확률 추론

I. 연속형 변수에 대한 확률

II. 열거에 의한 추론

I. 확률 밀도 함수

*연속 확률 변수*는 연속적인 값 집합을 결과 값으로 갖음

- 내일 최고 기온
- 노이즈나 불확실성이 포함된 힘 센서의 판독값
- 항공편의 실제 도착 시간 등

연속 확률 변수 X의 *누적 분포 함수*:

$$D(x) = P(X \leq x)$$

누적 분포 함수의 도함수는 확률 밀도 함수, *probability density function (pdf)*:

$$p(x) = \frac{d}{dx} D(x)$$

$$= \lim_{dx \to 0} \frac{P(X \leq x + dx) - P(X \leq x)}{dx}$$

$$= \lim_{dx \to 0} \frac{P(x < X \leq x + dx)}{dx}$$

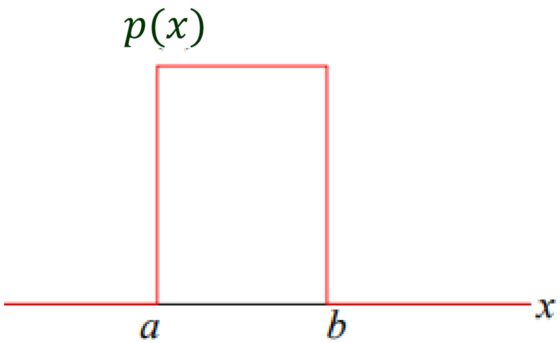
균등 분포

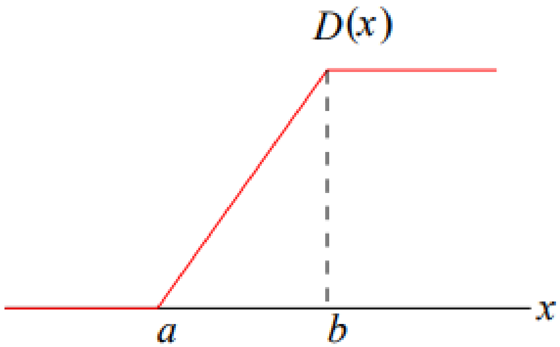
*균등 분포*는 일정한 확률 밀도 함수를 갖음

Example $[a, b]$ 구간에 걸쳐 분포하는 연속 균등 분포

$$p(x) = \begin{cases} 0, & \text{for } x < a, \\ \frac{1}{b-a}, & \text{for } a \le x < b, \\ 0, & \text{for } x \ge b. \end{cases}$$

$$D(x) = \begin{cases} 0, & \text{for } x < a, \\ \frac{x-a}{b-a}, & \text{for } a \le x < b, \\ 1, & \text{for } x \ge b. \end{cases}$$





기대값

확률 변수 X의 *기대값*(또는 *평균*) $E(X)$는 많은 확률 실험을 거친 후의 평균값

- n 회차의 실험
- m 개의 서로 다른 값: $x_1, \ldots, x_m$
- $1 \le i \le m$에 대해, 값 x_i가 각각 n_i 번 발생할 때 // $n = n_1 + \cdots + n_m$

$$E(X) = \frac{1}{n}\sum_{i=1}^{m} x_i \cdot n_i$$

Example 주사위를 무한히 여러 번 굴리면 각 숫자는 전체 횟수의 1/6 비율로 나타남

$$E(X) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{6} i \cdot \frac{n}{6} = \frac{7}{2}$$

분산과 표준 편차

$1 \le i \le m$에 대해 x_i 값의 확률이 p_i 라고 가정하면
X의 평균 μ 는

$$\mu = E(X) = p_1 x_1 + \cdots + p_m x_m.$$

확률 변수 X의 *분산*은 $x_1, \ldots, x_n$ 값들이 μ 주변에 퍼져 있는 정도를 측정함

$$\mathrm{var}(X) = E\left((X - E(X))^2\right) = \sum_{i=1}^{m} p_i (x_i - \mu)^2$$

$$= E((X-\mu)(X-\mu))$$

$$= E(X^2 - 2X\mu + \mu^2)$$

$$= E(X^2) - 2\mu^2 + \mu^2$$

$$= E(X^2) - \mu^2$$

확률 변수X의 *표준편차*는 $\sigma = \sqrt{\mathrm{var}(X)}$.

