선수 지식 - 통계
결합 확률과 주변 확률

결합 확률과 주변 확률 | 딥러닝의 기초가 되는 확률 개념 알아보기

강사 나동빈

선수 지식 - 통계

결합 확률과 주변 확률

독립(Independent)

- $P(X \cap Y) = P(X)P(Y)$인 경우, 두 사건 X와 Y는 서로 독립이다. (필요충분조건)
- 두 변수가 서로 영향을 주지 않는다는 의미다.
- 예를 들어 로또 1등 당첨 확률은 항상 1 / 8,145,060이다.
- **예시)** ① 1번째 자동 **로또**가 1등에 당첨되는 사건과 ② 2번째 자동 **로또**가 1등에 당첨되는 사건

[독립 사건에 대해서 생각해 보기]

- 아래 두 사건은 독립 사건일까?

① 내일 내가 수학 쪽지 시험에서 100점 맞는 사건

② 내일 샌프란시스코에 비가 오는 사건

종속(Dependent)

- 두 사건 X와 Y가 있다고 가정하자.
- 한 사건의 결과가 다른 사건에 영향을 줄 때 X와 Y를 **종속 사건**이라고 한다.
- **(상관관계)** 일반적으로 수학 성적이 높으면 영어 성적도 높다.
- ① **수학을 100점 맞는 사건**과 ② **영어를 100점 맞는 사건**을 생각해 보자.

배반 사건

- **배반 사건**은 “교집합이 없는“ 사건을 의미한다.
- **예시)** ① 나의 수학 성적이 50점 이상인 사건과 ② 나의 수학 성적이 50점 미만인 사건

배반 vs. 독립

- **배반 사건**과 **독립 사건**을 비교한 표는 다음과 같다.
- 독립 사건에서 $P(Y|X) = P(Y)$의 의미는 무엇일까?

→ "사건 X의 발생 여부와 상관없이, 사건 Y가 발생할 확률은 동일하다."

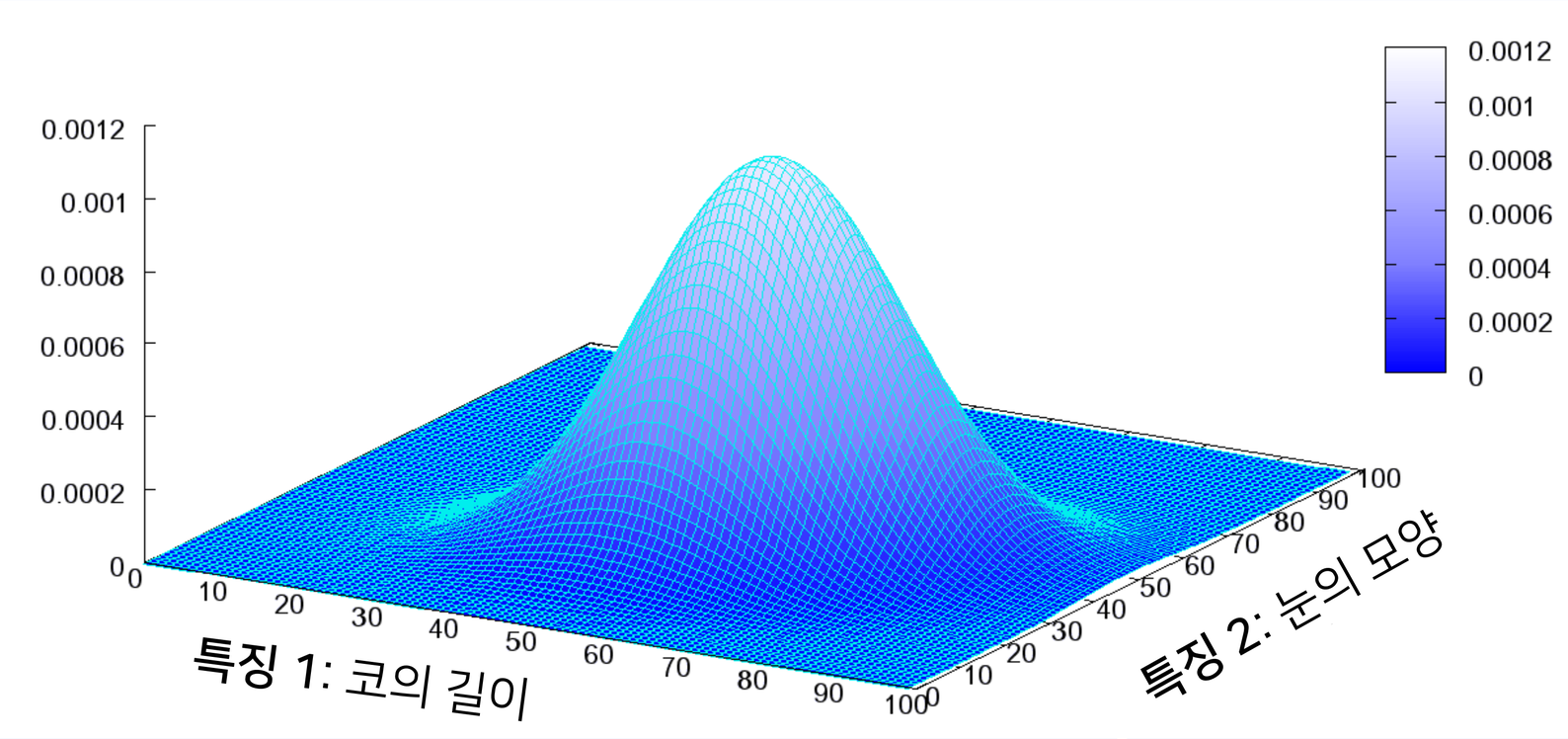
	배반 사건	독립 사건
정의	$X \cap Y = \phi$	$P(Y\|X) = P(Y)$
의미	두 사건이 동시에 일어나지 않는다.	두 사건이 동시에 일어날 때 서로 영향을 주지 않는다.
판단 방법	$X \cap Y = \phi$라면, 두 사건은 서로 배반 사건	$P(X \cap Y) = P(X)P(Y)$라면, 두 사건은 서로 독립 사건

