베이즈데이터분석 / 이재용 교수

01 강

베이즈 추론의 배경

한국방송통신대학교

목차

- 베이즈 추론의 아버지
  : **베이즈와 라플라스**

- 확률과 확률분포

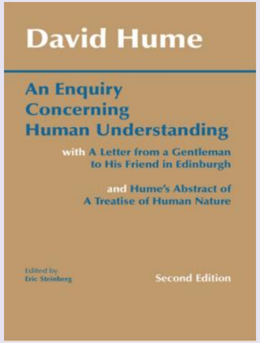
목차

- 베이즈 추론의 아버지
  : **베이즈와 라플라스**

- 확률과 확률분포

토마스 베이즈(Thomas Bayes)

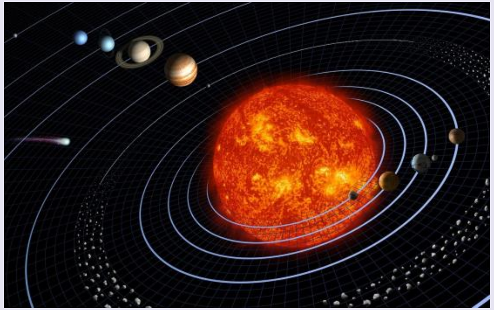
- 토마스 베이즈는 1702년에 태어나 1761년 4월 17일에 사망하였다.
- 베이즈는 영국 장로교 목사이자 아마추어 수학자였다.
- 흄(David Hume)은 인과관계를 인간이 파악할 수 있다는 믿음을 의심했다. 이에 대한 반박으로 베이즈는 논문을 썼다.



한국방송통신대학교

피에르 시몬 라플라스(Pierre-Simon Laplace)

- 라플라스는 1749년 3월 23일 노르망디의 작은 마을에서 태어나고, 1827년 3월 5일에 78세를 일기로 사망하였다.
- 태양계의 안정성 문제를 확률로 다룰 수 있다고 생각하고, 확률에 대한 연구를 시작하였다.



라플라스의 업적

- 베이즈 정리의 재발견
  - 원인의 확률이라는 개념 발견.
- 중심극한정리(Central Limit Theorem)의 발견
- 남녀출생성비의 계산
- “나는 그 가설이 필요하지 않습니다.”

목차

확률

함수

어떤 집합 $\mathcal{X}$상에 정의된 확률 P는 $\mathcal{X}$의 모든 부분 집합에
0과 1 사이의 값을 대응시킨 다음의 조건들을 만족하는 함수이다. 즉,

㉮
$$P : \mathcal{B} \to [0,1]$$

(i) $P(B) \geq 0, \forall B \in \mathcal{B};$

(ii) $P(\mathcal{X}) = 1;$

(iii) (가산 가법성) $\mathcal{B}$의 부분 집합으로 이루진 집합들의 열 (B_n)이 서로소일 때,

$$P(\cup_{n=1}^{\infty} B_n) = \sum_{i=1}^{\infty} P(B_n).$$

를 만족하면 P를 확률이라 한다. 여기서 $\mathcal{X}$를 표본 공간(sample space), $(\mathcal{X}, \mathcal{B}, P)$를 확률 공간(probability space)라고 한다.
또한 $\mathcal{X}$의 부분 집합을 사건(event)라고 한다.

조건부 확률

조건부 확률

$A, B \subset \mathcal{X}$이고 $P(B) > 0$이라 하자. B가 주어졌을 때, A의 조건부 확률은

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

와 같이 정의된다.

- 조건부 확률은 B라는 사건이 일어난 상황에서 사건 A가 일어날 확률을 의미한다.
- 정의에서

$$P(B) > 0$$

조건이 필요하다. 이 조건이 만족하지 않으면 정의에서 분모가 0이 되어 조건부 확률이 정의되지 않는다.

사건의 독립성

두 개의 사건 A와 B가

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

를 만족하면, A와 B는 서로 독립이라 한다.

성별 분포의 예

- 자식이 둘인 가족을 고려하자.
  이 때 자식들의 가능한 성별 분포가 bb, bg, gb, gg이라고 하고,
  각 사건의 확률이 1/4라고 하자.
  이 때 다음에 답하시오.
  b는 아들을 g는 딸을 의미한다.

  (a) 아들이 있다고 할 때, 두 자녀 모두 아들일 확률은?

  (b) 첫번째 자녀가 아들일 때, 두번째 자녀도 아들일 확률은?

