* Gaussian Distribution (정규 분포, 가우시안 분포, 가우스 분포, 가우시안 확률 분포)

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

- $X \sim N(\mu, \sigma^2)$ μ는 분포 중심, 분산은 분포 폭 결정

Probability Density Function (PDF) $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ $(-\infty < x < \infty)$

* Markov Chain, MC (= Markov Process, MP) : Markov property를 지니는 이산 시간(discrete time) 확률 과정 (stochastic process)

- Markov property : 과거 상태들 $(s_1, s_2, \cdots, s_{t-1})$과 현재 상태 (s_t)가 주어졌을 때,
미래 상태 (s_{t+1})는 과거 상태와 독립적으로 현재 상태에 의해서만 결정

$P[s_{t+1} | s_t] = P[s_{t+1} | s_1, \cdots, s_t]$ (다음 state로 갈 확률은 항상 같다)

* Bayesian Rule : $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

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

$P(A|B)$: 사후확률 (posterior) 사건 B가 발생한 후 갱신된 사건 A의 확률
$P(A)$: 사전확률 (prior) 사건 B가 발생하기 전에 가지고 있던 사건 A의 확률
$P(B|A)$: 가능도 (likelihood) 사건 A가 발생한 경우 사건 B의 확률
$P(B)$: 정규화 상수 (normalizing constant) 또는 증거 (evidence) 확률의 크기 조정

· 확장 1
전체 확률의 법칙 $P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_i P(A_i, B)} = \frac{P(B|A_i)P(A_i)}{\sum_j P(B|A_j)P(A_j)}$

· 확장 2
추가적인 사건 C가 발생했다면 $P(A|B,C) = \frac{P(C|A,B)P(A|B)}{P(C|B)}$

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

$\left(+ P(A|B \cap C) = \frac{P(A \cap B \cap C)}{P(B \cap C)} = \frac{P(A \cap B|C)P(C)}{P(B \cap C)} = \frac{P(A|C)P(B|C)P(C)}{P(B \cap C)} = \frac{P(A|C)P(B \cap C)}{P(B \cap C)} = P(A|C)\right)$

* Likelihood

Likelihood : 고정된 관측값이 어떠한 확률 분포에서 어느정도의 확률로 나타나는지에 대한 확률 (관측된 사건 고정, 확률 분포 변화)
$L(\theta|x)$ (θ : 확률 분포 구성 parameter, x : 관측값, data)

Probability : 고정된 확률 분포에서 어떠한 관측값이 나타나는지에 대한 확률 (확률 분포 고정, 관측된 사건 변화)
$P(x|\theta)$

- 이산 사건 $L(\theta|x) = p(x|\theta) = p_\theta(x) = p_\theta(X=x)$

  이산사건에서 확률값 - 확률질량 함수 PMF (Probability Mass Function)

  $L(\theta|x) \rightarrow p(x|\theta)$ 사건 x에 대한 확률값 y, $y>0$, $\sum f(x) = 1$
  데이터가 사건 동시에 발생 확률 $p(x|\theta) = \prod_{k=1}^{n} p(x_k|\theta)$
  $L(\theta|x)$: 관측값이 주어질 때, 변화되는 확률분포에서 주어진 관측값이 나올 확률 / $p(x|\theta)$: 확률분포가 주어질 때, 변화하는 관측값이 나올 확률

- 연속 사건

  연속사건에서 확률 분포 - 확률 밀도 함수 PDF (probability Density function)

  $L(\theta|x) \rightarrow p(x|\theta)$ 사건 x에 대한 확률값 y, $y>0$, $\int f(x)dx = 1$

  $L(\theta|x)$: 특정 단일 사건 x가 발생할 확률 → PDF의 y값
  $p(x|\theta)$ 사건 범위가 발생할 확률 → PDF의 밑의 면적값 ($p(a \le x \le b) = \int_a^b f(x)dx$)

* MLE 최대 우도 추정법 Maximum Likelihood Estimation : 관측되는 데이터들을 가장 잘 모델링하는 확률분포의 parameter를 찾는 알고리즘

