데이터과학과 AI를 위한 파이썬

09강. 확률변수와 확률분포

세종사이버대학교

김명배 교수

학습내용

- 집합
- 확률
- 조건부 확률
- 확률변수

학습목표

- 집합에 대해 설명할 수 있고 파이썬으로 집합을 표현할 수 있다.
- 확률 및 표본공간과 확률표본에 대해 설명할 수 있다.
- 결합확률과 조건부 확률, 베이즈 정리를 설명 할 수 있고, 베이즈 정리에 대해 파이썬으로 적용할 수 있다.
- 확률변수와 확률분포를 설명할 수 있고, 확률분포의 종류에 대해 설명할 수 있다.

1. 집합

1) 집합과 원소

- 구별가능한 객체의 모임을 집합이라고 하고, 집합에 포함된 객체를 원소라고 한다.
- 원소 x와 그 원소를 포함하는 집합 A의 관계 표현과 그 반대 표현

  $x \in A, \quad x \notin A$

[예시] A={1, 2, 3} 이면 $1 \in A,\ 4 \notin A$

[파이썬] set과 frozenset 자료형으로 집합을 표현할 수 있음

set : 내용을 변경할 수 있음(mutable)

frozenset : 자료를 변경할 수 없고(immutable) 딕셔너리 자료형의 키(key)나 set 자료형의 원소가 될 수 있음

- 무한집합은 파이썬의 set 또는 frozenset 자료형으로 표현할 수 없음

세종사이버대학교

1. 집합

2) 집합의 크기

- 집합이 가지는 원소의 수
- $|A|$ 또는 $card(A)$로 표현

[예시] $A = \{1, 2, 3\}$ 이면 $|\mathrm{A}| = card(A) = 3$

[파이썬] len() 함수 사용

세종사이버대학교

1. 집합

3) 합집합과 교집합

- 합집합(union) : 각집합의 원소를 모두 포함하는 집합

$A \cup B$

[파이썬] A.union(B) 또는 A | B

- 교집합(intersection) : 두 집합에 모두에 속하는 원소로 이루어진 집합

$A \cap B$

[파이썬] A.intersection(B) 또는 A & B

1. 집합

4) 전체집합, 부분집합

- 부분집합(subset) : 어떤 집합의 원소 중 일부만을 포함하는 집합
  (전체집합 : 원래의 집합)
  $A \subset B$
- 진부분집합(proper subset) : 원소의 크기가 더 작은 부분집합
  (자기 자신도 부분집합임)

[파이썬] A.issubset(B) 또는 A <= B

1. 집합

5) 차집합, 여집합

- 어떤 집합 A에 속하면서 다른 집합 B에 속하지 않은 원소로 이루어진 A의 부분집합을 A와 B의 차집합(difference)이라고 함

  $A - B$

- 전체 집합 Ω 중에서 부분집합 A에 속하지 않은 원소로 이루어진 부분집합을 A의 여집합(complement)

  $A^c = \Omega - A$

[파이썬] A.difference(B) 또는 A - B

세종사이버대학교

1. 집합

6) 공집합

- 공집합(null set) : 아무런 원소도 포함하지 않은 집합
  (집합의 크기가 0)
- 공집합의 성질
  - $A \cap \emptyset = \emptyset$
  - $A \cup \emptyset = A$
  - $A \cap A^c = \emptyset$

[파이썬] set([])으로 표현

세종사이버대학교

1. 집합

7) 부분집합의 수

- 원소의 개수가 N개인 집합은 2^N개의 부분집합을 가짐

[예시] $A = \{1, 2\}$

부분집합 : 공집합 : Ø

원소 1개 : $\{1\}$, $\{2\}$, $\{3\}$

원소 2개 : $\{1, 2\}$

[연습문제] 집합 $\Omega = \{\mathrm{HH}, \mathrm{HT}, \mathrm{TH}, \mathrm{TT}\}$에 대해

① 이집합의 부분집합의 개수는?