연속 변수의 평균과 분산

X가 확률 밀도 함수 $p(X)$를 가진 연속 변수일 때, 그 기대값:

$$\mu = E(X) = \int_{-\infty}^{\infty} x\, p(x)\, dx$$

분산:

$$\mathrm{var}(X) = \int_{-\infty}^{\infty} p(x)(x - \mu)^2\, dx$$

가우시안 분포

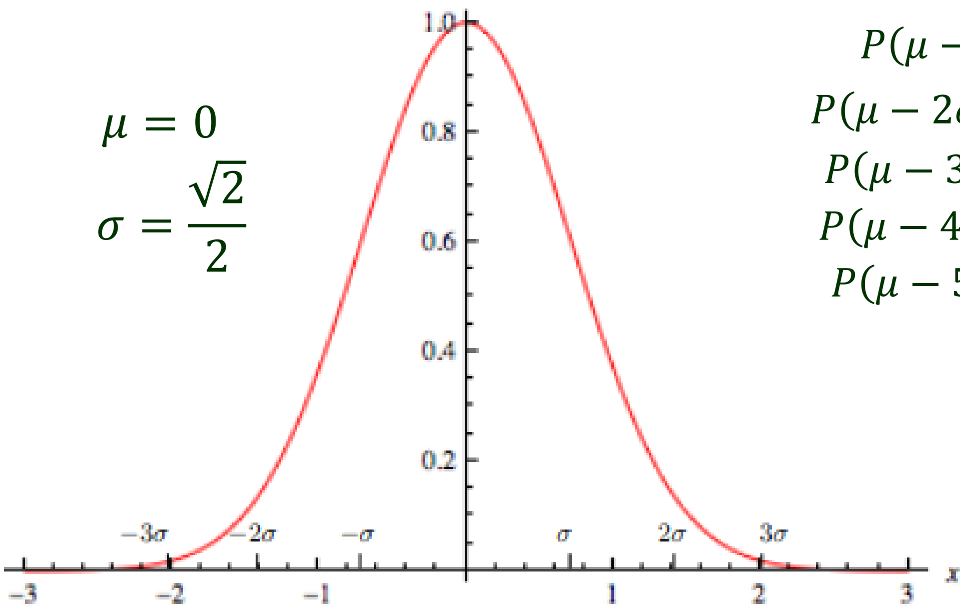
평균이 μ이고 분산이 σ^2이며 다음과 같은 확률 밀도 함수를 갖는 연속 확률 변수 X는 *가우시안 분포* (또는 *정규 분포*)를 따른다.

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

$\sqrt{\pi}\, p(x) = e^{-x^2}$

$\mu = 0$

$\sigma = \frac{\sqrt{2}}{2}$



$P(\mu - \sigma \le X \le \mu + \sigma) \approx 0.6826895$

$P(\mu - 2\sigma \le X \le \mu + 2\sigma) \approx 0.9544997$

$P(\mu - 3\sigma \le X \le \mu + 3\sigma) \approx 0.9973002$

$P(\mu - 4\sigma \le X \le \mu + 4\sigma) \approx 0.9999366$

$P(\mu - 5\sigma \le X \le \mu + 5\sigma) \approx 0.9999994$

실무에서는 주로 "3시그마 규칙" 적용:
(μ−3σ, μ+3σ) 구간 내의 모집단을 고려

가우시안 분포의 중요성

- 대부분의 자연 현상을 잘 모델링해주는 가장 중요한 분포 (예: 체중, 키, 체온 등 인간의 특성).

- 이는 n개의 독립적인 랜덤 변수 $X_1, \ldots, X_n$, 의 합 $X_1 + \cdots + X_n$이 $n \to \infty$ 일 때의 극한 분포(중심 극한 정리). 이는 다수의 독립적인 요인에 의해 영향을 받는 특성을 설명해줌

- 최소 제곱법, 칼만 필터 등 통계와 공학 분야에서 사용되는 중요한 방법들의 기초 이론

II. 열거에 의한 추론

확률적 추론: 관측된 증거를 기반으로 질의 명제에 대한 사후 확률을 계산

명제 ϕ의 확률을 계산: $$P(\phi) = \sum_{\omega \in \phi} P(\omega)$$

Joint distribution table:

	$toothache$		$\neg toothache$	
	$catch$	$\neg catch$	$catch$	$\neg catch$
$cavity$	0.108	0.012	0.072	0.008
$\neg cavity$	0.016	0.064	0.144	0.576

주변화

	$toothache$		$\neg toothache$	
	$catch$	$\neg catch$	$catch$	$\neg catch$
$cavity$	0.108	0.012	0.072	0.008
$\neg cavity$	0.016	0.064	0.144	0.576

$\neg$*toothache* (치통이 없는 경우)의 무조건(또는 주변) 확률:

$P(\neg toothache)$

주변화(Marginalization): 다른 변수들의 각 가능한 값에 대한 확률을 합산

≥ 1 variables　　　≥ 1 values

$$\boldsymbol{P(Y) = \sum_z P(Y, Z = z)}$$

변수 집합 $\boldsymbol{Z}$의 값에 대한 모든 가능한 조합을 합산

조건화(Conditioning)

Abbreviate $\mathbf{P}(Y, Z = z)$ as $\mathbf{P}(Y, z)$.

