3. 확 률(Probability)

담당교수 : 김 덕 기

toby123@cbnu.ac.kr

확률의 기본개념

➔ 확률(Probability) : 確(확실할 확), 率(비율 률)

Ex1) Lotto 당첨확률, 한국 야구시리즈에서 특정 팀이 우승할 확률, 폐암에 걸릴 확률

Ex2) 내일 비가 올 확률 60% ➔ 이 정보를 어떻게 유용하게 사용할 것인가 ?

➔ 확률은 불확실성을 이해하는데 중요한 개념

의사결정과 예상 손실

의사결정	예상손실
우산을 가져간다	m
우산을 가져가지 않는다	$0.6\times r+0.4\times 0=0.6r$

➔ 불확실성을 계량화하는 수단.

➔ 우산을 가져가서 겪게 되는 불편~m원의 비용, 안 가져가서 젖게 되는 손실~r비용

$m < 0.6r$?　　$m > 0.6r$?　　$m=0.6r$?

확률 – 객·주관적 방식

확률에 대한 두 가지 해석 방법

- 만일 어떤 석유공학자가 포항 앞 바다에 석유가 매장되어있을 가능성이 60%라고 하였다고 하자.

- 위 주장에 대한 해석을 크게 다음 두 가지로 정리할 수 있다.
  1. 포항 앞바다의 지층과 유사한 과거 여러 지층에서 탐사한 경우를 종합해보면 대략 60%는 석유가 발견되었다.
  2. 석유공학자는 60% 가능성으로 석유가 있다고 개인적으로 믿는다.

- 첫 번째는 확률의 빈도 개념으로 해석(frequency interpretation)
  - 동일한 조건의 실험 반복에 따른 상대돗수의 개념(과학자들)
- 두 번째는 확률의 주관적 해석(subjective interpretation)
  - 확률을 언급하는 개인의 주관적인 생각(철학자 또는 경제적 의사결정자)

- 두 해석방법의 차이에도 실제 확률의 계산은 동일한 규칙을 따름

확률에 대한 두 시각 차이

<u>확률에 대한 두(빈도론Frequentism vs. 베이지안Bayesianism) 시각 차이</u>

- 빈도론과 베이지안이 확률probability을 바라보는 관점 차이

→ 주사위를 던져서 3이 나오는 확률은 1/6이다.

(빈도론) 1000번을 던지면 166번, 10000번 던지면 1666번 3이 등장(경험적 사실, 객관적인 입장)

(베이지안) 주사위를 던질 때 3이 나온다고 16.66% 확신할 수 있음(지식이나 판단 정도)

- 모든 사건들을 경험적으로 확률화 하는 것은 어려움

→ 사전지식을 활용하여 우리가 알고 싶은, 경험적으로 얻기 힘든 사건에 대해 확률을 추정하는 것이 필요

예) EPL (영국 프리미엄리그)에서 토트넘이 이번 시즌 우승할 확률은 얼마일까?

→ (베이지안 추론) 여러 사전 지식과 정보를 활용하여 우승확률을 계산하고 도출함

확률의 기본개념-2

(1) 확률 실험(random experiment) :

➔ 모든 가능한 결과는 알고 있지만 어떤 결과가 나올지 예측할 수 없고 동일한 조건에서 반복 시행할 수 있는 실험.

(2) 표본 공간(sample space) : 이산표본공간, 연속표본공간

➔ 확률실험에서 나올 수 있는 가능한 모든 결과들의 집합.

[EX1] 공정한 동전을 연속해서 두 번 던지는 실험(확률실험)에서 앞면이 나오면 H, 뒷면이 나오면 T라 할 경우 표본 공간 : S={(H,H), (H,T), (T,H), (T,T)}

[EX2] 전구의 수명을 검사하는 실험에서 전구의 수명을 t라 했을 때 표본 공간 :