② 이 집합의 모든 부분집합을 frozenset 자료형 객체로 만들고, 이 부분집합들을 원소로 가지는 set 객체를 만드세요. → 파이썬 실습

1. 집합

8) 합잡합과 교집합의 분배 법칙

- $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
- $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

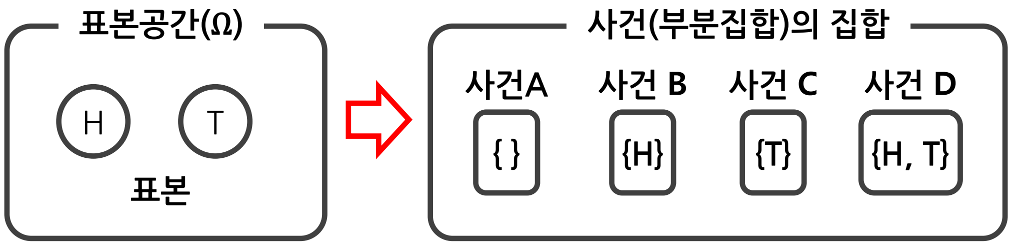
[연습문제]

다음 세 집합 A, B, C에 대해 두 가지 분배법칙이 성립하는지 파이썬 코드로 증명하라.

2. 확률

1) 표본공간과 확률표본 그리고 사건

- 확률표본(표본) : 풀고자하는 확률적 문제에서 발생할 수 있는 현상
- 표본공간 : 가능한 모든 표본의 집합
- 사건 : 표본공간의 부분집합



2. 확률

1) 표본공간과 확률표본 그리고 사건

[예제]

① 동전을 두 번 던지는 문제를 확률론적으로 접근할 때 표본공간 Ω_1를 구하세요.

표본공간 $\boldsymbol{\Omega}_1$={HH, HT, TH, TT}

② 결정된 날짜가 31인지 아닌지를 확률론적으로 접근할 때 표본공간 Ω_2를 구하세요.

표본공간 $\boldsymbol{\Omega}_2$={31일이다(True), 31일이 아니다(False)}

2. 확률

1) 표본공간과 확률표본 그리고 사건

[예제]

③ 회전하는 원판에 화살을 쏘고 화살이 박힌 위치와 기준선과의 각도를 결정하는 문제의 표본공산 Ω_3을 구하세요. 또 표본공간의 표본 개수는?

정답

$\Omega_3 = \{0 \leq \theta \leq 360\}$

표본의 개수는 무한대

2. 확률

2) 확률의 정의

- 확률이란, 어떤 실험에서 사건 A가 일어날 가능성을 수(數)로 나타낸 것

$$P(A) = \frac{A\text{가 일어날 경우의 수}}{\text{모든 경우의 수}} = \frac{n(A)}{n(S)}$$

- 모든 각각의 사건(부분집합)에 어떤 숫자(확률값)를 할당하는 함수
  확률은 표본이 아닌 사건을 입력으로 가지는 함수

[확률의 기본성질]

① 임의의 사건 A에 대해 $0 \leq P(A) \leq 1$인 실수

② 반드시 일어나는 사건 S에 대해 $P(S) = 1$

③ 절대로 일어나지 않는 사건 C에 대해 $P(C) = 0$

2. 확률

2) 확률의 정의

[콜모고로프(kolmogorov)의 공리]

① 모든 사건에 대해 확률은 실수이고 0또는 양수이다.

$P(A) \geq 0$

② 표본공간(전체집합)이라는 사건(부분집합)에 대한 확률은 1이다.

$P(\Omega) = 1$

③ 공통 원소가 없는 두 사건의 합집합의 확률은 사건별 확률의 합니다.

$A \cap B = \emptyset \quad \rightarrow P(A \cup B) = P(A) + P(B)$

2. 확률

3) 확률의 의미

빈도주의 관점의 확률	베이지안 관점의 확률
▪ 모든 것은 실험으로만 알 수 있다. ▪ 여러 번 반복 실험으로 얻은 객관적인 정보다. ▪ 반복적으로 선택된 표본이 사건 A의 원소가 될 경향	▪ 실험하지 않아도 주어진 정보들을 이용해서 활률적으로 표현될 수 있다. ▪ 선택된 표본이 특정한 사건에 속한다는 가설, 명제 혹은 주장의 '신뢰도'

예시)
- 새는 날 수 있다.
- 동전의 앞면이 나올 확률이 0.5다

예시)
- 새가 날 수 있는 가능성은 95%다.
- '앞면이 나왔다'는 주장의 신뢰도는 0.5이다.

