Bayes Theorem

$$p(\theta \mid x) = \frac{p(x|\theta)\, p(\theta)}{p(x)}$$

- reveal how to update our belief about model parameters θ after observing data x
  - $p(x \mid \theta)$: likelihood
  - $p(\theta)$: prior

    - $p(\theta \mid x)$: posterior
    - $p(x)$: marginal likelihood (evidence)

Likelihood

- function of parameters
    - measure how well a set of parameters θ explains the observed data
    - not a probability distribution over the data x
- Gaussian model: $x \sim \mathcal{N}(\mu, \sigma^2)$

  Likelihood: $p(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

Maximum Likelihood Estimation (MLE)

- choose parameters that maximizes the likelihood of the observed data
    - without any prior belief
    - how to define the objective
        - not optimization method itself (e.g., gradient descent)
- usually maximize log-likelihood: $\arg\max_\theta \ \log p(x \mid \theta)$
    - usually corresponds to minimizing a negative log-likelihood loss
        - e.g., MSE, cross-entropy

Maximum A Posterior (MAP)

- similar to MLE but incorportate a prior belief about parameters θ
    - also likely under a prior distribution (regularization)
- prior: uniform → MAP: MLE

- concentrating on a few dominant directions
  - keeping top singular components (≃ SVD perspective)
  - capturing the most important adaptation directions efficiently
- memory efficient training
  - fine-tuning large pretrained models

Probability Distribution

- modeling uncertainty of random variables
  - mapping outcomes to probabilities
- discrete v.s. continuous distributions
  - probability mass function (PMF) v.s. probability density function (PDF)

Bernoulli Distribution

$$P(X = x) = p^x(1-p)^{1-x}, \quad x \in \{0, 1\}$$

- single binary trial (e.g., success / failure)
  - $X \in \{0, 1\}$
  - success probability p

Binomial Distribution

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- number of successes k in a fixed number of trials n
  - independent Bernoulli trials
  - success probability p
- statistics

    - mean: np
    - variance: $np(1-p)$
- normal approximation when n large

Multinomial Distribution

$P(\mathbf{X}=\mathbf{x}) = \frac{n!}{x_1!\cdots x_K!}\prod_{i=1}^{K} p_i^{x_i}$

- generalization of Bernoulli / Binomial to multiple categories
- used in multi-class classification
    - categorical likelihood
    - softmax output

Normal(Gaussian) Distribution

$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

- mean μ
- variance σ^2
- commonly used as a noise model in regression

Exponential Distribution

단위 시간 당 λ번의 사건 발생

$p(x) = \lambda e^{-\lambda x}, \quad x \geq 0$

다음 사건이 언제 발생할지 (걸리는 시간)

- models waiting time until the next event
    - time between events in a Poisson process
    - rate $\lambda > 0$
- mean: $1/\lambda$
- variance: $1/\lambda^2$
- memoryless property
    - $P(X > s+t \mid X > s) = P(X > t)$

- future waiting time independent of the past

Poisson Distribution

$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$ 단위 시간에 k번 사건이 발생할 확률

- event counts over time or space
  - unknown number of trials
- $X \in \{0, 1, 2, \ldots\}$
- event rate λ 단위 시간에 평균적으로 λ번 사건 발생
- Binomial limit when $n \to \infty, p \to 0$
- independent event assumption
- mean = variance = λ

Gamma Distribution

$p(x) = \frac{\lambda^k}{\Gamma(k)} x^{k-1} e^{-\lambda x}, \quad x \geq 0$

- generalization of the Exponential distribution
- models waiting time until the k-th event
  - shape k
  - rate λ
- mean: k/λ
- variance: k/λ^2
- when $k = 1 \to$ Gamma = Exponential

Central Limit Theorem

$X_1, \ldots, X_n \quad \text{i.i.d.}$

$\mathbb{E}[X_i] = \mu, \mathrm{Var}(X_i) = \sigma^2 < \infty$