다변수 확률 변수(Multivariable Random Variable)

- **확률 변수**가 두 개 이상 있는 경우를 말한다.
- 이 경우 개별적인 확률 변수에 대한 확률 분포를 고려할 수도 있다.
- 두 확률 변수를 모두 고려한 "복합적인" 확률 분포를 계산할 수 있다.

다변수 확률 분포(Multivariable Probability Distribution)

- 딥러닝 분야 분포는 일반적으로 **다변수 확률 분포(변수가 여러 개)**에 해당한다.
- 얼굴(face) 특징에 대한 확률 분포 예시를 고려해 보자.



* https://en.wikipedia.org/wiki/Multivariate_normal_distribution

결합 확률(Joint Probability)

- **두 개의 사건이 동시에 일어날 확률**로, 두 확률 변수의 교집합이 발생할 확률이다.
- 마찬가지로 확률은 항상 0과 1 사이의 값을 가진다.
- $P(X, Y)$ 혹은 $P(X \cap Y)$ 형태로 표현한다.

결합 확률 함수(Joint Probability Function)

- 이산확률변수 X, Y에 대한 결합 확률 함수 $f_{XY}(x_i, y_j)$란 무엇일까?

[특징]

- $f_{XY}(x_i, y_j) = P(X = x_i, Y = y_j)$이다.
- X가 x_1, x_2, …의 값을 가질 수 있고, Y가 y_1, y_2, …의 값을 가질 수 있다고 가정한다.
- 결과적으로 단순히 $f(x, y)$라고 쓰기도 한다. (X, Y의 결합 확률 분포)
- X와 Y가 가진 범위에서 결합확률함수의 값을 모두 더하면 1이다.

결합확률질량함수 예시

- 랜덤으로 1부터 9까지의 수 중에서 하나를 출력하는 기계가 있다.
- 이 기계를 한 번 동작 시켰다고 가정하자.

1. 얻은 수가 짝수이면 $X = 0$, 홀수이면 $X = 1$이다.
2. 얻은 수가 소수가 아니면 $Y = 0$, 소수이면 $Y = 1$이다.

- 얻을 수 있는 수에 따른 X와 Y의 값을 확인해 보자.

	1	2	3	4	5	6	7	8	9
X	1	0	1	0	1	0	1	0	1
Y	0	1	1	0	1	0	1	0	0

결합확률질량함수 예시

- 결합 확률 함수는 다음과 같다.
- $f(0,0) = 3/9$
- $f(0,1) = 1/9$
- $f(1,0) = 2/9$
- $f(1,1) = 3/9$

Y \ X	0	1
0	3 / 9	2 / 9
1	1 / 9	3 / 9

결합확률질량함수 예시

- 수학 점수는 등급 형태로 1부터 5까지의 값을 가진다.
- 영어 점수는 등급 형태로 1부터 5까지의 값을 가진다.
- 수학 등급 확률 변수 X
- 영어 등급 확률 변수 Y

[궁금한 점] 수학 등급이 높으면, 영어 등급도 높을까?