2. 확률

4) 확률에서 사용하는 기본 용어

용어	설명	표현
실험	동일한 조건에서 여러 번 반복할 수 있고 그 결과가 우연으로 결정되는 관찰이나 실험	
표본공간	한 실험에서 나올 수 있는 모든 가능한 결과의 집합	Ω
확률표본 (근원사건)	표본 공간을 이루는 개개의 결과	$\omega_1, \omega_2, \ldots$
사건	근원사건의 집합, 표본 공간의 부분집합	
합사건	두 사건 A와 B의 합집합으로 표현할 수 있는 사건	$A \cup B$
곱사건	두 사건 A와 B의 교집합으로 표현할 수 있는 사건	$A \cap B$
여사건	사건 A가 일어나지 않는 사건	A^c
배반사건	사건 A와 B가 동시에 일어나지 않는 사건	$A \cap B = \emptyset$

2. 확률

5) 확률의 성질

① 공집합의 확률

$P(\emptyset) = 0$

② 여집합의 확률

$P(A^c) = 1 - P(A)$

③ 덧셈규칙(포함-배제 원리)

$P(A \cap B) = P(A) + P(B) - P(A \cap B)$

④ 전체 확률의 법칙

복수의 사건 C_i에 대해 $C_i \cap C_j = \emptyset$이고 $C_1 \cap C_2 \cap \cdots = \Omega$일 때

$$P(A) = \sum_i P(A \cap C_i)$$

3. 조건부 확률

1) 결합확률과 조건부확률

- 사건 A와 B가 동시에 발생할 확률을 결합확률(joint probability) 이라고 함 → A와 B의 교집합의 확률이 같음

$$P(A \cap B)\ or\ P(A, B)$$

- 반대로, 개별 사건의 확률 P(A) 또는 P(B)를 주변확률(marginal probability)라고 함
- B가 사실일 경우의 사건 A에 대한 확률을 사건 B에 대한 사건 A의 조건부확률(conditional probability)이라고 함

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A,B)}{P(B)}$$

- A와 B는 각각 '가정과 그 가정에 따른 조건부 결론', '원인과 결과', '근거와 추론'으로 생각할 수 있음

세종사이버대학교

3. 조건부 확률

1) 결합확률과 조건부확률

[연습문제] 범인을 찾는 문제

Case1	머리카락이 길다 : $P(B) = 0.5$	머리카락이 길지 않다 : $P(B^c) = 0.5$	계
범인이 남자다 : $P(A) = \mathbf{0.6}$	3명 $P(A,B)$=3/20	9명 $P(A,B^c)$=9/20	12명
범인이 여자다 : $P(A^c) = 0.4$	7명 $P(A^c,B)$=7/20	1명 $P(A^c,B^c)$=9/20	8명
계	10명	10명	

➔ $\boldsymbol{P(A|B)}$

$$= \frac{\boldsymbol{P(A,B)}}{\boldsymbol{P(B)}} = \frac{\mathbf{3/20}}{\mathbf{10/20}}$$

$$= \frac{\mathbf{3}}{\mathbf{10}} = \mathbf{0.3}$$

Case2	머리카락이 길다 : $P(B) = 0.5$	머리카락이 길지 않다 : $P(B^c) = 0.5$	계
범인이 남자다 : $P(A) = \mathbf{0.6}$	6명 $P(A,B)$=6/20	6명 $P(A,B^c)$=6/20	12명
범인이 여자다 : $P(A^c) = 0.4$	4명 $P(A^c,B)$=4/20	4명 $P(A^c,B^c)$=4/20	8명
계	10명	10명	

➔ $\boldsymbol{P(A|B)}$

$$= \frac{\boldsymbol{P(A,B)}}{\boldsymbol{P(B)}} = \frac{\mathbf{6/20}}{\mathbf{10/20}}$$

$$= \frac{\mathbf{6}}{\mathbf{10}} = \mathbf{0.6}$$

3. 조건부 확률

2) 독립사건과 종속사건

- 두 사건 A와 B에서 한 사건의 결과가 다른 사건에 영향을 주지 않을 때 A와 B는 독립사건이라고 함

$P(A|B) = P(A|B^c) = P(A)$ 또는 $P(B|A) = P(B|A^c) = P(B)$

- 두 사건이 서로 독립일 수 있는 필요충분조건은

$P(A \cap B) = P(A)P(B)$ (단 $P(A) > 0,\ P(B) > 0$)

