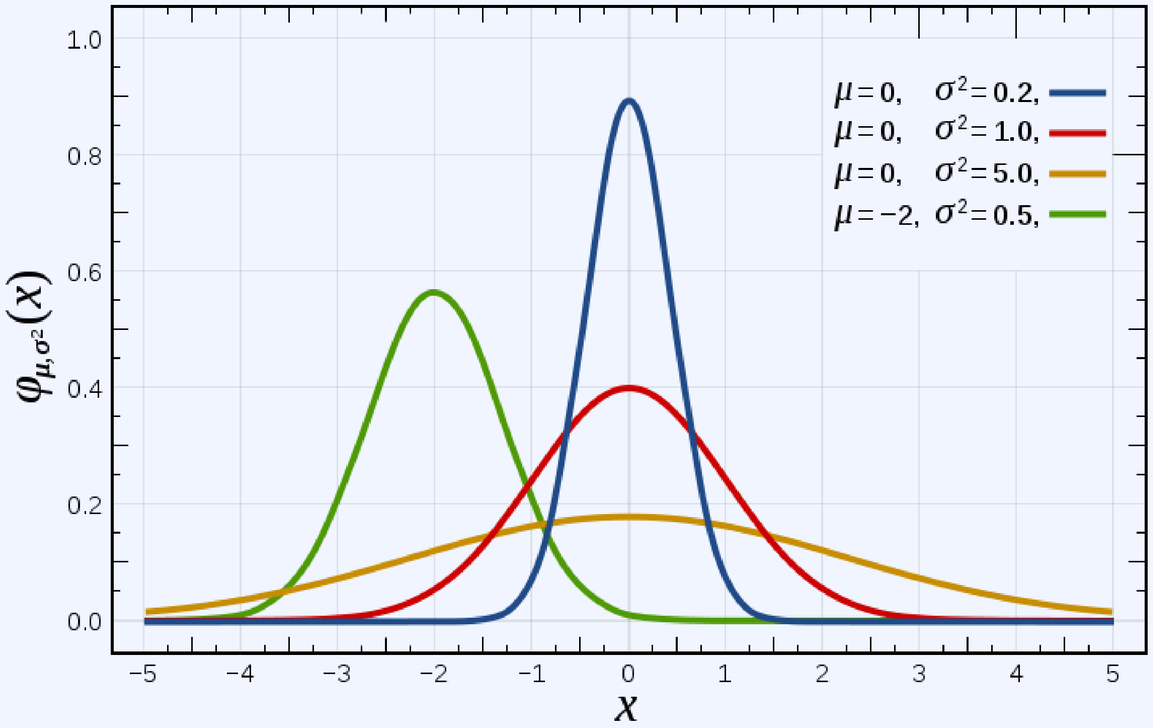
- 주어진 데이터를 이용해 확률 분포를 계산하는 대표적인 두 가지 방법이 존재한다.

1. 모멘트 방법
2. 최대 가능도 추정

fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

확률 분포를 나타내기 위한 모수(parameter)

- 우리는 지금까지 **모수(parameter)**를 가지는 확률 분포에 대해서 알아보았다.
- 정규 분포는 **평균**과 **분산** 두 가지 파라미터를 적절히 조합해, 다양한 정규 분포를 표현할 수 있다.



* https://ko.wikipedia.org/wiki/%EC%A0%95%EA%B7%9C_%EB%B6%84%ED%8F%AC

fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

모멘트(Moment) = 적률

- 확률분포에서 계산한 특징 값의 일종으로, n차 모멘트는 다음과 같이 정의된다.

$$M_n = E[X^n] = \int_{-\infty}^{\infty} x^n f(x) dx$$

- 1차 모멘트는 **평균**, 2차 모멘트는 **분산**에 해당한다.
- 1차부터 무한대 차수에 이르기까지 두 확률 분포의 모든 모멘트 값이 같다면?

→ 두 확률 분포는 같다.

fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

모멘트 방법(Method of Moment)

- 1차 모멘트는 데이터의 **평균**(mean)과 같다.

$$\mu = E[X] \triangleq \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- 2차 모멘트는 데이터의 **분산**(variance)과 같다.

$$\sigma^2 = E[(X-\mu)^2] \triangleq \frac{1}{N}\sum_{i=1}^{N} (x_i - \bar{x})^2$$

fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

모멘트 방법을 활용한 정규 분포 추정

- 모멘트 방법을 이용해 정규 분포의 $parameter$를 추정할 수 있다.
- 학생 성적이 7등급의 구간으로 나뉘는 경우를 생각해 보자.

등급	1등급	2등급	3등급	4등급	5등급	6등급	7등급	합계
학생 수	3	5	7	10	6	6	3	40

- **평균(mean)**: (3 + 10 + 21 + 40 + 30 + 36 + 21) / 40 = 4.025
- **분산(variance)**: 2.774

fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

모멘트 방법을 활용한 정규 분포 추정

- 정규 분포의 함수에서 평균(μ)과 표준편차(σ)를 넣어 확률 값을 계산할 수 있다.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

모멘트 방법을 활용한 정규 분포 추정


```python
import math

arr = [1] * 3 + [2] * 5 + [3] * 7 + [4] * 10 + [5] * 6 + [6] * 6 + [7] * 3

mean = 0 # 평균
for x in arr:
    mean += x / len(arr)
variance = 0 # 분산
for x in arr:
    variance += ((x - mean) ** 2) / len(arr)
std = math.sqrt(variance) # 표준 편차

print(f"평균: {mean:.3f}")
print(f"분산: {variance:.3f}")
print(f"표준 편차: {std:.3f}")
```

[실행 결과]

```
평균: 4.025
분산: 2.774
표준 편차: 1.666
```

fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

모멘트 방법을 활용한 정규 분포 추정


```
import numpy as np
import matplotlib.pyplot as plt
plt.rcParams["figure.figsize"] = [10, 8]

x = np.linspace(mean - 10, mean + 10, 1000) # 평균(mean)을 중심으로 다수의 x 데이터 생성
# 정규 분포의 확률 밀도 함수(probability density function)
y = (1 / (np.sqrt(2 * np.pi) * std)) * np.exp(-1 / (2 * (std ** 2)) * ((x - mean) ** 2))
plt.plot(x, y)
plt.xlabel("$x$")
plt.ylabel("$f_X(x)$")
plt.show()
```

fast campus Copyright FASTCAMPUS Corp.All Rights Reserved

모멘트 방법을 활용한 정규 분포 추정