→ 결합 확률 분포를 확인하여 경향을 확인할 수 있다.

결합확률질량함수 예시

- 테이블의 각 원소를 **수학 등급**(X)이 x, **영어 등급**(Y)이 y인 학생의 수라고 해보자.
- 다음과 같이 표현된다. 예를 들어 수학 등급이 3등급, 영어 등급이 2등급인 학생의 수는 4명이다.

Y \ X	1	2	3	4	5
1	2	1	0	0	0
2	1	3	4	0	0
3	1	3	5	2	0
4	0	0	0	3	2
5	0	0	0	1	2

결합확률질량함수 예시

- 결합 확률 질량 함수는 모든 확률 값을 더했을 때 1이다.
- 따라서 다음과 같이 결합 확률 질량 함수로 표현할 수 있다. (원소의 총 합은 1이다.)

Y \ X	1	2	3	4	5
1	2 / 30	1 / 30	0 / 30	0 / 30	0 / 30
2	1 / 30	3 / 30	4 / 30	0 / 30	0 / 30
3	1 / 30	3 / 30	5 / 30	2 / 30	0 / 30
4	0 / 30	0 / 30	0 / 30	3 / 30	2 / 30
5	0 / 30	0 / 30	0 / 30	1 / 30	2 / 30

결합확률질량함수(Joint Probability Mass Function)

- 이산확률변수가 두 개 이상인 확률질량함수다.
- 확률은 $P_{XY}(x,y) = P(X = x, Y = y)$ 형태로 표현한다.
- 또한 이때 $\sum_i \sum_j P(X = x_i, Y = y_i) = 1$이다. (원소의 총 합은 1이다.)
- 전체 학생 수가 30명일 때, 수학이 1등급이면서 영어가 2등급인 학생이 **1명** 있다면?

$$P_{XY}(1,2) = P(X = 1, Y = 2) = 1/30$$

파이썬 소스 코드 예시) 결합 확률 질량 함수

- 수학 성적(X)과 영어 성적(Y)에 대한 결합 확률 질량 함수를 나타낼 수 있다.

```
import pandas as pd


scores = [1, 2, 3, 4, 5]
data = [
    [2, 1, 0, 0, 0],
    [1, 3, 4, 0, 0],
    [1, 3, 5, 2, 0],
    [0, 0, 0, 3, 2],
    [0, 0, 0, 1, 2]
]

df = pd.DataFrame(data, index=scores, columns=scores)
df.columns.name = "X"
df.index.name = "Y"
pmf = df / df.values.sum()
print(pmf)
```

[실행 결과]

X	1	2	3	4	5
Y					
1	0.066667	0.033333	0.000000	0.000000	0.000000
2	0.033333	0.100000	0.133333	0.000000	0.000000
3	0.033333	0.100000	0.166667	0.066667	0.000000
4	0.000000	0.000000	0.000000	0.100000	0.066667
5	0.000000	0.000000	0.000000	0.033333	0.066667

