인공지능을 위한 확률과 통계

1.2 확률

서울대 통계학과 장원철

확률의 정의

동전던지기

- 동전을 던졌을때 앞면이 나올 확률은 얼마인가?
- 동전 던지기의 역사
  - Count Buffon (1707-1788): 총 4,040번을 던져서 앞면이 2,048번 (50.69%) 나옴
  - Karl Pearson (1857 - 1936): 총 24,000번을 던져서 앞면이 12,012번 (50.05%) 나옴
  - John Kerrich (1903 - 1985): 2차 세계대전 당시 포로수용소에서 총 10,000번을 던져서 5,067번 앞면이 나옴

확률이란?

- 어떤 사건 (event)이 일어날 가능성을 나타내는 개념
  - 표본공간 (Sample Space, S): 어떤 시행 (experiment)에서 얻을 수 있는 가능한 모든 결과 (outcome)들의 집합
  - 사건: 표본공간의 부분집합으로 보통 집합 A, B, C 등으로 표현

표본공간과 사건: 예시

- 일어날 모든 가능한 사건의 집합
  - 혈액검사를 했을때 나올 혈액형의 표본공간은? $S = \{A, AB, B, O\}$
  - 두명의 자녀가 있다고 하자. 자녀성별의 표본공간은?
    $S = \{(M, M), (M, F), (F, M) . (F, F)\}$
- 두명의 자녀의 성별이 다른 사건을 *A*라고 하자.
  $A = \{(M, F), (F, M)\}$
- 이번학기 본인의 통계학 성적의 표본공간은?
  $S = \{A_+, A_0, A_-, B_+, B_0, B_-, C_+, C_0 . C_-, D_+, D_0, D_-, F\}$
- 본인 성적이 B_0 이상일 사건을 *E* 라고 하자.
  $E = \{A_+, A_0, A_-, B_+, B_0,\}$

사건의 연산

- 집합 연산의 기호를 사용
  - 합사건: $A \cup B$
  - 곱사건: $A \cap B$
  - 여사건: A^c
  - 배반사건: $A \cap B = \emptyset$ 이면 A와 B는 서로 배반

확률의 정의 - 등확률모형의 경우

- 표본공간의 모든 원소들이 일어날 확률이 동일할 경우 사건 A가 일어날 확률 $\Pr(A)$는 다음과 같이 정의할 수 있다.
- $$\Pr(A) = \frac{\text{사건 A에 속하는 원소의 개수}}{\text{표본공간 전체 원소의 개수}}$$
- 예시: 주사위를 던졌을 때 홀수가 나올 확률

확률의 공리를 이용한 확률의 정의

Pr (A) = “사건 *A*가 일어날 확률”이라고 하면 확률은 다음의 공리를 따른다.

1. 표본공간 *S*에서 임의의 사건 A에 대하여 $0 \leq \Pr(A).$

2. $\Pr(S) = 1.$

3. 서로 배반인 사건 $A_1, A_2, \ldots,$ 에 대하여

$$\Pr(A_1 \cup A_2 \ldots) = \Pr(A_1) + \Pr(A_2) + \ldots$$

확률의 성질

- 확률의 공리를 사용하면 다음과 같은 성질을 보일 수 있다.
- $\Pr(\varnothing) = 0$
- $A \subset B$ 이면 $\Pr(A) \leq \Pr(B)$
- $0 \leq \Pr(A) \leq 1$
- $\Pr(A^c) = 1 - \Pr(A)$
- (덧셈법칙) $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$

예제

- 다섯개의 주사위를 던져 숫자의 합이 7이상인 확률은?

조건부 확률 (Conditional Probability)