일반 값

$$\mathbf{P}(\mathit{Cavity}) = \mathbf{P}(\mathit{Cavity}, \mathit{toothache}, \mathit{catch}) + \mathbf{P}(\mathit{Cavity}, \mathit{toothache}, \neg\mathit{catch})$$
$$+ \mathbf{P}(\mathit{Cavity}, \neg\mathit{toothache}, \mathit{catch}) + \mathbf{P}(\mathit{Cavity}, \neg\mathit{toothache}, \neg\mathit{catch})$$

확률 변수는
대문자

$$= \langle 0.108, 0.016 \rangle + \langle 0.012, 0.064 \rangle + \langle 0.072, 0.144 \rangle + \langle 0.008, 0.576 \rangle$$

$$= \langle 0.2, 0.8 \rangle$$

cavity   *¬cavity*

굵은 폰트 아님

조건화 규칙:

$$\mathbf{P}(Y) = \sum_{z} \underbrace{\mathbf{P}(Y \mid z) P(z)}_{= \mathbf{P}(Y, z)}$$

	toothache		*¬toothache*	
	catch	*¬catch*	*catch*	*¬catch*
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

조건부 확률 계산

1. $P(a \mid b) = \dfrac{P(a \wedge b)}{P(b)}$ $\Longrightarrow$ 조건부 확률을 무조건 확률을 사용하여 표현
2. 전체 결합 분포에서 표현식을 평가

$$P(cavity \mid toothache) = \frac{P(cavity \wedge toothache)}{P(toothache)}$$

$$= \frac{0.108 + 0.012}{0.108 + 0.012 + 0.016 + 0.064} = 0.6$$

$$P(\neg cavity \mid toothache) = \frac{P(\neg cavity \wedge toothache)}{P(toothache)}$$

$$= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4$$

	toothache		*¬toothache*	
	catch	*¬catch*	*catch*	*¬catch*
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

정규화

	toothache		¬toothache	
	catch	¬catch	catch	¬catch
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

$$P(cavity \mid toothache) = \frac{P(cavity \wedge toothache)}{P(toothache)} = 0.6$$

$$P(\neg cavity \mid toothache) = \frac{P(\neg cavity \wedge toothache)}{P(toothache)} = 0.4$$

- $P(toothache)$는 $\mathbf{P}(Cavity \mid toothache)$분포에 대한 정규화 상수, 그 합은 1

- $P(toothache)$ 를 몰라도 $\mathbf{P}(Cavity \mid toothache)$ 계산 가능

$$\mathbf{P}(Cavity \mid toothache) = \alpha\, \mathbf{P}(Cavity, toothache)$$
$$= \alpha\, [\mathbf{P}(Cavity, toothache, catch) + \mathbf{P}(Cavity, toothache, \neg catch)]$$
$$= \alpha\, [\langle 0.108, 0.016\rangle + \langle 0.012, 0.064\rangle]$$
$$= \alpha\, \langle 0.108 + 0.012, 0.016 + 0.064\rangle = \alpha\, \langle 0.12, 0.08\rangle$$
$$= \langle 0.6, 0.4\rangle$$

정규화 상수 $\alpha = 5 = 1/P(toothache)$.

추론 프로시저

- X: 단일 질의 변수 (e.g., *Cavity*).
- $\boldsymbol{E}$: 증거변수 (e.g., *Toothache*).
- $\boldsymbol{e}$: 증거변수에 포집되는 관측 값
- $\boldsymbol{Y}$: 관측되지 않는 변수 (e.g., *Catch*).

$$\boldsymbol{P}(X \mid \boldsymbol{e}) \leftarrow \alpha \boldsymbol{P}(X, \boldsymbol{e}) = \alpha \sum_{\boldsymbol{y}} \boldsymbol{P}(X, \boldsymbol{e}, \boldsymbol{y})$$

♣ 시그마 합산은 $\boldsymbol{Y}$ 변수가 가질 수 있는 값들의 모든 가능한 조합 $\boldsymbol{y}$를 대상으로 함

♣ $\boldsymbol{P}(X, \boldsymbol{e}, \boldsymbol{y})$은 전체 결합 분포에서 선택된 확률의 부분집합

♣ 전체 결합 분포는 변수의 개수에 대해 지수 적으로 증가하기 때문에 실제로 계산하는 경우가 거의 없음

독립성

날씨라는 네 번째 변수 추가. 날씨 도메인은 { 해, 비, 구름, 눈 }

$$\mathbf{P}(Toothache, Catch, Cavity, Weather)$$

$$P(toothache, catch, cavity, cloud)$$

$$= P(cloud \mid toothache, catch, cavity)\, P(toothache, catch, cavity)$$

$$P(cloud \mid toothache, catch, cavity) = P(cloud)$$

$$= P(cloud)\, P(toothache, catch, cavity)$$

$$\mathbf{P}(Toothache, Catch, Cavity, Weather) = \mathbf{P}(Toothache, Catch, Cavity)\, \mathbf{P}(Weather)$$

