Statics

빅데이터분석 기초 통계

류영표 강사
youngpyoryu@dongguk.edu

강사 소개

류영표
Youngpyo Ryu

동국대학교 수학과/응용수학 석사수료
現 Upstage AI X 네이버 부스트 캠프 1기,2기 멘토
現 Innovation on Quantum & CT(IQCT) 이사
前 한국파스퇴르연구소 Image Mining 인턴(Deep learning)
前 ㈜셈웨어(수학컨텐츠, 데이터 분석 개발 및 연구인턴)

강의 경력

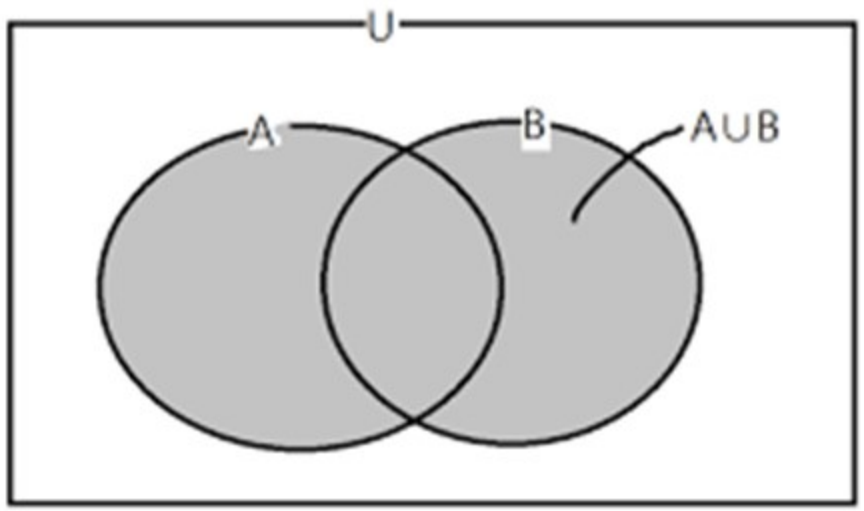
- 현대자동차 연구원 강의 (인공지능/머신러닝/딥러닝/강화학습)
- 딥러닝 집중 교육과정 강사
- (재)웰튼블록체인 6일 과정 (파이썬기초, 크롤링,머신러닝)
- ㈜ 모두의연구소 Aiffel 1기 퍼실리테이터(인공지능 교육)
- 인공지능 자연어처리(NLP) 기업데이터 분석 전문가 양성과정 멘토
- 고려대학교 선도대학 소속 30명 딥러닝 집중 강의
- 메가 IT 아카데미(파이썬, 빅데이터 강사)
- 이젠 종로 아카데미(파이썬, ADSP 강사)
- 최적화된 도구(R/파이썬)을 활용한 애널리스트 양성과정(국비과정) 강사
- 한화, 하나금융사 교육
- 인공지능 신뢰성 확보를 위한 실무 전문가 자문위원

주요 프로젝트 및 기타사항

- 제1회 인공지능(AI)기반 데이터사이언티스트 전문가 양성과정 최우수상 수상(Q&A 챗봇)
- 인공지능(AI)기반 데이터사이언티스트 전문가 양성과정 1기 수료
- 제 1회 산업 수학 스터디 그룹 (질병에 영향을 미치는 유전자 정보 분석)
- 제 4,5회 산업 수학 스터디 그룹 (피부암, 유방암 분류)
- 빅데이터 여름학교 참석 (혼잡도를 최소화하는 새로운 노선 건설 위치의 최적화 문제)