S는 $S = \{t, t \geq 0\}$

(3) 사건(event) : 단순 사건, 복수 사건

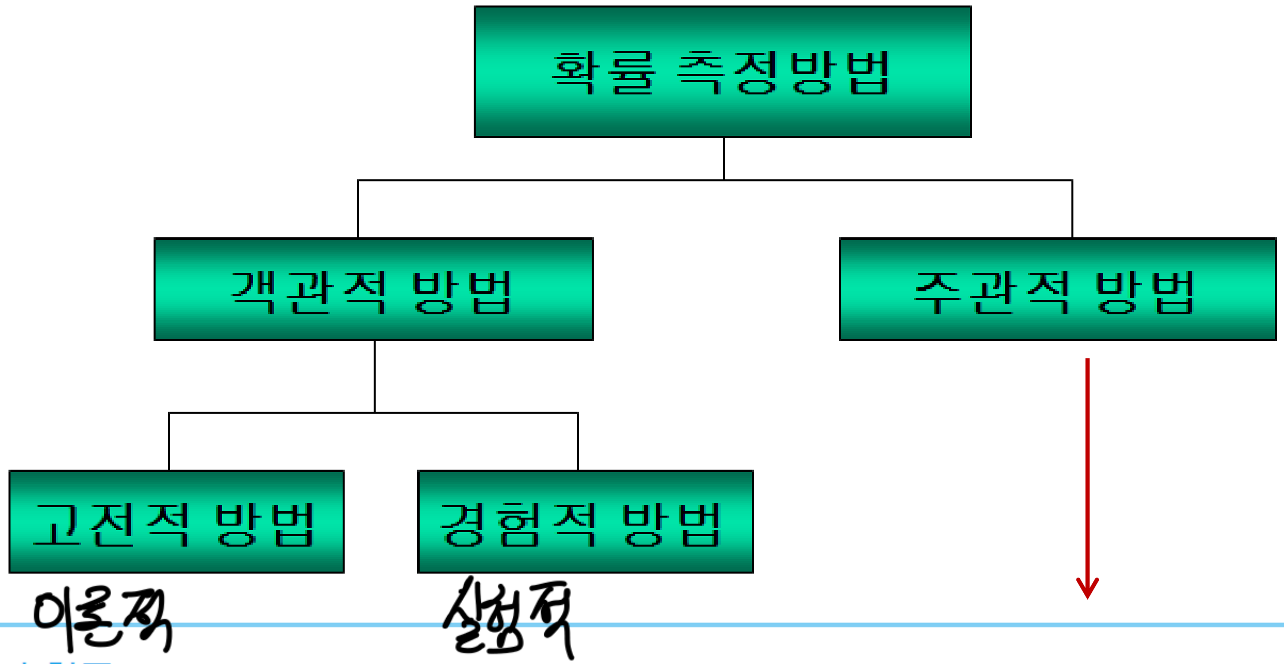
➔ 표본공간에서 특정한 결과에 관한 집합, 표본 공간의 부분집합.

[EX3] 공정한 동전을 두 번 던질 때 동일한 면이 나오는 경우를 사건 A라 하면

A={(H,H), (T,T,)}

확률의 기본개념-3

- 확률의 측정방법



주관적 확률

사건 A가 발생할 확률을 개인적인 경험이나 지식에 근거한 추측 또는 간접적인 증거에 의하여 의사결정자가 주관적으로 정의하는 것을 주관적 확률이라고 한다.

확률의 기본개념-4

고전적 확률의 정의

사건 A의 확률의 고전적 정의는 표본공간 안의 n개의 원소들이 일어날 가능성은 모두 같고, 그 가운데 반드시 한 개는 일어날 때, 사건 A에 속한 모든 원소들의 개수를 표본공간 안의 원소의 개수 n으로 나눈 값이 된다.

$$P(A) = \frac{\text{사건 } A\text{에 속하는 모든 원소의 개수}}{\text{표본공간 안의 모든 원소의 개수}}$$

수학적 확률

상대도수의 극한적 확률의 정의

사건 A의 확률의 상대도수의 극한적 정의는 어떤 통계적 실험을 무한히 반복하였을 때, 사건 A가 일어난 상대도수를 의미하고 다음과 정의한다.

$$P(A) = \lim_{n \to \infty} \frac{\#(A)}{n}$$

통계학적 확률

여기서 n은 시행횟수, $\#(A)$는 n번의 시행 중 사건 A가 일어나는 횟수를 나타낸다.