- 사건 *A*가 주어졌을때 사건 *B*의 조건부확률은 $\Pr(B \mid A)$로 나타내고 $\Pr(A) > 0$이라는 가정하에

$$\Pr(B \mid A) = \frac{\Pr(A \cap B)}{\Pr(A)}$$

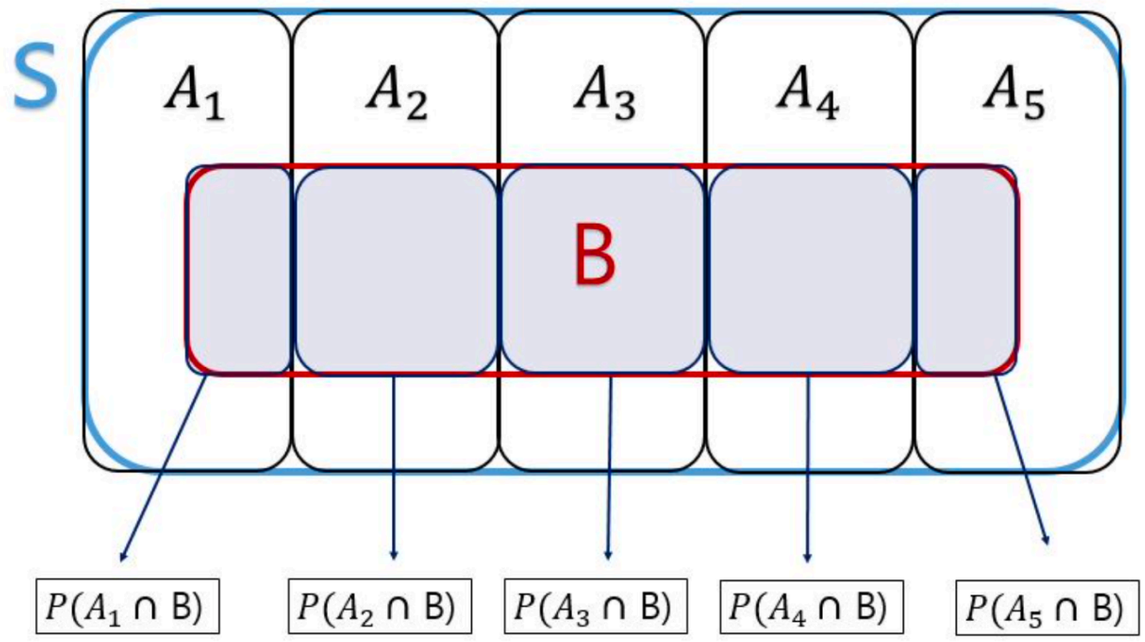
- 사건 *A*를 새롭게 축소된 표본공간으로 간주했을때, 사건 *B*가 일어날 확률
- 예제: 세개의 동전을 차례로 던지는 경우, 앞면이 나온 경우가 두번일때, 첫번째 동전에서 앞면이 나올 확률은?

곱셈법칙

- $Pr(A) > 0, \quad Pr(B) > 0$ 이면
  $Pr(A \cap B) = Pr(A) \cdot Pr(B) = Pr(A \mid B) \cdot Pr(B) = Pr(B \mid A) \cdot Pr(A)$
- 예제: 빨간 구슬 10개와 파란구슬 90개가 들어있는 상자에서 구슬 2개를 임의로 추출할때, 2개 모두 빨간구슬일 확률을 구하여라.

전확률공식

- 어떤 사건 B의 확률 $\Pr(B)$을 구할때, 표본공간의 분할정보를 이용하는 공식



$$\Pr(B) = \Pr(B \mid A_1) \cdot \Pr(A_1) + \Pr(B \mid A_2) \cdot \Pr(A_2) + \Pr(B \mid A_2) \cdot \Pr(A_2) + \Pr(B \mid A_2) \cdot \Pr(A_2) + \Pr(B \mid A_2) \cdot \Pr(A_2)$$

전확률공식

- 표본공간 S의 분할 $\{A_1, \ldots, A_n\}$을 생각하자. 표본공간의 분할은 다음을 만족한다.

$$A_i \cap A_j = \varnothing \quad (i \neq j), \quad A_1 \cup A_2 \cup \cdots \cup A_n = S$$

- 이때, 전확률공식은

$$\Pr(B) = \Pr(B \mid A_1) \cdot \Pr(A_1) + \cdots + \Pr(B \mid A_n) \cdot \Pr(A_n)$$

전확률공식

예제

- 현대중공업 직원의 60%는 이공계, 30% 인문사회계, 10%는 예체능계 학위 소지자라고 하자. 이공계 출신의 50%, 인문사회계 20%, 예체능계 10%가 서울대학교 인공지능 과정 이수생이다. 과정 이수생중 한 명을 임의 추출했을때 그 사원이 인문사회계 학위 소지자일 확률을 구하여라.