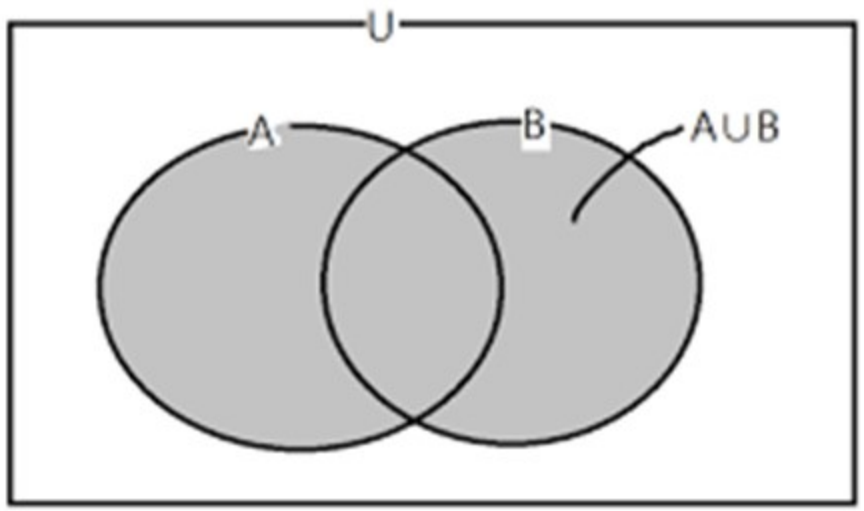
고등학교 통계

- 같은 조건에서 여러 번 반복할 수 있고, 그 결과가 우연에 의하여 결정되는 실험이나 관찰을 시행이라고 부름. 또한 어떤 시행에서 얻어지는 결과를 사건이라고 함.(통계학적으로는 어떤 시행에서 일어날 수 있는 모든 경우의 집합을 S라고 할 때, S의 부분집합을 사건이라고 부름)



고등학교 통계

- 같은 조건에서 여러 번 반복할 수 있고, 그 결과가 우연에 의하여 결정되는 실험이나 관찰을 시행이라고 부름. 또한 어떤 시행에서 얻어지는 결과를 사건이라고 함.(통계학적으로는 어떤 시행에서 일어날 수 있는 모든 경우의 집합을 S라고 할 때, S의 부분집합을 사건이라고 부름)



고등학교 통계

- 서로 다른 n개에서 r개를 택하여, 일렬로 나열하는 것을 n개에서 r개를 택하는 순열(Permutation)이라 함.
- 서로 다른 n개에서 순서를 생각하지 않고, r개를 택하는 것을 n개에서 r개를 택하는 조합(Combination)이라고 함.

$$ {}_{n}\mathrm{P}_{r} = {}_{n}\mathrm{C}_{r} \times r! $$

$$ {}_{n}\mathrm{C}_{r} = {}_{n}\mathrm{P}_{r} \,/\, r! = \frac{n!}{(n-r)! \times r!} $$

고등학교 통계

- 평균(mean,average) : 주어진 수의 합을 측정개수로 나눈 값.
- 분산(Variance) : 변량들이 퍼져 있는 정도, 분산이 크면 불안정
- 표준편차 : 분산의 제곱근

$$(\text{평균}, E(X), m) = \frac{x_1 f_1 + x_2 f_2 + \ldots + x_n f_n}{N} = \frac{\sum_{i=1}^{n} x_i f_i}{N} = \sum_{i=1}^{n} x_i \frac{f_i}{N} = \sum_{i=1}^{n} x_i p_i$$

$$(\text{분산}, V(X)) = \frac{1}{N}\sum_{i=1}^{n}(x_i - m)^2 f_i = (\text{분산}, V) = \sum_{i=1}^{n}(x_i - m)^2 \frac{f_i}{N} = \sum_{i=1}^{n}(x_i - m)^2 p_i$$

$$(\text{표준편차}, \sigma(X)) = \sqrt{V(X)}$$

고등학교 통계

- 어떤 시행에서 사건 A가 일어날 가능성을 수로 나타낸 것을 A가 일어날 확률이라고 부르고 P(A)로 나타낸다.