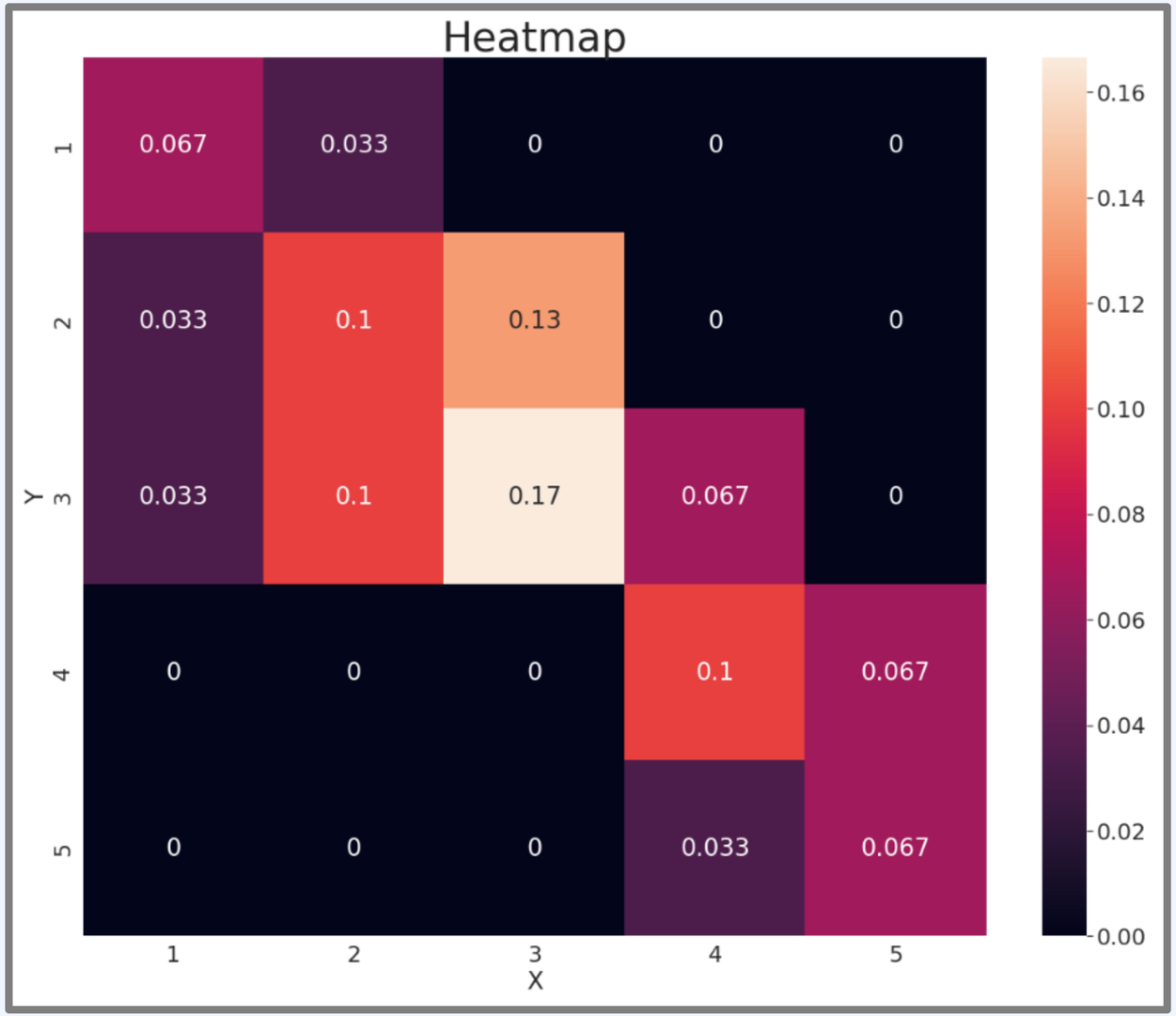
파이썬 소스 코드 예시) 히트맵(Heatmap)

- Python을 이용해 특정한 결합 확률 질량 함수를 히트맵으로 표현할 수 있다.