확률 : 상대도수의 극한값

확률의 의미와 공리

- P(A): 실험을 반복할 때 사건 A가 발생하는 비율의 극한
  - 동전을 던질 때 앞면이 나올 확률은 $\frac{1}{2}$

- P(A): 사건 A가 일어날 가능성에 대한 믿음을 수치화한 것
  - 10년 이내에 통일이 될 확률

- 두 사건 A와 B에 대하여 $A \cap B = \emptyset$일 때, A와 B는 서로 배반(disjoint)이라고 한다.

종속

- 확률의 공리
  1. 모든 사건 A에 대하여 $0 \le P(A) \le 1$
  2. $P(\Omega) = 1$
  3. 사건 $A_1, A_2, \ldots$ 가 서로 배반일 때 (즉, 서로 다른 i, j에 대하여 $A_i \cap A_j = \emptyset$)

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

확률의 법칙–2

- 사건 $A_1, A_2, \dots, A_n$ 가 서로 배반일 때 (즉, 서로 다른 i, j에 대하여 $A_i \cap A_j = \emptyset$)

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$$

- $P(\emptyset) = 0$
- $P(A^c) = 1 - P(A)$: 여사건의 확률법칙
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$: 합 사건의 확률법칙

disjoint ➔ $P(A_1 \cup A_2 \cup A_3 \cup \cdots) = P(A_1)+P(A_2)+P(A_3)+\cdots$

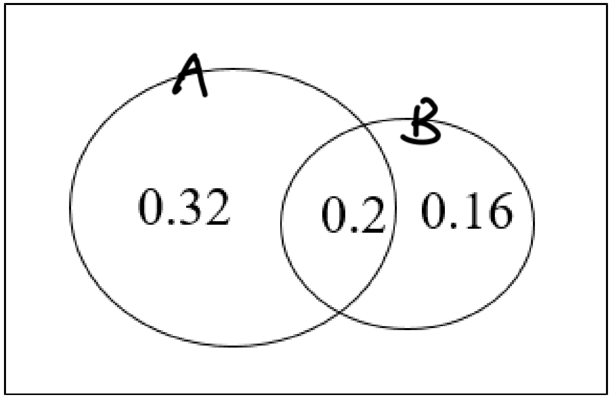
$= \sum_{n=1}^{\infty} P(A_n)$ ☆

확률의 법칙–예제

연습문제 3.7

→ no disjoint

- $P(A)=0.52, P(B)=0.36, P(A\cap B)=0.2$일 때, $P(A\cap B^c)$, $P(A^c\cap B), P(A^c\cap B^c)$의 값은?



- $P(A)=P(A\cap B)+P(A\cap B^c)$이므로 $P(A\cap B^c)=0.52-0.2=0.32$
- $P(B)=P(A\cap B)+P(A^c\cap B)$이므로 $P(A^c\cap B)=0.36-0.2=0.16$
- $P(A^c\cap B^c)=1-P(A\cup B)=1-\left(P(A)+P(B)-P(A\cap B)\right)$ $=1-(0.52+0.36-0.2)=0.32$

확률의 법칙-조건부 확률

- 사건 B가 발생했다는 정보가 주어졌을 때, 사건 A가 발생할 확률을 조건부확률이라고 하며 $\boldsymbol{P(A|B)}$라고 나타낸다.

- 일반적으로 아무런 정보가 없는 상태에서 사건 A가 발생할 확률 $P(A)$와 사건 B가 발생했다는 조건 하에서 사건 A가 발생할 조건부확률인 $P(A|B)$는 같지 않다.

➔ 기존의 정보가 추가정보하에서 갱신
➔ P(A)=P(A|B) ~ 두 사건은 독립

- 사건 B가 발생했다는 조건 하에서 사건 A가 발생할 조건부확률은

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

- 조건부확률 $P(A|B)$는 사건 B 안에서 A가 차지하는 비율이라고 생각할 수 있다. 즉 조건부확률 $P(A|B)$를 구할 때는 표본공간이 B로 한정된다고 생각하고, 이 안에서 사건 A가 발생할 확률을 구하면 된다.

