선수 지식 - 통계
데이터 추출

데이터 추출 | 딥러닝의 기초가 되는 확률 개념 알아보기

강사 나동빈

선수 지식 - 통계

데이터 추출

데이터 추출(Data Sampling)

- 기계 학습 분야에서는 데이터를 랜덤으로 추출하는 경우가 많다.
- 예를 들어 100개의 학습 데이터 중에서 랜덤으로 1개를 추출해야 한다면?

리스트 내에서 1개의 원소만 랜덤으로 추출하기

- *choice*() 메서드를 사용해 리스트 내에서 **1개의 원소를 랜덤으로 추출**할 수 있다.

```
import random

arr = [1, 2, 3, 4, 5]
sampled = random.choice(arr)
print(sampled)
```

[실행 결과]

1부터 5 사이의 정수 중 하나

리스트에서 여러 가지 원소를 랜덤 추출하기 (중복 허용 X)

- $sample()$ 메서드를 사용해 k개의 데이터를 **중복 없이** 추출할 수 있다.
- 단, 데이터의 개수를 초과할 수 없다. (아래 예시에서는 5를 초과할 수 없다.)

```
import random

arr = [1, 2, 3, 4, 5]
sampled = random.sample(arr, 3)
print(sampled)
```

[실행 결과]

1부터 5 사이의 정수 3개를 포함한 리스트

리스트에서 여러 가지 원소를 랜덤 추출하기 (중복 허용)

- 리스트 내에서 k개의 데이터를 **중복을 허용하여** 추출할 수 있다.
- **첫 번째 방법**은 $choice()$ 메서드를 이용하는 방식이다.

```
import random

arr = [1, 2, 3, 4, 5]
sampled = [random.choice(arr) for i in range(3)]
print(sampled)
```

[실행 결과]

1부터 5 사이의 정수 3개를 포함한 리스트 (중복 가능)

리스트에서 여러 가지 원소를 랜덤 추출하기 (중복 허용)

- 두 번째 방법은 *choices*() 메서드를 사용하는 방법이다.

```
import random

arr = [1, 2, 3, 4, 5]
sampled = random.choices(arr, k=3)
print(sampled)
```

[실행 결과]

1부터 5 사이의 정수 3개를 포함한 리스트 (중복 가능)

리스트에서 여러 가지 원소를 랜덤 추출하기 (중복 허용)

- $choices()$ 메서드는 중복을 허용하기 때문에, k가 원소의 개수보다 클 수 있다.

```
import random

arr = [1, 2, 3, 4, 5]
sampled = random.choices(arr, k=7)
print(sampled)
```

[실행 결과]

1부터 5 사이의 정수 7개를 포함한 리스트 (중복 가능)

균등 분포(Uniform Distribution)에서 추출

- [0,1] 범위의 **균등 분포**에서 5개의 데이터를 추출한다.

```python
import numpy as np

sampled = np.random.uniform(0, 1, 5)
print(sampled)
```

[실행 결과 예시]

```
[
  0.81220421,
  0.83316333,
  0.16066372,
  0.22085026,
  0.72513752
]
```

정규 분포(Normal Distribution)에서 추출

- **표준 정규 분포**(평균: 0, 표준편차: 1)에서 5개의 데이터를 추출한다.

```python
import numpy as np

sampled = np.random.normal(0, 1, 5)
print(sampled)
```

[실행 결과 예시]

```
[
  -1.71076484,
  0.13390148,
  0.29766549,
  1.34645879,
  -0.06277921
]
```