인공지능 수학: 통계학
3. 확률분포

윤상민

E-mail : smyoon@kookmin.ac.kr
Office : 02-910-4645

확률 변수 (random variable)

- 랜덤한 실험 결과에 의존하는 실수
  - 즉 표본 공간의 부분 집합에 대응하는 실수
- 주사위 2개를 던지는 실험
  - 주사위 숫자의 합: 하나의 확률변수
  - 주사위 숫자의 차: 하나의 확률 변수
  - 두 주사위 숫자 중 같거나 큰 수: 하나의 확률 변수
- 동전 10개를 던지는 실험
  - 동전의 앞면의 수
  - 첫번째 앞면이 나올 때가지 던진 횟수
- 보통 표본 공간에서 실수로 대응되는 함수로 정의
- 보통 X나 Y같은 대문자로 표시

확률 변수 (random variable)

- 이산확률변수
  - discrete random variable
  - 확률변수가 취할 수 있는 모든 수 값들을 하나씩 셀 수 있는 경우
  - 주사위, 동전과 관련된 앞의 예
- 연속확률변수
  - continuous random variable
  - 셀 수 없는 경우
  - 어느 학교에서 랜덤하게 선택된 남학생의 키

확률 분포

- Probability Distribution
- 확률변수가 가질 수 있는 값에 대해 확률을 대응시켜주는 관계
- 예)
  - 어떤 확률 변수 X가 가질 수 있는 값: 0,1,3,8
  - 각 값이 나올 확률
  - $P(X = 0) = 0.2$
  - $P(X = 1) = 0.1$
  - $P(X = 3) = 0.5$
  - $P(X = 8) = 0.2$

확률 분포

- 확률분포의 표현: 매우 다양함
  - 표
  - 그래프
  - 함수
  - …

x	0	1	3	8
$P[X = x]$	0.2	0.1	0.5	0.2

확률 분포

- 주사위 2개를 던지는 실험
  - 확률 변수 X: 주사위 숫자의 합
    - X가 가질 수 있는 값
      - 2, 3, …, 12
    - $P(X = 12) = \frac{1}{36}$
  - 확률 변수 Y: 주사위 숫자의 차
    - Y가 가질 수 있는 값
      - 0,1,2,…,5
    - $P(Y = 5) = \frac{2}{36} = \frac{1}{18}$

확률 분포

- 주사위 2개를 던지는 실험
  - 확률 변수 X:
    - 주사위 숫자의 합
  - 주사위를 던질 때마다 X 의 값이 달라질 수 있음
  - n번 실험하면, n개의 숫자가 나옴
  - 이 n개의 숫자의 평균과 분산을 계산할 수 있음
- 확률 변수 X도 평균과 분산을 가짐
  - 이 평균과 분산을 모집단의 평균과 분산이라고 할 수 있음

이산확률변수

- 이산확률변수의 확률분포
  - 보통 함수로 주어짐
  - 확률변수 X가 x라는 값을 가질 확률
    - $P(X = x) = f(x)$
    - 확률질량함수
  - 예)
    - 확률변수 X가 가질 수 있는 값: 0, 2, 5
    - $P(X = x) = f(x) = \frac{x+1}{10}$
      - $P(X = 0) = 0.1$
      - $P(X = 2) = 0.3$
      - $P(X = 5) = 0.6$

이산확률변수

- 이산확률변수의 평균
  - 기대값 (expected value)라고도 함
  - $E(X) = \sum_x xP(X = x) = \sum_x xf(x)$
  - $E(X) = 0 \times 0.1 + 2 \times 0.3 + 5 \times 0.6 = 3.6$

  - 예를 들어 100,000 번의 실험을 했다면,
    - 0이 대략적으로 10,000 번 나옴
    - 2가 대략적으로 30,000 번 나옴
    - 5가 대략적으로 60,000 번 나옴
    - 따라서 평균은

$$\frac{(0 \times 10{,}000 + 2 \times 30{,}000 + 5 \times 60{,}000)}{100{,}000}$$

$$= 0 \times 0.1 + 2 \times 0.3 + 5 \times 0.6 = 3.6$$

이산확률변수

- 이산확률변수의 분산
  - 실험을 할 때마다 확률변수의 값이 달라질 수 있음.
  - 따라서 그 변동의 정도인 분산을 계산할 수 있음.
  - 예를 들어 100,000 번의 실험을 했다면,
    - 평균: 3.6
    - $(0-3.6)^2$이 대략적으로 10,000 번 나옴
    - $(2-3.6)^2$이 대략적으로 30,000 번 나옴
    - $(5-3.6)^2$이 대략적으로 60,000 번 나옴

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$$

$$\sigma^2 = \frac{\left((0-3.6)^2 \times 10{,}000 + (2-3.6)^2 \times 30{,}000 + (5-3.6)^2 \times 60{,}000\ \right)}{100{,}000} = 3.24$$

이산확률변수

- 이산확률변수의 분산
  - $(X-\mu)^2$의 평균
  - $\sigma^2 = E[(X-\mu)^2] = \sum_x (x-\mu)^2 P(X=x)$

  - $\sigma^2 = (0-3.6)^2 \times 0.1 + (2-3.6)^2 \times 0.3 + (5-3.6)^2 \times 0.6 = 3.24$
  - $\mathrm{Var}(X)$라고도 함

이산확률변수

- 이산확률변수의 표준편차
  - 분산의 양의 제곱근
  - $\sqrt{\sigma^2} = \sigma$
  - $\mathrm{SD}(X)$라고도 함

이산확률변수

- 확률변수 X의 확률분포

x	**0**	**1**	**2**	**3**
$P[X = x]$	0.2	0.3	0.1	0.4

- 확률변수 X의 평균, 분산, 표준편차?