독립 (Independence)

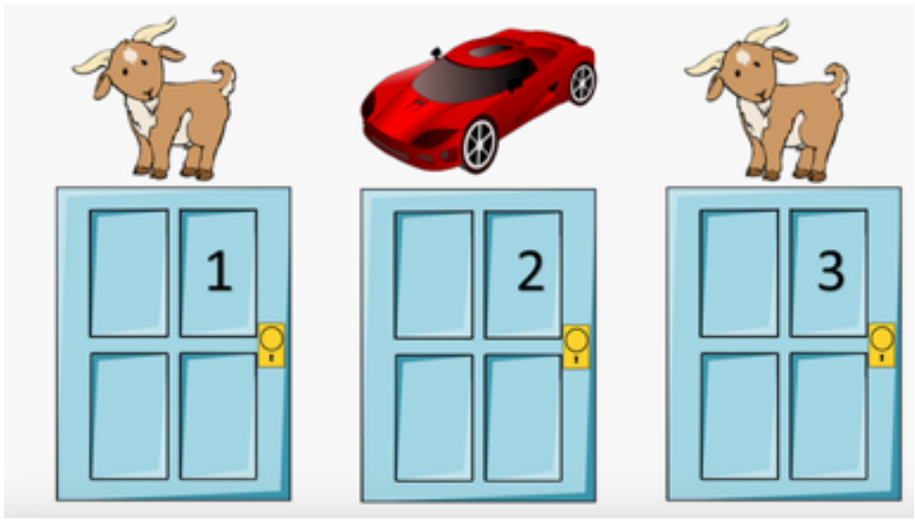
- 한 사건이 일어나는 경우가 다른 사건이 일어나는 경우에 영향을 미치지 않은 경우 두 사건이 독립이라고 한다
- 주사위를 두개 던질때 하나의 주사위에서는 짝수가 관측되고 다른 주사위에서는 홀수가 관측되는 경우 두 사건은 독립이다
- 사건 *A*와 *B*가 독립이면 $Pr(A \text{ and } B) = Pr(A) \times Pr(B)$
  - 참고: $A \cap B = \emptyset$인 두 사건 *A*와 *B*는 서로 배반 (disjoint), 즉 두 사건이 동시에 일어날 수 없음을 의미하고 *A*와 *B*는 서로 독립이 아니다.
- 한국의 왼손잡이 비율은 5%정도라고 알려져 있다. 임의로 2명을 뽑았을 때 모두 왼손잡이일 확률은?

예제

- 장원철 교수님은 자제가 2명이다. 한명이 아들이라는 사실을 알고 있을때 다른 한명이 아들일 확률은?

몬티 홀 문제

- 미국 게임 쇼 ``Let's Make a Deal'' 에서 유래한 퍼즐이다.
- 퍼즐의 이름은 이 게임 쇼의 진행자 몬티 홀의 이에서 따온 것이다.



https://theuijunkie.com/monty-hall-problem-explained/

몬티 홀 문제

- 퍼즐의 내용은 다음과 같다. 세 개의 문 중에 하나를 선택하여 문 뒤에 있는 선물을 가질 수 있는 게임쇼에 참가했다. 한 문 뒤에는 자동차가 있고, 나머지 두 문 뒤에는 아무것도 없다. 이때 어떤 사람이 예를 들어 1번 문을 선택했을 때, 게임쇼 진행자는 남아있는 두문 중 하나를 열어 문뒤에 아무것도 없음을 보여주면서 1번 대신 다른문을 선택하겠냐고 물었다.
- 선택을 바꾸는 것이 유리한지 여부를 토론해보자

몬티 홀 문제

- 여기서 표본공간 $S = \{(\omega_1, \omega_2) : \omega_i \in \{1,2,3\}\}$라 정의하고 ω_1을 자동차가 있는 문, ω_2를 도전자가 처음 선택하는 문이라 하자.
- 조건부 확률을 이용하여 설명해 보자
- 표본공간은 다음과 같다.
  $S = \{(1,1), (1,2), (1,3), (2,1), (2,2), (2,3), (3,1), (3,2), (3.3)\}$
- 선택을 바꾸지 않을 경우 당첨될 경우는 $(1,1), (2,2), (3,3)$
- 나머지 경우는 선택을 바꾸면 당첨이 된다!
- 다음 Web Applet을 사용하여 가상실험을 해보자.

http://www.rossmanchance.com/applets/MontyHall/Monty04.html

검사의 오류

- DNA 검사결과의 오류는 일반적으로 백만분의 1이라고 알려져 있다. 범행 현장에서 발견된 오직 유일한 증거인 DNA를 바탕으로 경찰에서 가지고 있는 10만명의 데이터베이스와 비교하여 1명이 일치하여 체포하였다.
- 법정에서 검사는 백만분의 1의 오차만 허용하는 DNA검사의 특성상 용의자가 범인이 확실하다고 주장한다. 과연 이 주장의 근거는 확실한가?