- 확률 분포 · 관측값 x를 입력으로 확률값 y를 출력하는 확률 밀도 함수 PDF로 정의
- parameter (θ)를 변수로 두었을 때, peak (미분했을 때 0이 되는 지점) 1개
- Likelihood를 계산하기 위한 확률분포들은 iid를 따름 (독립 항등 분포 · variable의 확률 분포가 independent, identically distribution)

• Exponential Distribution

확률 분포 parameter 1개 / $y = \lambda e^{-\lambda x}$ (λ : 확률 분포 parameter)

λ에 대한 L의 수학적 정의 : $L(\lambda|x_1, x_2, \dots x_n) = L(\lambda|x_1)L(\lambda|x_2)\cdots L(\lambda|x_n)$ (iid)

$$= \lambda e^{-\lambda x_1} \lambda e^{-\lambda x_2} \cdots \lambda e^{-\lambda x_n} = \lambda^n [e^{-\lambda x_1} e^{-\lambda x_2} \cdots e^{-\lambda x_n}]$$
$$= \lambda^n [e^{-\lambda(x_1 + x_2 + \cdots + x_n)}]$$

Likelihood의 최댓값 (peak) → 미분했을 때 0인 지점

$$\frac{d}{d\lambda} L(\lambda|x_1, x_2, \cdots, x_n) = \frac{d}{d\lambda} \lambda^n [e^{-\lambda(x_1 + x_2 + \cdots + x_n)}]$$

$$\frac{d}{d\lambda} \log(\lambda^n [e^{-\lambda(x_1 + x_2 + \cdots + x_n)}]) = \frac{d}{d\lambda} \log(\lambda^n) + \log[e^{-\lambda(x_1 + x_2 + \cdots + x_n)}]$$
$$= \frac{d}{d\lambda} n \log(\lambda) - \lambda(x_1 + x_2 + \cdots + x_n)$$
$$= n\frac{1}{\lambda} - (x_1 + x_2 + \cdots + x_n)$$

$$(x_1 + x_2 + \cdots + x_n) = n\frac{1}{\lambda}$$
$$\lambda(x_1 + x_2 + \cdots + x_n) = n$$
$$\therefore \lambda = \frac{n}{(x_1 + x_2 + \cdots + x_n)}$$

$\mu = \frac{1}{\lambda}$, $\sigma^2 = \frac{1}{\lambda^2}$

- Gaussian Distribution

  확률 분포 parameter 2개 : 평균(μ), 분산편차(σ) / $p(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$

  $\ln[L(\mu,\sigma|x)] = -\frac{n}{2}\ln(2\pi) - n\ln(\sigma) - \frac{(x_1-\mu)^2}{2\sigma^2} - \cdots - \frac{(x_n-\mu)^2}{2\sigma^2}$

  μ, σ에 대해 각각 편미분 $\frac{\partial}{\partial\mu}\ln[L(\mu,\sigma|x_1,\cdots,x_n)]$, $\mu = \frac{(x_1+\cdots+x_n)}{n}$

  $\frac{\partial}{\partial\sigma}\ln[L(\mu,\sigma|x_1,\cdots,x_n)]$, $\sigma = \sqrt{\frac{(x_1-\mu)^2+\cdots+(x_n-\mu)^2}{n}}$

- Log Likelihood : 미분할 때 0이 되는 지점에 영향 X, 미분 수식 계산 용이
- NLL (Negative Log Likelihood) : 최소값을 찾는 식

* 정보이론 기초

- Entropy

  - 개별정보 $H = n\log(s) = \log(s^n)$ (s: 각 선택에서 가능한 결과의 가지수, $\log s$: 구분해내기 위한 질문 개수, n: 구분하고자 하는 자수)

  - Entropy : 불확실성의 측정, H (bits)
    최선의 전략 하에서 그 사건을 판단하는 데 필요한 질문개수. 질문 개수에 대한 기대값
    모든 사건이 같은 확률로 일어날 때 그 최댓값을 갖는다
    entropy 감소 = 질문 개수 감소 = 정보량 감소