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


sns.set(font_scale=2)
plt.rcParams["figure.figsize"] = [20, 16]
ax = sns.heatmap(pmf, annot=True,
    xticklabels=[1, 2, 3, 4, 5],
    yticklabels=[1, 2, 3, 4, 5]
)
plt.title("Heatmap", fontsize=40)
plt.show()
```

파이썬 소스 코드 예시) 히트맵(Heatmap)



주변확률질량함수(Marginal Probability Mass Function)

- 두 확률 변수 중에서 하나의 확률 변수에 대해서만 확률 분포를 나타낸 함수다.
- $P_{\mathrm{X}}(\mathrm{x}) = \sum_{y_i} P_{XY}(x, y_i)$
- $P_Y(y) = \sum_{x_i} P_{XY}(x_i, y)$
- 예를 들어 수학 등급이 1등급일 확률은 다음과 같다.
- $P_X(1) = P_{\mathrm{XY}}(1,1) + P_{\mathrm{XY}}(1,2) + P_{\mathrm{XY}}(1,3) + P_{\mathrm{XY}}(1,4) + P_{\mathrm{XY}}(1,5)$
- 즉, $X = 1$로 고정하고, 모든 Y 변수에 대하여 확률 값을 더한 것이다.
- 이를 모든 X에 대하여 표현하면, **주변 확률 질량 함수**가 된다.

주변확률질량함수 예시

- 주변확률질량함수의 예시는 다음과 같다.

Y \ X	1	2	3	4	5	$P_Y(y)$
1	2 / 30	1 / 30	0 / 30	0 / 30	0 / 30	3 / 30
2	1 / 30	3 / 30	4 / 30	0 / 30	0 / 30	8 / 30
3	1 / 30	3 / 30	5 / 30	2 / 30	0 / 30	11 / 30
4	0 / 30	0 / 30	0 / 30	3 / 30	2 / 30	5 / 30
5	0 / 30	0 / 30	0 / 30	1 / 30	2 / 30	3 / 30
$P_X(x)$	4 / 30	7 / 30	9 / 30	6 / 30	4 / 30	

주변확률질량함수 예시

- Python을 이용해 특정한 주변 확률 질량 함수를 나타낼 수 있다.

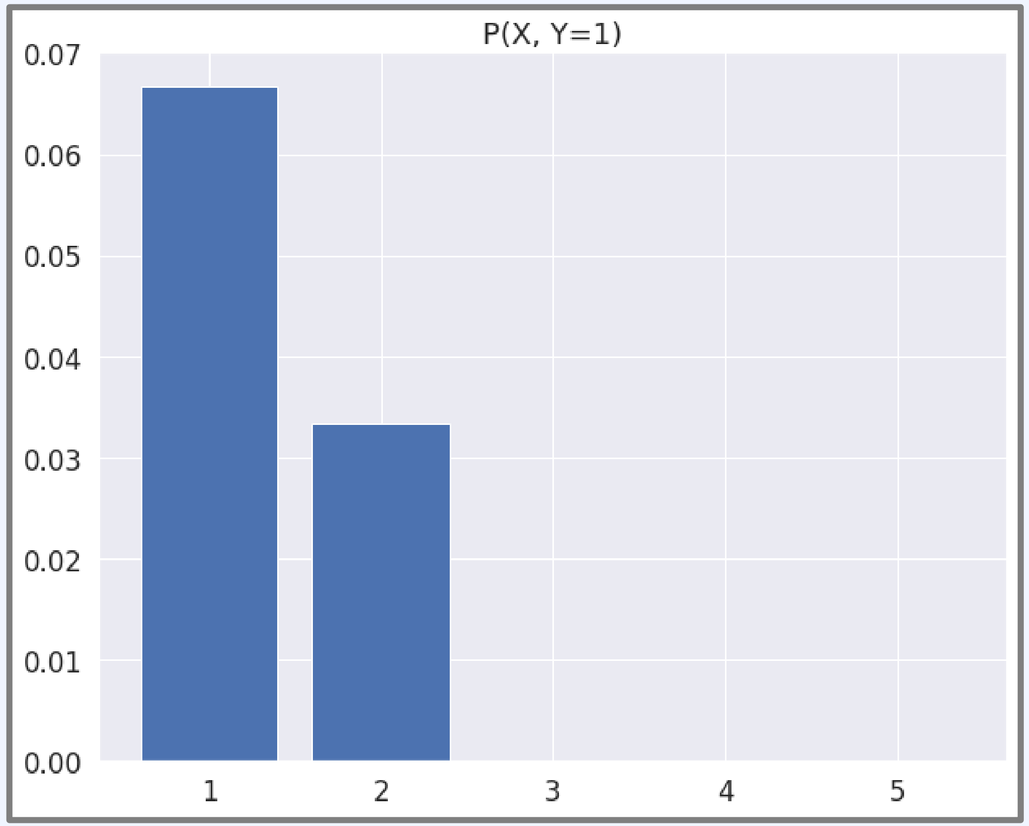
```
index = 0
x = [0, 1, 2, 3, 4]
plt.bar(x, pmf.iloc[index])
plt.xticks(x, ["1", "2", "3", "4", "5"])
plt.title(f"P(X, Y={index + 1})")
plt.show()

marginal_pmf_x = pmf.sum(axis=0)
print(marginal_pmf_x)

marginal_pmf_y = pmf.sum(axis=1)
print(marginal_pmf_y)
```

주변확률질량함수 예시

- Python을 이용해 특정한 주변 확률 질량 함수를 나타낼 수 있다.



주변확률질량함수 예시

- Python을 이용해 특정한 주변 확률 질량 함수를 나타낼 수 있다.

```
marginal_pmf_x = pmf.sum(axis=0)
print(marginal_pmf_x)

marginal_pmf_y = pmf.sum(axis=1)
print(marginal_pmf_y)
```

[실행 결과]

```
X
1   0.133333
2   0.233333
3   0.300000
4   0.200000
5   0.133333
dtype: float64
Y
1   0.100000
2   0.266667
3   0.366667
4   0.166667
5   0.100000
dtype: float64
```