임신테스트

- 특정 임신테스트가 99% 정확하다고 하자.
- 제약회사에서는 다른 정밀한 방법으로 임신여부가 확인된 여성들을 대상으로 테스트 기계가 얼마나 임신여부를 정확히 예측하는냐에 따라 정확도를 보고한다.
- 즉 $\Pr(\text{test+} \mid \text{임신 O}) = \Pr(\text{test-} \mid \text{임신 X}) = 0.99$
- 하지만 내가 알고 싶은 것은 테스트 결과가 양성일때 실제로 임신했을 확률 $\Pr(\text{임신 O} \mid \text{test +})$이다.

임신테스트

	임신 O	임신 X
Test +	99	99
Test -	1	9801

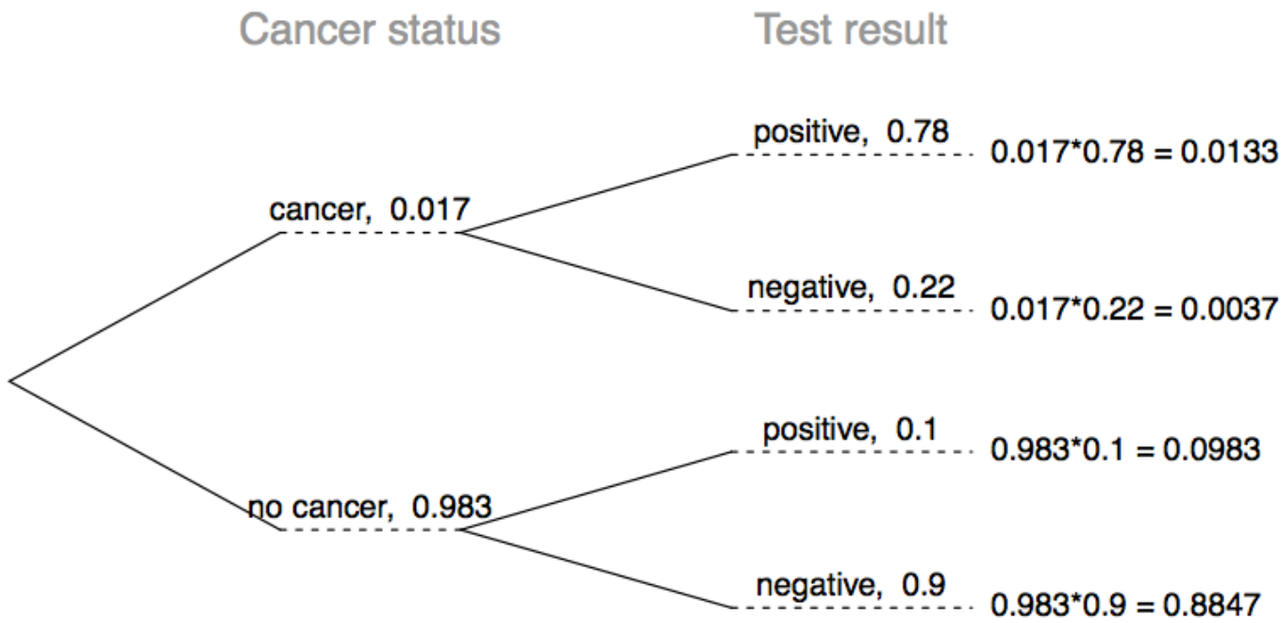
- 제약회사에서 총 10,000명의 가임기 여성을 대상으로 임신 테스트기의 정확성을 조사하고자 한다.
- 이중 실제 임신한 여성은 100명이다.
- 임신한 여성중 테스트 결과가 양성일 확률은 99/100=0.99이며 임신하지 않은 여성을 대상으로 할 경우 테스트 결과가 음성일 확률은 $\frac{9801}{9900} = 0.99$
- 하지만 테스트 결과가 양성인 198명중 실제로 임신한 경우는 99명으로 $\Pr(\text{임신 O} \mid \text{test +}) = 0.5$에 불과하다.