- 두 사건 A와 B에서 한 사건의 결과가 다른 사건에 영향을 줄 때 A와 B를 종속사건이라고 함

$P(B|A) \neq P(B|A^c)$

세종사이버대학교

3. 조건부 확률

3) 베이즈 정리(Bayesian rule)

- 조건부 확률에 의한 정리

우도, 가능도
(likelihood)

사전확률
(prior probability)

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

사후확률
(posterior probability)

증거(evidence)
정규화 상수

세종사이버대학교

3. 조건부 확률

3) 베이즈 정리(Bayesian rule)

① 베이즈 정리의 확장1

- 사건 A_i에 대해 $A_i \cap A_j = 0$, $A_1 \cup A_2 \cup \cdots = \Omega$ 일 때

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_i P(A_i,B)} = \frac{P(B|A_i)P(A_i)}{\sum_i P(B|A_i)P(A_i)}$$

- 이는 다중분류 문제에서 베이즈 정리를 사용할 수 있음을 보임
- 특히 $A_i = A, A_2 = A^c$인 경우(이분형 분류 문제)

여기에 수식을 입력하십시오.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B,A)+P(B,A^c)} = \frac{P(B|A)P(A)}{P(B|A)P(A)+P(B|A^c)P(A^c)}$$

$$= \frac{P(B|A)P(A)}{P(B|A)P(A)+P(B|A^c)(1-P(A^c))}$$

[파이썬] pgmpy 패키지의 BayesianModel 클래스 적용

3. 조건부 확률

3) 베이즈 정리(Bayesian rule)

② 베이즈 정리의 확장2

- 사건 A의 확률이 사건 B에 의해 갱신된 확률을 계산하는데, 이 상태에서 또 추가적인 사건 C가 발생했다면,

$$P(A|B,C) = \frac{P(C|A,B)P(A|B)}{P(C|B)}$$

[과제] 몬티 홀 문제에 대한 고찰

- 세 문 중에 하나를 선택하여 문 뒤에 있는 선물을 가지는 게임에서
- 참가자가 1번 문을 선택하고 사회자가 선물이 없는 문을 열어 보였을 때
- 참가자는 선택을 바꾸는 것이 유리한가? 그렇지 않은가?

세종사이버대학교

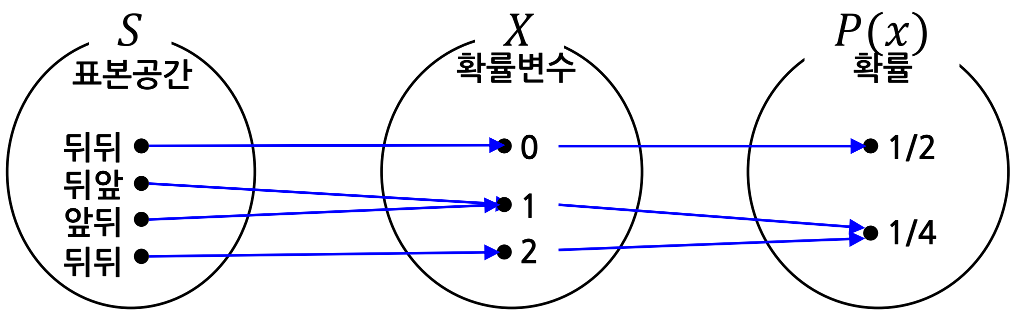
4. 확률변수와 확률분포

1) 확률변수(random variable)

- 실험결과에 따라 표본공간의 각 원소에 실수 값 하나를 대응시킨 것을 확률변수라고 함
- 표본 공간을 정의역, 실수를 공역으로 가지는 함수로 정의할 수 있음

[예시]