확률의 법칙-조건부 확률(예제)

예1) 어느 회사의 직원 100명 중에서 성별과 흡연 여부에 따른 분할 표

예2) 어느 회사 직원의 체중과 혈압에 따른 비율 표

	흡연	비 흡연	합계
남	30	30	60
녀	15	25	40
합계	45	55	100

	비만	정상	정상 이하	합계
고혈압	0.10	0.08	0.02	0.20
정상 혈압	0.15	0.45	0.20	0.80
합계	0.25	0.53	0.22	1.00

- 임의로 한 명을 선택할 때, 선택된 사람이 흡연자일 확률은? 선택된 사람이 남성이라는 정보가 주어졌을 때, 이 사람이 흡연자일 확률은?
- A: 선택된 사람이 흡연자인 사건
  B : 선택된 사람이 남성인 사건
- $P(A) = \frac{45}{100}$: 전체 직원 중에서 흡연자의 비율
  $P(A|B) = \frac{30}{60}$: 남성 중에서 흡연자의 비율
- $P(A|B) = \frac{30}{60} = \frac{30/100}{60/100} = \frac{P(A \cap B)}{P(B)}$: 남성 중에서 흡연자의 비율

- 한 명을 임의로 선택할 때, 이 사람이 고혈압일 확률은?
- 임의로 선택된 사람이 비만일 때, 이 사람이 고혈압일 확률은?

- A: 선택된 사람이 고혈압인 사건
  B : 선택된 사람이 비만인 사건
- $P(A) = 0.20, P(B) = 0.25, P(A \cap B) = 0.10$
  $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.10}{0.25} = 0.4$: 비만인 사람 중에서 고혈압인 비율

확률의 법칙-곱셈 법칙 & 예제 ☆

확률의 곱셈정리 예측할때

표본공간 S 안의 두 사건 A와 B가 일어난다면 교사건의 확률 $P(A \cap B)$에 대하여 다음이 성립한다.

$$P(A \cap B) = \begin{cases} P(A)P(B \mid A): \ P(A) > 0\text{일 때} \\ P(B)P(A \mid B): \ P(B) > 0\text{일 때} \end{cases}$$

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

- 예제 10) 한 신용카드 회사에서 주요 고객 25명에 대해 그들의 대금결제현황을 조사하였더니, 그 중 20명은 매달 착실하게 결제하고 있었고 나머지 5명은 항상 기일을 넘기는 것으로 드러났다. 이런 25명의 고객 중에 2명을 임의추출한다고 하였을 때 두 사람 모두 대금결제 기일을 어길 확률?

  - F_1: 첫 번째로 선택된 사람이 대금결제 기일을 어길 사건
    F_2: 두 번째로 선택된 사람이 대금결제 기일을 어길 사건
  - $P(F_1 \cap F_2) = P(F_1)P(F_2|F_1) = \frac{5}{25} \times \frac{4}{24} = \frac{1}{30}$

독립 사건(independent)

- 두 사건 A와 B에 대하여 $P(A|B) = P(A)$이 성립한다면, 이는 어떤 의미인가?
  - 사건 A가 발생할 확률과 사건 B가 발생했다는 조건 하에서 사건 A가 발생할 조건부확률이 같다.
  - $P(A|B) = P(A)$이 성립하면 $P(A|B^c) = P(A)$도 성립한다. (why?)
  - 사건 A가 발생할 확률은 사건 B가 발생하든 발생하지 않든 상관없이 같은 값을 가진다.
  - 사건 B가 발생했다는 정보가 사건 A의 확률을 구하는데 영향을 주지 않는다.
  - 사건 A와 사건 B는 서로 영향을 주지 않는다.