8-element table + 4-element table

	toothache		*¬toothache*	
	catch	*¬catch*	*catch*	*¬catch*
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

~~A 32-element table.~~

독립 변수

- 두 명제 a와 b가 *독립*이면, 다음 조건을 만족

$$\underbrace{P(a \mid b) = P(a) \quad \text{or} \quad P(b \mid a) = P(b) \quad \text{or} \quad P(a \wedge b) = P(a)P(b)}_{\text{동치}}$$

- 두 확률 변수 X와 Y가 *독립*이면, 다음 조건을 만족

$$\boldsymbol{P}(X \mid Y) = \boldsymbol{P}(X) \quad \text{or} \quad \boldsymbol{P}(Y \mid X) = \boldsymbol{P}(Y) \quad \text{or} \quad \boldsymbol{P}(X \wedge Y) = \boldsymbol{P}(X)P(Y)$$

- 두 연속 확률 변수 X와 Y가 독립이면, 모든 $x, y \in \mathbb{R}$에 대해 다음 조건을 만족

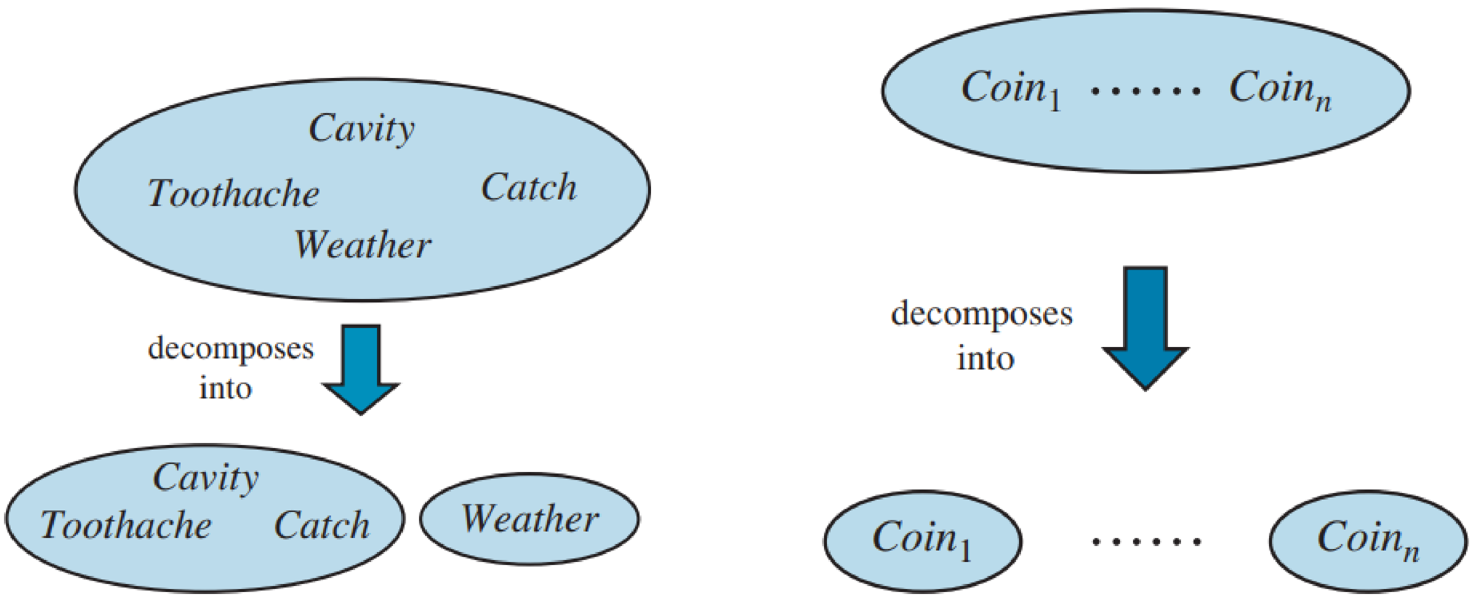
$$P(X \leq x \wedge Y \leq y) = P(X \leq x) \cdot P(Y \leq y)$$

결합 확률 밀도 함수 $p(x,y)$는 다음을 만족

$$p(x,y) = \int_{-\infty}^{\infty} p(x,y)\, dy \cdot \int_{-\infty}^{\infty} p(x,y)\, dx$$

결합 확률 분포의 분해

전체 결합 확률 분포는 *독립적인 변수의 부분 집합*에 대한 *개별 결합 분포*로 분해할 수 있음



(a) (b)

날씨와 치과 문제는 독립적임

동전 던지기는 매회 독립적임