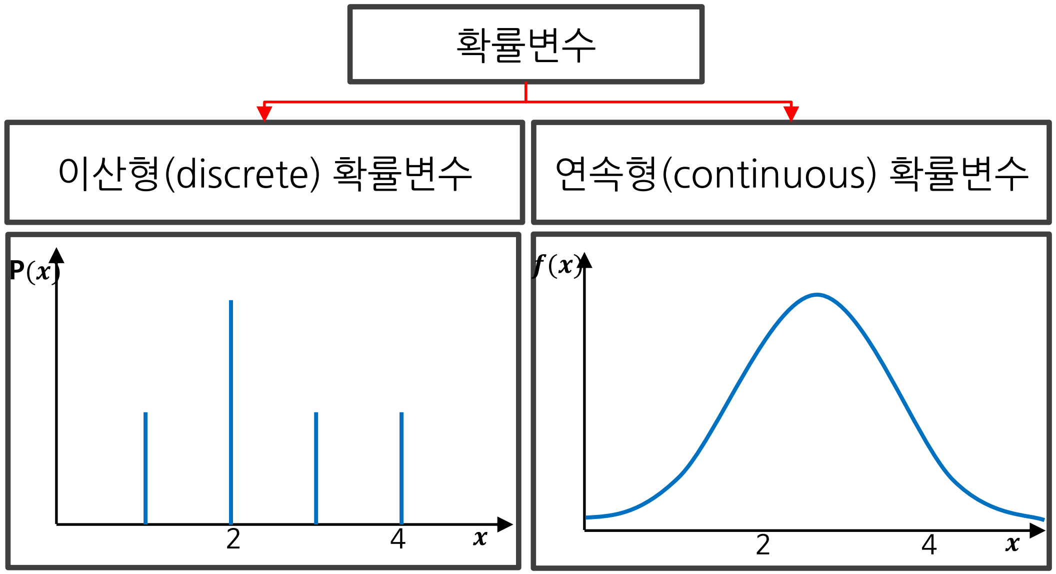
동전 하나를 두 번 던지는 실험에서 앞면이 나온 횟수를 X라고 하면



세종사이버대학교

4. 확률변수와 확률분포

1) 확률변수(random variable)



4. 확률변수와 확률분포

1) 확률변수(random variable)

① 이산형 확률변수

- 확률변수 X가 어느 구간의 모든 실수값을 택하지 않고 0, 1, 2, ⋯ 등의 이산형인 경우

[예시]

주사위를 두 개 던 졌을 때 두 눈의 합 → {2, 3, 4, 6, ⋯ , 11, 12}

② 연속형 확률변수

- 확률변수 X가 어떤 구간의 모든 실수 값을 갖는 연속형인 경우

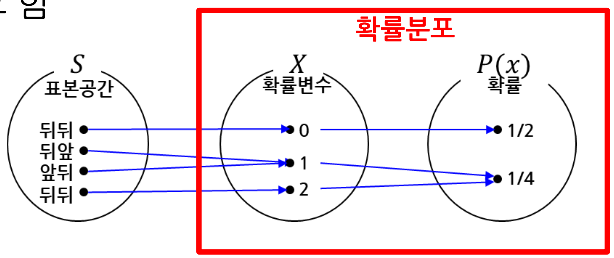
[예시]

홍길동의 키가 178cm와 가깝다 → $P(177<X<179)$

4. 확률변수와 확률분포

2) 확률분포(probability distribution)

- 확률변수의 조합으로 생기는 확률 값의 분포를 나타낸 것
- 확률이 어디에 어느 정도 분포되어 있는지를 수학적으로 명시하고 전달하는 도구
- 표로 표현하면 확률분포표라고 함



[확률분포함수의 종류]

- 확률질량함수
- 누적분포함수
- 확률밀도함수

세종사이버대학교

4. 확률변수와 확률분포

2) 확률분포(probability distribution)

① 확률질량함수(PMF; Probability Mass Function)

- 확률변수가 취할 수 있는 값이 유한 개 이거나 자연수처럼 셀 수 있는 이산형 확률변수일 때, 확률을 나타내는 함수

$$f(x) = P[X = x]$$

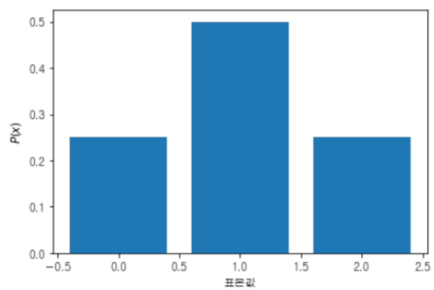
[예시] 동전 한 개를 두 번 던지는 실험에서 앞면이 나오는 횟수를 확률변수 X라고 할 때,