이산확률변수

- 확률변수 X의 평균, 분산, 표준편차?
  - $E(X) = \sum_x xP(X = x) = 0 \times 0.2 + 1 \times 0.3 + 2 \times 0.1 + 3 \times 0.4 = 1.7$
  - $\sigma^2 = \sum_x (x - \mu)^2 P(X = x) = (0 - 1.7)^2 \times 0.2 + (1 - 1.7)^2 \times 0.3 + (2 - 1.7)^2 \times 0.1 + (3 - 1.7)^2 \times 0.4 = 1.41$
  - $\sigma = \sqrt{1.41} = 1.187$

이산확률변수

- 확률변수 X의 분산: 간편식

$$\sigma^2 = \sum_x (x-\mu)^2 P(X=x) = \sum_x (x^2 - 2\mu x + \mu^2) P(X=x)$$

$$= \sum_x x^2 P(X=x) - \sum_x 2\mu x P(X=x) + \sum_x \mu^2 P(X=x)$$

$$= E(X^2) - 2\mu \sum_x x P(X=x) + \mu^2 \sum_x P(X=x)$$

$$= E(X^2) - 2\mu^2 + \mu^2 = E(X^2) - \mu^2 = \boldsymbol{E\left(X^2\right) - \{E(X)\}^2}$$

이산확률변수

- 확률변수 X의 분산

x	0	1	2	3
$P[X=x]$	0.2	0.3	0.1	0.4

- $E(X^2) = 0^2 \times 0.2 + 1^2 \times 0.3 + 2^2 \times 0.1 + 3^2 \times 0.4 = 4.3$
- $E(X) = 1.7$
- $\sigma^2 = E(X^2) - \{E(X)\}^2 = 4.3 - (1.7)^2 = 1.41$

결합확률 분포

- joint probability distribution
- 두 개 이상의 확률 변수가 동시에 취하는 값들에 대해 확률을 대응시켜주는 관계
- 확률변수 X
  - 한 학생이 가지는 휴대폰의 수
- 확률변수 Y
  - 한 학생이 가지는 노트북의 수

X

Y		0	1	2
	0	0.1	0.2	0
	1	0	0.4	0.3

결합확률 분포

- 결합확률분포를 통해 각 확률변수의 확률 분포를 도출 할 수 있음
  - 주변확률분포 (marginal probability distribution)

- X의 확률분포

x	0	1	2
$P[X = x]$	0.1	0.6	0.3

- Y의 확률분포

y	0	1
$P[Y = y]$	0.3	0.7

공분산

- 고등학교 1학년 학생들
  - 확률변수 X : 키
  - 확률변수 Y : 몸무게
  - 확률변수 Z : 수학성적
  - $(X-\mu_X)(Y-\mu_Y)$: 양일 가능성 높음
  - $(X-\mu_X)(Z-\mu_Z)$: 양과 음이 될 가능성이 반반

  - $(X-\mu_X)(Y-\mu_Y)$와 $(X-\mu_X)(Z-\mu_Z)$
    - 각각 확률변수
    - 따라서 평균과 분산을 구할 수 있음.

공분산

- Covariance
- 확률변수 X와 Y의 공분산
  - $(X-\mu_X)(Y-\mu_Y)$의 평균

$$\mathrm{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$$
$$= E(XY) - \mu_X\mu_Y = E[XY] - E[X]E[Y]$$

공분산

- 확률변수 X와 Y의 공분산?

X

Y		0	1	2
	0	0.1	0.2	0
	1	0	0.4	0.3

- $E[XY] = 1 \times 1 \times 0.4 + 2 \times 1 \times 0.3 = 1.0$
- $E[X] = 1 \times 0.6 + 2 \times 0.3 = 1.2$
- $E[Y] = 1 \times 0.7 = 0.7$
- $\text{Cov}(X, Y) = E[XY] - E[X]E[Y] = 1.0 - 1.2 \times 0.7 = 0.16$

상관계수 (correlation coefficient)

- 공분산은 각 확률 변수의 절대적인 크기에 영향을 받음
  - 단위에 의한 영향을 없앨 필요

$$\rho = \mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

상관계수 (correlation coefficient)

- $\mathrm{Var}[X] = E[X^2] - [E(X)]^2 = 1^2 \times 0.6 + 2^2 \times 0.3 - (1.2)^2 = 0.36$
- $\sigma_X = \sqrt{\mathrm{Var}[X]} = \sqrt{0.36} = 0.6$
- $\mathrm{Var}[Y] = E[Y^2] - [E(Y)]^2 = 1^2 \times 0.7 - (0.7)^2 = 0.21$
- $\sigma_Y = \sqrt{\mathrm{Var}[Y]} = \sqrt{0.21} = 0.458$
- $\rho = \mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{0.16}{0.6 \times 0.458} = 0.582$