◇ 수학적 확률의 정의

어떤 실험이나 관찰에서 각각의 경우가 일어날 가능성이 같다**(동일한 가능성)**고 할 때, 일어날 수 있는 모든 경우의 수를 n, 어떤 사건 A가 일어날 경우의 수를 a라고 하면 사건 A가 일어날 수학적 확률 p는

$$p = \frac{(\text{사건 } A\text{가 일어날 경우의 수})}{(\text{모든 경우의 수})} = \frac{a}{n}$$

고등학교 통계

- 어떤 시행에서 사건 A가 일어날 가능성을 수로 나타낸 것을 A가 일어날 확률이라고 부르고 P(A)로 나타낸다.

> 통계적 확률이란 동일한 시행을 n번 반복하여 사건 A가 일어날 횟수가 r_n이라 할 때, n을 크게 함에 따라 $\frac{r_n}{n}$이 일정한 값 P에 가까워지면 $\lim_{n\to\infty} \frac{r_n}{n} = P(A)$ 이고 P을 사건 A의 통계적 확률 이라고 한다.

§ 확률의 기본성질

❶ 임의의 사건 A에 대하여 $0 \le P(A) \le 1$

❷ 사건 A가 반드시 일어날 때, $P(A)=1$

❸ 사건 A가 절대로 일어나지 않을 때, $P(A)=0$

- 사건 A가 주어졌을 때 사건 B의 조건부 확률

$$P(B|A) = \frac{P(A \cap B)}{P(A)}, \quad P(A) > 0$$

- $P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$
- $P(B) = P(A \cap B) + P(A^c \cap B)$
  $= P(A)P(B|A) + P(A^c)P(B|A)$

종속사건과 독립사건

1. 종속 사건 $P(B \mid A) \neq P(B)$

사건 A, B에 대하여 사건 A가 일어날 경우와 사건 A가 일어나지 않을 경우에 따라 사건 B가 일어날 확률이 다를 때 사건 B는 사건 A에 종속한다.

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

2. 독립 사건 $P(B \mid A) = P(B \mid A^c) = P(B)$

사건 A, B에 대하여 사건 A가 일어나든 일어나지 않든 사건 B가 일어날 확률이 달라지지 않을 때 사건 A와 B는 독립이라 한다.

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

고등학교 통계

※ 확률의 곱셈정리의 중요성질

① 확률의 곱셈정리 :

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

이때 A, B가 독립 일 때

곱셈정리는 $P(A \cap B) = P(A) \cdot P(B)$

② A, B가 독립

- A, B^c도 독립 $\cdots P(A \cap B^c) = P(A) \times P(B^c)$
- A^c, B도 독립 $\cdots P(A^c \cap B) = P(A^c) \times P(B)$
- A^c, B^c도 독립 $\cdots P(A^c \cap B^c) = P(A^c) \times P(B^c)$

