BAYESIAN INFERENCE

Frequentist view vs Bayesian view :

Example: You try to evaluate if a coin is fair : $P(H) = P(T) = \frac{1}{2}$

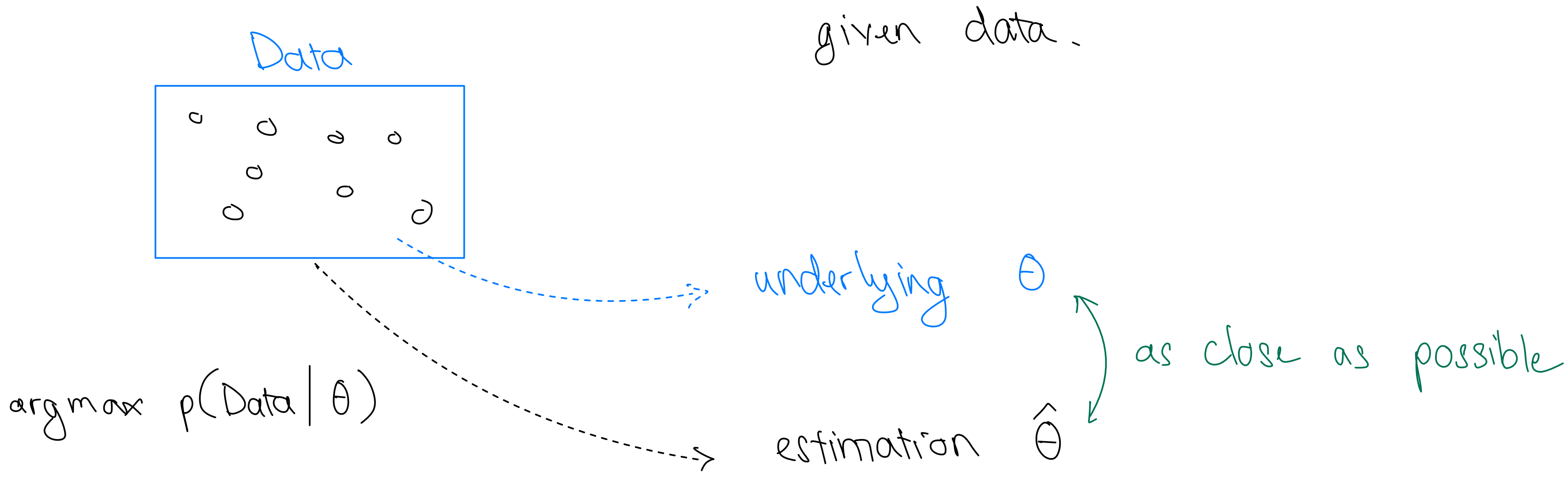
Frequentist	Bayesian
◦ Run experiments : flip the coin 1000 times	◦ Assume the coin is fair
◦ Collect data	◦ Collect data, update your assumption
◦ Evaluate if the coin is fair	◦ By 1000 times, conclude if it is fair or not

Main Idea of Bayesian Inference :

◦ Recall Maximum Likelihood Estimation, we trying to find the underlying θ given the dataset $p(\theta \mid D)$ by maximizing the probability of seeing the data given θ :

$$\underbrace{\hat{\theta}}_{\text{estimation}} = \underset{\theta}{\operatorname{argmax}} \; p(\underbrace{D}_{\text{given}} \mid \underbrace{\theta}_{\text{unknown}})$$

This is the frequentist view : estimating unknown parameter through given data.



◦ Look at the same problem under Bayesian View :

$$\underbrace{p(\theta \mid D)}_{\text{posterior}} = \frac{\underbrace{p(D \mid \theta)}_{\text{likelihood}} \cdot \underbrace{p(\theta)}_{\text{prior}}}{\underbrace{p(D)}_{\text{Normalization factor}}}$$

Normalization factor : $p(D) = \int p(D \mid \theta)\, p(\theta)\, d\theta$

$p(D)$ is treated as constant since we only care ab θ. So :

$$\boxed{p(\theta \mid D) \underbrace{\propto}_{\text{"proportional to"}} p(D \mid \theta) \cdot p(\theta)}$$

Bayes' rule

Example: Why Bayesian Inference is better than MLE (frequentist)

- In this example, we try to evalute the accuracy of an alarm that goes off when the sun explodes.