확률변수

- 확률변수는 표본공간에서 실수로 가는 함수로 표본공간의 원소를 실수로 대응시킨다.

$$x : \mathcal{X} \rightarrow \mathbb{R}$$

와 같이 표현한다.

- 확률 변수 x가 실수 상의 집합 B에 포함될 확률로 실수 상에 정의된 확률 P_x을

$$P(x \in B) = P_x(B)$$

와 같이 정의할 수 있다.
이 때 P_x를 확률 변수 x의 분포(distribution)이라고 한다.

- 확률 P는 표본 공간$\mathcal{X}$에 정의된 확률이고,
x의 분포 P_x 는 $\mathbb{R}$상에 정의된 확률이다.

확률분포 표현의 세 가지 방법

- (확률) x의 확률 분포 P는 x가 포함된 집합에 대한 확률을 표시한다.

  예를 들면 $A \subset R$일 때, $x \in A$인 확률을 $P(A)$ 혹은 $P(x \in A)$라 표시한다.

- (누적분포함수)

$$F_\theta(t) := P_\theta(-\infty, t], t \in \mathbb{R}$$

를 x의 누적분포함수(cumulative distribution function, cdf)
혹은 분포함수(distributionfunction)라 한다.

확률분포 표현의 세 가지 방법

- (확률밀도함수, 확률질량함수) 모든 $A \subset \mathbb{R}$에 대해서 확률을 0 보다 큰 값을 갖는 함수 $f(x)$의 적분 혹은 합으로

$$P(A) = \begin{cases} \int_A f(x)\,dx \\ \sum_{x \in A} f(x) \end{cases} \quad \forall A \subset \mathbb{R}$$

와 같이 표현할 수 있을 때 $f(x)$를 x(혹은 P)의 확률밀도함수(probability density function) 혹은 밀도함수(density function)라 한다. 확률을 확률 밀도 함수의 적분으로 표현할 수 있는 확률 변수를 연속형이라 하고, 합으로 표현할 수 있는 확률 변수를 이산형이라 한다. 이산형 확률 변수의 밀도 함수를 특별히 확률 질량 함수(probability mass function)이라 하기도 한다.

밀도 함수의 변수 변환

- 확률 변수 x의 확률 밀도 함수가 $f_x(x)$이라 하자.
  새로운 확률 변수가 $y = u(x)$와 같이 정의되고,
  u는 일대일 함수이고 연속적으로 미분 가능이라 하자.

- y의 확률 밀도 함수$f_y(y)$를 다음과 같이 구할 수 있다.

$$f_x(x)\,dx = f_x\left(u^{-1}(y)\right)\left|\frac{dx}{dy}\right| dy \tag{1}$$
$$= f_y(y)\,dy\,.$$

- dx표현을 쓰면 y에 대한 적분을

$$\mathbb{P}(y \in A) = \int_{y\in A} f_x(x)dy = \int_{y\in A} f_x\left(u^{-1}(y)\right)\left|\frac{dx}{dy}\right| dy = \int_{y\in A} f_y(y)dy$$

와 같이 쓸 수 있다.

예

확률변수 x의 밀도함수가

$$f(x) = 2xI(0 < x < 1)$$

일 때, $y = x^2$의 밀도함수를 구하라.

기댓값과 분산

- (기대값)확률 변수 x의 확률 밀도 함수가 $f(x)$일 때, $u(x)$의 기대값은

$$\mathbb{E}u(x) = \begin{cases} \int u(x)f(x)dx, x\text{가 연속형일 때,} \\ \sum u(x)f(x), x\text{가 이산형일 때} \end{cases} \quad \forall A \subset \mathbb{R}$$

와 같이 정의된다. 기대값은 확률 측도 혹은 누적 분포 함수를 이용해서

$$\mathbb{E}u(x) = \int u(x)P(dx)\text{혹은} \int u(x)F(dx)$$

와 같이 쓰기도 한다.

- (분산) 확률 변수 x의 분산은

$$\mathbb{V}ar(x) = \mathbb{E}(x - \mathbb{E}x)^2$$

으로 정의되고, x의 표준 편차는

$$sd(x) = \sqrt{\mathbb{V}arx}$$

로 정의된다.

예

확률변수 u의 밀도함수가

$$f(u) = I(0 < u < 1)$$

일 때, u의 평균과 분산을 구하라.

다음시간

02 강

베이즈 추론