유방암 검사

- 미국 암협회에서 성인여자의 1.7%가 유방암환자라고 추정하고 있다.
- Susan G. Komen for the Cure Foundation에의하면 mammogram 검사를 통해서 유방암 환자중 78%를 정확히 진단할 수 있다.
- 하지만 2003년 Health Service Research에 발표된 연구에 의하면 mammogram 검사결과 10%의 false positive (정상인데 유방암이라고 오진하는 경우)가 발생한다.

확률나무 (Probability Tree)

만약 mammogram 결과가 양성이라면 실제로 환자가 유방암에 걸린 경우의 확률은 얼마나 될까?



$$\Pr(C|+) = \frac{\Pr(C \cap +)}{\Pr(+)} = \frac{0.0133}{0.0133 + 0.0983} = 0.12$$

베이즈 정리

- $A_1, \ldots, A_n$이 서로 배반이고 $\Pr(A_1 \cup A_2 \cdots \cup A_n) = 1$이라고 하자.
- 베이즈 정리:

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

임신테스트

베이즈 룰

- 위가 같은 경우 베이즈 룰을 이용하여 실제 계산을 할 수 있다!

$$\Pr(\text{임신 O}|\text{test +}) = \frac{\left(\frac{99}{100}\right)\cdot\left(\frac{100}{10000}\right)}{\left(\frac{99}{100}\right)\cdot\left(\frac{100}{10000}\right)+\left(\frac{99}{9900}\right)\cdot\left(\frac{9900}{10000}\right)} = 0.5$$

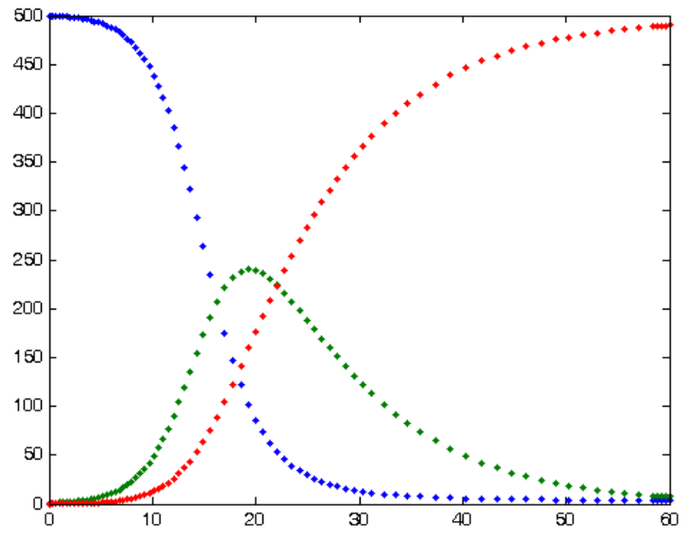
- 질병검사의 경우 유병률이 매우 낮고 (0.01%) 검사의 정확성이 99%정도라고 하면 검사의 결과가 양성이라고 해도 실제 질병에 걸려있을 확률은 1%밖에 되지 않는다!

- 비슷하게 DNA 검사의 경우 증거가 없기 때문에 10만명의 용의자가 범인일 확률은 동일하다고 가정하면 Pr(용의자가 진범) = 10만분의 1 이며
  Pr(용의자가 진범 | DNA 일치) = 0.9

- 만약 데이터베이스가 100만명일 경우 확률은 1/2로 줄어든다

예제: SIR model

- SIR model은 ordinary differential equation model로서 역학(epidemiology)에서 질병 전파를 설명하는 가장 보편적인 모형이다.
- SIR은 질병 전파 3단계인 susceptible, infected, recovered의 앞글자를 따온 것이다.



https://en.wikipedia.org/wiki/Compartmental_models_in_epidemiology

연습문제: SIR model

- 신종플루 유행시즌중 60%의 인구가 susceptible, 10%는 infected, 30%는 recovered라고 가정하자.
- 신종플루 검사의 정확성은 다음과 같다.
- susceptible: 95% (음성 판정비율)
- infected: 99% (양성판정 비율)
- recovered: 65% (음성 판정비율)
- 만약 검사결과가 양성일 경우 실제로 신종플루에 감염되었을 확률을 구하여 보아라.