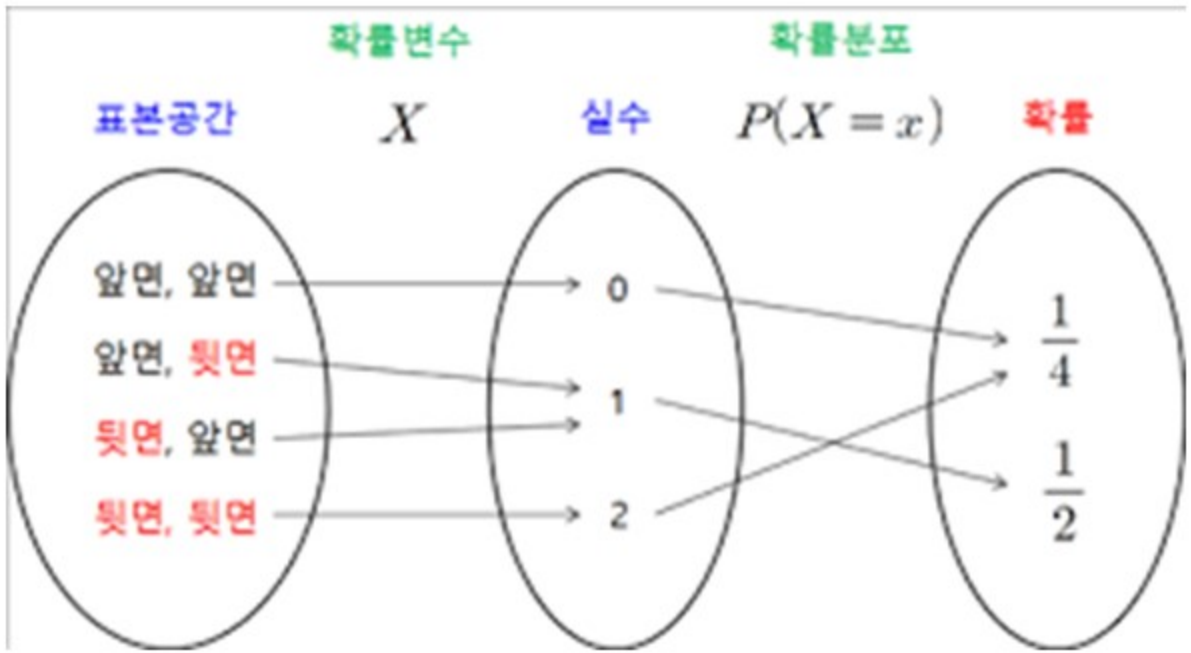
③ 배반사건과 독립사건의 관계

	배반사건	독립사건
정의	$A \cap B = \phi$	$P(B \mid A) = P(B \mid A^c) = P(B$
판단	$P(A \cup B) = P(A) + P(B$	$P(A \cap B) = P(A) \times P(B)$
결론	A, B가 배반사건이면 A, B는 서로 종속사건이다.	

④ A, B가 배반사건이면 $P(B \mid A) = 0$이다.