- 확률변수 X = {0,1,2}

$P(X = 0) = 1/4$

$P(X = 1) = 1/2$

$P(X = 2) = 1/4$



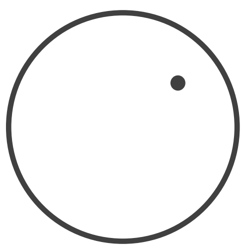
4. 확률변수와 확률분포

2) 확률분포(probability distribution)

② 누적분포함수(CDF; Cumulative Distribution Function)

- 연속형 확률변수는 확률변수가 취할 수 있는 값이 연속적이고 무한하기 때문에 확률변수 하나의 값에 대응하는 확률은 0이 됨
- 따라서 연속형 확률변수는 구간으로 확률을 표현 함

[예시] 회전하는 원반에 화살을 쏘았을 때 특정 한점에 꽂힐 확률



- 표본은 무한하기 때문에 A에 화살이 꽂힐 확률은 0
- 즉, 표본이 무한한 연속형인 경우 하나의 값에 대응되는 확률은 0

4. 확률변수와 확률분포

2) 확률분포(probability distribution)

② 누적분포함수(CDF; Cumulative Distribution Function)

- 시작점을 음의 무한대(-∞)로 고정하여 특정 값까지의 확률을 표현한 함수를 누적분포함수라고 함

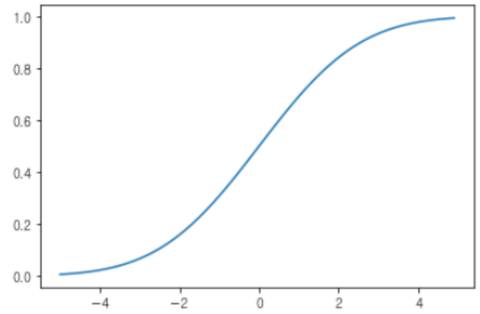
$$F(x) = P[X \leq x]$$

- 특정 구간사건의 확률은 콜모고로프의 공리를 이용하여

$$P(a, b) = F(b) - F(a)$$

[누적분포함수의 특징]

- $F(-\infty) = 0$
- $F(+\infty) = 1$
- $x > y \rightarrow F(x) \geq F(y)$



4. 확률변수와 확률분포

2) 확률분포(probability distribution)

③ 확률밀도함수(PDF; Probability Density Function)

- 누적분포함수는 어떤 확률변숫값이 더 자주 나오는지에 대한 정보를 알기 어려움
- 누적분포함수를 미분하여 구한 도함수의 확률밀도함수라고 함

$$f(x) = \frac{dF(x)}{dx}$$

$$F(x_2) - F(x_1) = \int_{x_1}^{x_2} f(x)\,dx, \qquad F(x) = \int_{-\infty}^{x} f(x)\,dx$$

- 특정한 구간의 확률이 다른 구간에 비해 상대적으로 어떤가를 나타내는 것이며, 값 자체가 확률값이 아님

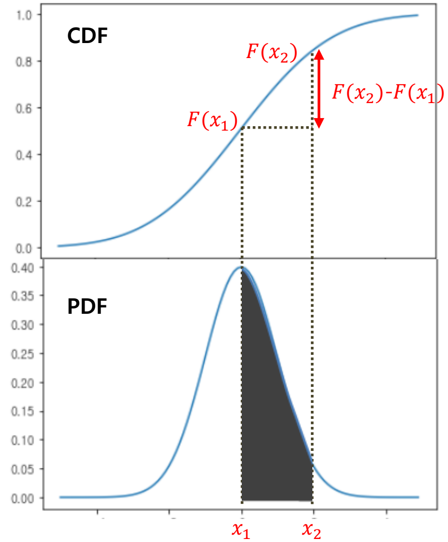
4. 확률변수와 확률분포

2) 확률분포(probability distribution)

③ 확률밀도함수(PDF; Probability Density Function)

[확률밀도 함수의 특징]

- $f(x) \geq 0$
- $\int_{-\infty}^{+\infty} f(x)\,dx = 1$



세종사이버대학교

정리하기

1. 집합

- 구별가능한 객체의 모임을 집합이라고 하고, 집합에 포함된 객체를 원소라고 함
- 집합의 표현 : $x \in A, \quad x \notin A$
- 파이썬 : set(), frozenset()
- 집합의 크기 : $|A|$ 또는 $card(A)$ (파이썬 : len())
- 합집합 : union(), A|B
- 교집합 : intersection(B), A&B
- 부분집합 : issubset(B), A<=B
- 차집합 : difference(B), A-B
- 공집합 : set([])
- 부분집합의 수 : 원소개수가 N개 이면 부분집합은 2^N개
- 합잡합과 교집합의 분배 법칙