$$\boldsymbol{P(A \cap B) = P(A)P(B)}$$

결합 주변

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

if A_i ~ 독립 → $\prod_{i=1}^{n} P(A_i)$

독립 사건(independent) : 예제

- 예제 12) 성별과 길거리 금연에 대한 찬반 의견 조사

	금연 찬성	금연 반대	합계
남	24	16	40
녀	14	6	20
합계	38	22	60

이 중에서 임의로 한 명을 선택할 때,
M : 선택된 사람이 남성인 사건
A : 선택된 사람이 금연에 찬성할 사건

if $P(M \cup A) = P(M)P(A)$,
then A, M ⇒ 독립
〃 $= 0$, 〃 ⇒ 배반
〃 $\neq$ 〃 ⇒ 종속

$P(M) = \frac{40}{60} = \frac{2}{3}, P(A) = \frac{38}{60} = \frac{19}{30}$

$P(M \cap A) = \frac{24}{60} = \frac{2}{5}$

→ 종속

$\rightarrow P(M \cap A) \neq P(M)P(A)$ 이므로 A와 M은 독립이 아니다.

혹은 $P(A) = \frac{38}{60}, P(A|M) = \frac{24}{40}$로 서로 다르므로 A와 M은 독립이 아니다.

$y = f(x) + e$

target feature

↳ 질적 ~ 분류
양적 ~ 예측) + 생성

분할과 전확률 공식

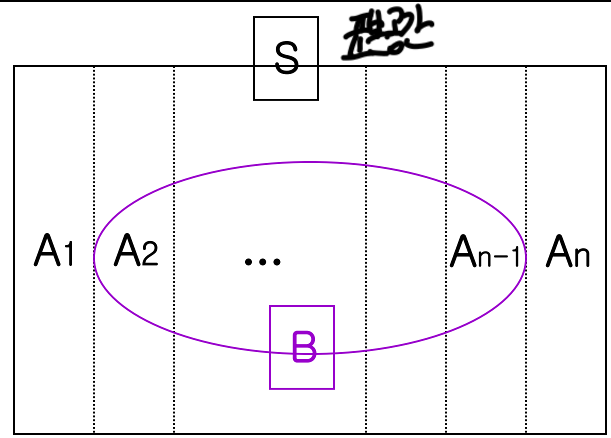
전확률 공식(Total Probability Formula)
집합 $A_1, A_2, \cdots, A_n$이 표본공간 S의 분할이고, 모든 i에 대해 $P(A_i) > 0$이면, 임의의 사건 B에 대해 $$P(B) = P(A_1)P(B \mid A_1) + \cdots + P(A_n)P(B \mid A_n)$$ 이 성립한다.

Proof

$$P(B) = P(S \cap B) = P\left[\left(\cup_{i=1}^{n} A_i\right) \cap B\right]$$

$$= P\left[\cup_{i=1}^{n} (A_i \cap B)\right] = \sum_{i=1}^{n} P(A_i \cap B)$$

$$= \sum_{i=1}^{n} P(A_i) P(B \mid A_i)$$



표본공간

$$p(B) = \sum_{i=1}^{n} P(A_i) P(B \mid A_i) = \sum_{i=1}^{n} P(A_i \cap B)$$) 전확률 공식

$$P(A_k \mid B) = \frac{P(A_k)P(B \mid A_k)}{\sum_{i=1}^{n} P(A_i)P(B \mid A_i)} = \frac{''}{P(B)}$$

사후 확률

베이즈 정리(Bayes Theorem)

조건부확률의 정의와 곱셈법칙, 전확률 공식으로부터, 임의의 사건 A_k 에 대한 조건부 확률 $P(A_k|B)$이 아래와 같이 성립함을 알 수 있다.