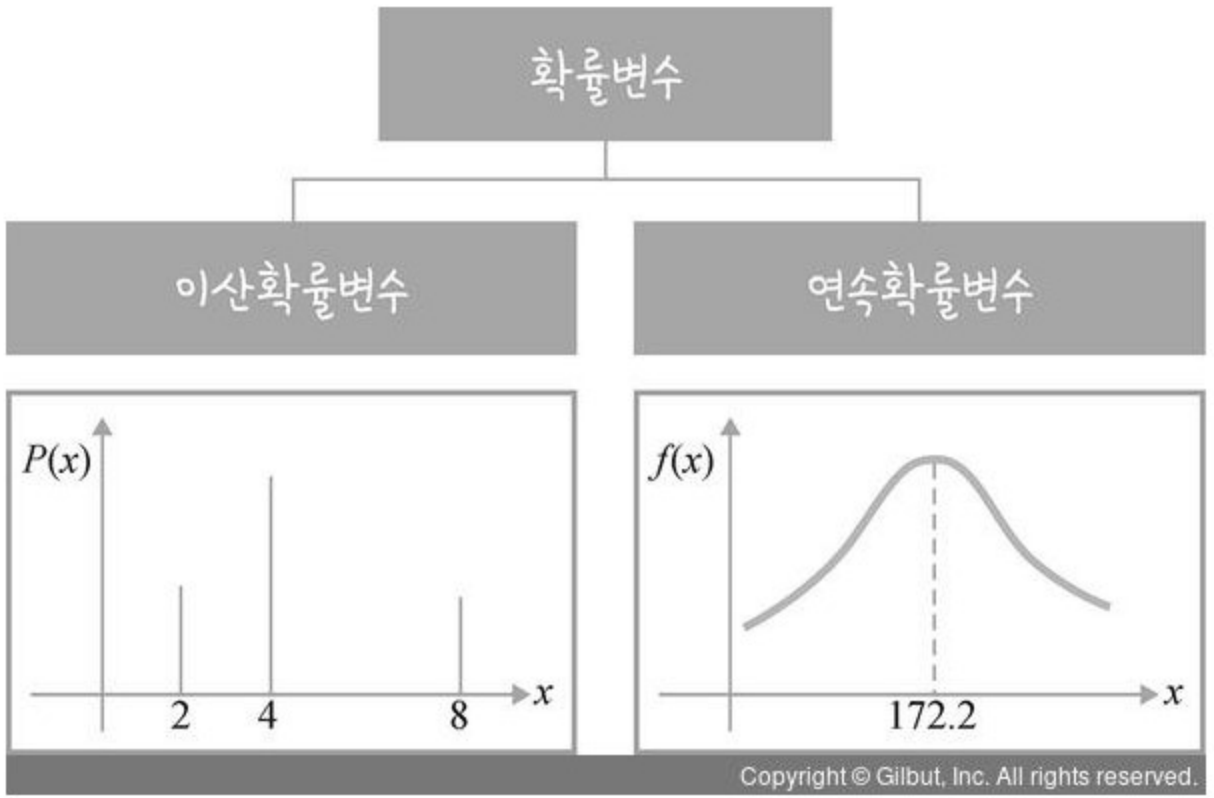
⑤ A와 B가 배반사건이면 $P(A) + P(B) \leq 1$이다

- 어떤 시행의 결과에 따라 표본공간의 각 원소에 하나의 실숫값을 대응시키고, 그 값에 확률이 각각 주어지는 변수를 확률변수라고 함.



고등학교 통계

- 이산(Discrete) : 확률변수에 속한 변량들이 서로 떨어져 분리 된 것
- 연속(Continuous) : 끊어지지 않고 연결된 선과 같이 변량이 무수히 많다.

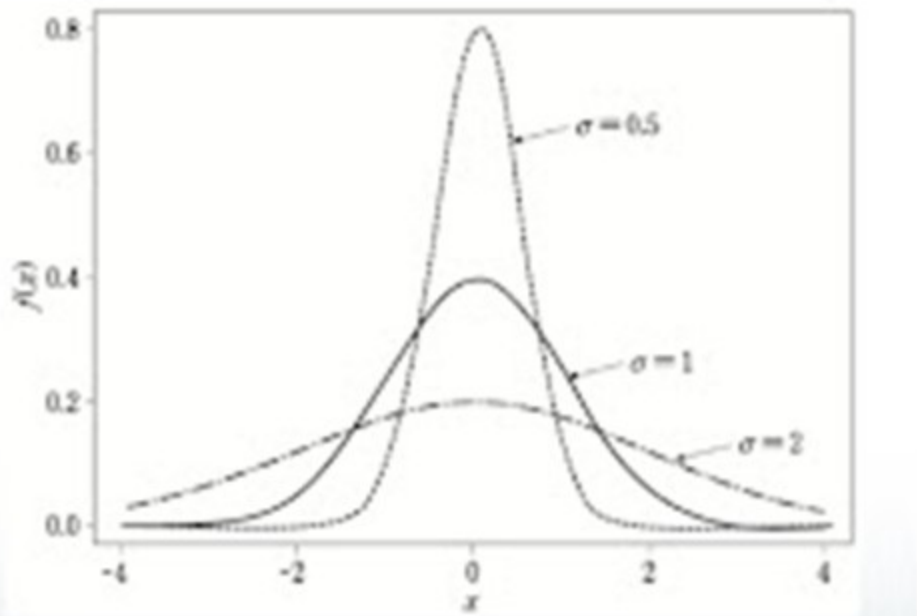


고등학교 통계

$N(\mu,\sigma^2)$: 평균 : μ, 분산 : σ^2

$$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

σ 값이 작을수록 중심에 더 몰려있는 확률분포
μ 값에 따라 x축으로 μ 만큼 평행



$N(0,1)$: 평균 : 0, 분산 : 1

$$f(x)=\frac{1}{\sqrt{2\pi}}e^{\frac{x^2}{2}}$$

$$P(\mu-\sigma\le X\le\mu+\sigma)=0.6827$$
$$P(\mu-2\sigma\le X\le\mu+2\sigma)=0.9545$$
$$P(\mu-3\sigma\le X\le\mu+3\sigma)=0.9973$$