 - $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

 - $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

정리하기

2. 확률

- 확률표본(표본) : 풀고자하는 확률적 문제에서 발생할 수 있는 현상
- 표본공간 : 가능한 모든 표본의 집합
- 사건 : 표본공간의 부분집합
- 확률이란, 어떤 실험에서 사건 A가 일어날 가능성을 수(數)로 나타낸 것

$$P(A) = \frac{A\text{가 일어날 경우의 수}}{\text{모든 경우의 수}} = \frac{n(A)}{n(S)}$$

- 콜모고로프(kolmogorov)의 공리 : $P(A) \geq 0,\ P(\Omega) = 1,\ A \cap B = \emptyset \ \rightarrow P(A \cup B) = P(A) + P(B)$

빈도주의 관점의 확률	베이지안 관점의 확률
▪모든 것은 실험으로만 알 수 있다. ▪여러 번 반복 실험으로 얻은 객관적인 정보다. ▪반복적으로 선택된 표본이 사건 A의 원소가 될 경향	▪실험하지 않아도 주어진 정보들을 이용해서 활률적으로 표현될 수 있다. ▪선택된 표본이 특정한 사건에 속한다는 가설, 명제 혹은 주장의 '신뢰도'

- 확률의 성질

$P(\emptyset) = 0,\ P(A^c) = 1 - P(A),\ P(A \cap B) = P(A) + P(B) - P(A \cap B),\ P(A) = \sum_i P(A \cap C_i)$

정리하기

3. 조건부확률

- 사건 A와 B가 동시에 발생할 확률을 결합확률(joint probability)이라고 함
- 개별 사건의 확률 P(A) 또는 P(B)를 주변확률(marginal probability)라고
- B가 사실일 경우의 사건 A에 대한 확률을 사건 B에 대한 사건 A의 조건부확률(conditional probability)이라고 함

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A,B)}{P(B)}$$

- 베이즈 정리(Bayesian rule)

우도, 가능도
(likelihood)

사전확률
(prior probability)

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

사후확률
(posterior probability)

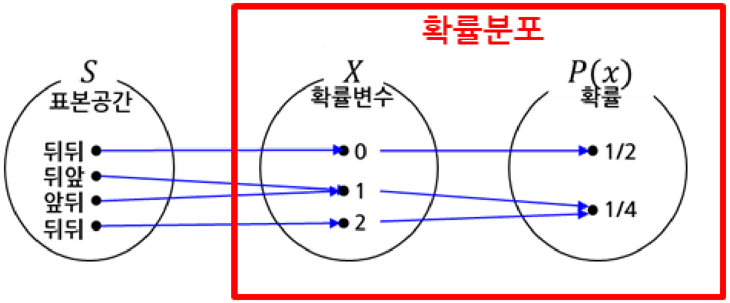
증거(evidence)
정규화 상수

- 베이즈 정리는 다중분류 문제와 사건이 추가되었을 때의 확률문에에 적용가능 함

정리하기

4. 확률변수와 확률분포

- 실험결과에 따라 표본공간의 각 원소에 실수 값 하나를 대응시킨 것을 확률변수라고 함
- 확률변수는 이산형 확률변수와 연속형 확률변수로 구분됨
- 확률분포란 확률변수의 조합으로 생기는 확률 값의 분포를 나타낸 것



- 확률분포함수의 종류 : 확률질량함수, 누적분포함수, 확률밀도함수
- 확률질량함수 : 확률변수가 취할 수 있는 값이 유한 개 이거나 자연수처럼 셀 수 있는 이산형 확률변수일 때, 확률을 나타내는 함수($f(x) = P[X = x]$)
- 누적분포함수 : 시작점을 음의 무한대(-∞)로 고정하여 특정 값까지의 확률을 표현한 함수
  ($F(x) = P[X \leq x]$)
- 확률밀도함수 : 누적분포함수를 미분하여 구한 도함수 ($f(x) = \frac{dF(x)}{dx}$)