  이산 확률 분포일 때, (질문 개수는 각 사건 발생 확률의 역수)

  $H = \sum(\text{사건 발생 확률}) \cdot \log\left(\frac{1}{\text{사건 발생 확률}}\right) = \sum_i p_i \log_2\left(\frac{1}{p_i}\right) = -\sum_i p_i \log_2(p_i)$

- Cross Entropy

  - Cross Entropy : 어떤 문제에 대해 최적 전략을 쓸 때 예상되는 질문개수에 대한 기대값 (정확 : 확률분포)
    확률 분포로 된 어떤 문제 p에 대해 확률 분포로 된 어떤 전략 g를 사용할 때의 질문 개수의 기대값

  이산형

  $H(p,g) = \sum_i p_i \log_2\frac{1}{g_i} = -\sum_i p_i \log_2(g_i)$

  연속형

  $-\int p(x)\log g(x)\,dx$

  (여기서의 p_i: 각 확률에 대한 참값 또는 목표 확률, g_i: 현재 추측한 확률)

  - Binary classification (1나 0 두가지 중 하나만 추정)

    $-y\log\hat{y} - (1-y)\log(1-\hat{y})$ $p = [y, 1-y]$, $g = [\hat{y}, 1-\hat{y}]$ 인 cross entropy 식
    → logistic regression 에서 쓰는 cost function.

  - Cross Entropy → Log loss → negative log likelihood : cross Entropy 최소화 하는 것은 log likelihood 최대화

    어떤 데이터가 0 또는 1로 예측될 확률은 $\hat{y}$, $1-\hat{y}$, likelihood 식 → $\hat{y}^y(1-\hat{y})^{(1-y)}$
    $y=1$일 때 $\hat{y}$ 최대화, $y=0$일 때 $(1-\hat{y})$ 최대화. log를 씌우면
    maximize $y\log\hat{y} + (1-y)\log(1-\hat{y})$ = minimize $-y\log\hat{y} - (1-y)\log(1-\hat{y})$

No.

• KL - divergence (Kullback - Leibler divergence) : 두 확률분포의 엔트로피 차이 계산

$$H(p,q) = -\sum_i p_i \log q_i$$
$$= -\sum_i p_i \log q_i - \sum_i p_i \log p_i + \sum_i p_i \log p_i$$
$$= H(p) + \sum_i p_i \log p_i - \sum_i p_i \log q_i$$
$$= H(p) + \sum_i p_i \log \frac{p_i}{q_i}$$

분포 p와 분포 q의 엔트로피 차이 → KL-divergence

$\sum_i p_i \log \frac{p_i}{q_i} = H(p,q) - H(p)$ p와 q의 cross entrophy에서 p의 entrophy를 뺀 값 : 두 분포의 차이

$$KL(p\|q) = H(p,q) - H(p)$$

$$KL(p\|q) = \begin{cases} \sum_i p_i \log \frac{p_i}{q_i} \text{ 또는 } -\sum_i p_i \log \frac{q_i}{p_i} & \text{(이산형)} \\ \int p(x) \log \frac{p(x)}{q(x)} dx \text{ 또는 } -\int p(x) \log \frac{q(x)}{p(x)} dx & \text{(연속형)} \end{cases}$$

- $KL(p|q) \geq 0$
- $KL(p|q) \neq KL(q|p)$

* generative model vs discriminative model

- 생성모델 : posterior를 간접적으로 도출 / 데이터 밀도의 분포를 학습
  데이터 X가 생성되는 과정을 두 개의 확률 모형 $P(Y)$, $P(X|Y)$로 정의하여 $P(Y|X)$를 간접적으로 도출
- 판별모델 posterior를 직접적으로 도출 / 클래스 간 (decision boundary)을 학습
  데이터 X가 주어졌을 때 레이블 Y가 나타날 조건부 확률 $P(Y|X)$를 직접적으로 반환 EX. 선형 회귀, 로지스틱 회귀