95% : $\pm1.96\sigma$ / 99% : $\pm2.58\sigma$

정규분포의 표준화

$X\sim N(\mu,\sigma^2)$ 일 때 $aX+b\Rightarrow N(a\mu+b,a^2\sigma^2)$

$X\sim N(\mu,\sigma^2)$ 일 때 $Z=\frac{X-\mu}{\sigma}\Rightarrow N(0,1)$

확률이란?

- 어떤 일이 일어날 가능성을 ‘수’로 나타낸 것
- 미래에 발생할 사건에 대한 믿음에 대한 추정
- 물리, 화학, 사회 과학 등에서 발생하는 관심 현상의 측정값을 불확실성에 의해 예측 할 수 없는 경우 사용

  ex) 1분간 당신 맥박 수, 다리가 무너지기 전 최대 하중 등

통계 분류

빈도 주의	베이지안 관점
반복적으로 선택된 표본이 사건(부분 집합) a의 원소가 될 경향(Propensity) 을 그 사건의 확률이라고 함.	선택된 표본이 특정한 사건(부분집합)에 속한다는 가설(hypothesis), 명제(Proposition) 혹은 주장(assertion)의 신뢰도(drgree of belief)라고 함. 단, 반복이라는 개념은 사용되지 않음.
Ex) 동전을 던져 '앞면이 나오는 사건 ' 의 확률값이 0.5라는 것은 빈도주의 관점에서는 실제로 동전을 반복하여 던졌을 경우 동전을 던진 전체 횟수에 확률값을 곱한 숫자만큼 해당 사건이 발생한다고 함.	같은 사건을 비교해보면, 베이지안 관점에서는 '앞면이 나왔다'는 주장의 신뢰도가 0.5이다. - 확률의 정의는 무언가 반복되는 것, 또는 빈도와는 전혀 관계가 없다. '확률값이 0.5이다 ' 라는 주장의 신뢰도 일 뿐

확률 공리

표본공간 S인 실험에서 임의의 사건 A에 대해 아래 조건을 만족하는 P(A)를 A의 확률(probability)이라고 정의하고 이를 확률의 공리(axiom)라 한다.

- 공리 1) $P(A) \geq 0$, 모든 사건의 확률 값은 0보다 크거나 같다.
- 공리 2) P(S) = 1, 표본공간의 확률 값은 1이다.
- 공리 3) 만약 임의의 두 사건 A,B가 상호 배반적이라면 ($A \cap B = \emptyset$), 합집합 확률은 $P(A \cup B) = P(A)+P(B)$로 정의 됨.

확률 공리(Axiom)

1. 표본공간 S인 실험에서 임의의 사건 A에 대해 아래 조건을 만족하는 P(A)를 A의 확률(probability)이라고 정의하고 이를 확률의 공리(axiom)라 한다.

2. 공리 1) $P(A) \geq 0$, 모든 사건의 확률 값은 0보다 크거나 같다.
3. 공리 2) P(S) = 1, 표본공간의 확률 값은 10이다.
4. 공리 3) 만약 임의의 두 사건 A,B가 상호 배반적이라면 ($A \cap B = \emptyset$), 합집합 확률은 $P(A \cup B) = P(A)+P(B)$로 정의 됨.

출처 : https://brunch.co.kr/@kpf10/290

Thank you.

빅데이터 기초 통계 / 류영표 강사
youngpyoryu@dongguk.edu