베이즈 정리(Bayes Theorem) 실전에서 주로 사용

사건 $A_1, A_2, \cdots, A_n$이 표본공간 S의 분할이고, 모든 i에 대해 $P(A_i) > 0$이고 $P(B) > 0$이면, 조건부 확률 $P(A_k \mid B)$은 다음과 같이 얻어진다.

$$P(A_k \mid B) = \frac{P(A_k)P(B \mid A_k)}{\sum_{i=1}^{n} P(A_i)P(B \mid A_i)}$$

사후 확률

➔ 곱셈정리: $P(A_k \cap B)$

➔ 전 확률 : $P(B)$

※ 베이즈 정리의 원리

사전적 확률 $P(A_k)$

추가 정보 $P(B|A_k)$

사후적 확률 $P(A_k|B)$

관측 전의 사건에 대한 가능성(사전적 확률)과 관측 후의 사건에 대한 가능성(사후적 확률) 사이의 관계를 규명

전확률 공식(예제)

- 예제 14) 어느 회사의 공장 A_1, A_2, A_3에서 주기판(motherboard)의 30%, 50%, 20%를 생산한다. 공장 A_1, A_2, A_3의 불량률이 각각 2%, 1%, 5%이다. 사전정보 추가정보
  1. 이 회사제품 중 임의로 하나를 선택하였을 때, 이 제품이 불량일 확률은?

     B: 제품이 불량인 사건

     사전 × 추가 + …

     $$P(B) = P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_3)P(B|A_3)$$
     $$= 0.3 \times 0.02 + 0.5 \times 0.01 + 0.2 \times 0.05 = 0.021$$

  2. 선택된 제품이 불량일 때, 이 제품이 공장 A_1에서 생산되었을 확률은?

     $$P(A_1|B) = \frac{P(A_1)P(B|A_1)}{P(B)} = \frac{0.3\times0.02}{0.021} = \frac{6}{21}$$

베이즈 정리(예제)

- **사전지식(사전확률)을 활용하여 관심있는 정보에 대한 확률(사후확률)을 추정하는** 것

예1) **(사전확률 정보)** 발병률이 0.01인 A질병이 있다고 하자. 실제로 A병에 걸린 사람이 검사하면 0.99의 비율로 양성판정, 병에 걸리지 않은 사람이 검사하면 0.10의 비율로 양성판정으로 나온다고 한다. 그렇다면 이 질병 검사에서 양성판정이 나왔을때,

(사후확률 추정) 검사 대상자가 진짜 질병에 걸렸을 확률은 ?

$P(A) = 0.01$, $P(A^c) = 0.99$

B ~ 양성, $P(B|A) = 0.99$, $P(B|A^c) = 0.1$

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

$= \frac{\text{곱셈법칙}}{\text{전확률}}$

양성 적중률의 계산 식

$$P(\text{질병}|\text{양성}) = \frac{P(\text{양성}|\text{질병})P(\text{질병})}{P(\text{양성})} = \frac{0.99 \times 0.01}{0.01 \times 0.99 + 0.99 \times 0.10} = 0.091$$

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

$P(B) \longrightarrow P(B|A) \cdot P(A) + P(B|A^c) \cdot P(A^c)$

베이즈 정리(예제1~3) : H.W

예1) (사전확률 정보) 발병률이 0.05인 A질병이 있다고 하자. 실제로 A병에 걸린 사람이 검사하면 0.90의 비율로 양성판정, 병에 걸리지 않은 사람이 검사하면 0.10의 비율로 양성판정으로 나온다고 한다. 그렇다면 이 질병 검사에서 양성판정이 나왔을 때, (사후확률 추정) 검사 대상자가 진짜 질병에 걸렸을 확률은 ?

예2) 실험자 학생이 거짓말을 하는지 알아보기 위해 거짓말 탐지기를 사용하기로 했다. (사전확률 정보) 학생이 거짓말을 할 확률이 0.25이고, 진실을 말할 확률은 0.75이다. 그리고 경험적으로 거짓말 탐지기의 작동이 맞을 확률은 0.85이고 틀릴 확률은 0.15이다. (사후확률 추정) 거짓말 탐지지가 거짓말이라고 응답했을 때, 실제로 이 학생이 진실을 말했을 확률은 얼마인가 ?

예3) 한 대학교 재학생의 35%는 여학생이다. (사전정보 확률) 이 학교의 남학생의 70%, 여학생의 20%가 흡연을 한다고 조사되었다. (사후정보 추정) 이 학교의 학생 한 명을 임으로 선발하였을 때, 선출된 학생이 흡연자라면, 이 학생이 여학생일 확률은 얼마인가 ?