Given:
$$\begin{cases} \theta \in \{0,1\} & \text{is indicator if the sun explodes} \\ x \in \{0,1\} & \text{is indicator if the alarm fires} \\ \alpha \in [0,1] & \text{is the error rate of the alarm} \end{cases}$$

So:
$$\begin{cases} p(x=\theta \mid \theta) = 1-\alpha & \rightarrow \text{correctly fired} \\ p(x = 1-\theta \mid \theta) = \alpha & \rightarrow \text{incorrectly fired} \end{cases}$$

- Problem: If the alarm fires, should we believe the sun has exploded or not? Assume that this alarm is very accurate $\alpha = 0.0001$

1. Find Θ using MLE

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ p(x=1 \mid \theta) = \begin{cases} \alpha & \text{if } \theta = 0 \\ 1-\alpha & \text{if } \theta = 1 \end{cases}$$

$$\Rightarrow \hat{\Theta} = 1$$

This suggest that we should trust the alarm completely!

2. Find θ using Bayesian Inference

- Step 1: Determine prior (probability of θ without observing any data)

$$p(\theta) = \begin{cases} \beta & \text{if } \theta = 1 \\ 1-\beta & \text{if } \theta = 0 \end{cases}$$

$< \beta$ is very small, $\approx 0 >$

prior knowledge is what makes Bayesian better

- Step 2: Determine posterior

$$p(\theta \mid x=1) = \frac{p(x=1 \mid \theta) \cdot p(\theta)}{p(x=1)}$$
$$\propto p(x=1 \mid \theta) \cdot p(\theta)$$
$$= \begin{cases} (1-\alpha) \cdot \beta & \text{if } \theta = 1 \\ \alpha(1-\beta) & \text{if } \theta = 0 \end{cases}$$

- Step 3: Decide value of θ
  - If we predict $\theta = 1$, then:

$$p(\theta = 1 \mid x = 1) > p(\theta = 0 \mid x = 1)$$
$$\Leftrightarrow (1-\alpha)\beta > \alpha(1-\beta)$$
$$\Leftrightarrow \frac{\beta}{1-\beta} > \frac{\alpha}{1-\alpha}$$

This is not true since $\beta \simeq 0$ and $\alpha = 0.0001$

So the opposite holds, meaning $\theta = 0$

Example: Bayesian Inference on Gaussian Distribution

- You just moved to a new apartment.
- Your friend told you the commute time is 30 ± 10 min.
- You drove yourself and recorded the time: 25, 45, 30, 50.

Problem: How should you predict the commute time?

Solution:

- Let θ be the time (treat θ as random variable)

Step 1: Determine prior, we assume θ is Normally distributed

$$P(\theta) \sim N(\mu_0, \sigma_0^2) \quad \text{where} \begin{cases} \mu_0 = 30 \\ \sigma_0 = 10 \end{cases}$$

assign a reasonable value

Step 2: Observe the data collected.

- Let D be the dataset $\{x_i\}_{i=1}^n$
- Assume the data observed have some noise:

$$x_i = \underbrace{\theta}_{\text{ground truth (true parameter)}} + \underbrace{\sigma_1 \varepsilon_i}_{\text{noise}} \quad \text{where} \begin{cases} \varepsilon_i \sim N(0,1) \\ \sigma_1 : \text{variance} = \text{some value} \end{cases}$$

- From above assumption, we can say that each data point drawn i.i.d. from Gaussian Distribution:

$$P(x_i \mid \theta) \sim N(\theta, \sigma_1^2) \Rightarrow P(x_i \mid \theta) = \frac{1}{\sqrt{2\pi}\,\sigma_1} \exp\left(-\frac{(x_i - \theta)^2}{2\sigma_1^2}\right)$$

Step 3: Determine likelihood function

$$P(D \mid \theta) = \prod_{i=1}^n P(x_i \mid \theta)$$

Step 4: Determine posterior

$$P(\theta \mid D) = \frac{P(D \mid \theta) \cdot P(\theta)}{\underbrace{P(D)}_{\text{constant}}}$$

$$\propto P(D \mid \theta) \cdot P(\theta)$$

$$= \left[\prod_{i=1}^n P(x_i \mid \theta)\right] \cdot P(\theta)$$

$$\propto \left[\prod_{i=1}^n \exp\left(-\frac{(x_i - \theta)^2}{2\sigma_1^2}\right)\right] \cdot \exp\left(-\frac{(\theta - \mu_0)^2}{2\sigma_0^2}\right)$$

$$\propto \exp\left[-\sum_{i=1}^n \left(\frac{(\theta - x_i)^2}{2\sigma_1^2}\right) - \frac{(\theta - \mu_0)^2}{2\sigma_0^2}\right]$$

$$= \exp\left[-\frac{1}{2}\left(A\,\theta^2 - 2B\,\theta + \text{Const}\right)\right]$$

Why remove this even though σ_1 is a r.v.? Because the exponent part is much more influential

$$\boxed{\begin{aligned} A &= \sum^n \frac{1}{\sigma_1^2} + \frac{1}{\sigma_0^2} \\ B &= \sum^n \frac{x_i}{\sigma_1^2} + \frac{\mu_0}{\sigma_0^2} \end{aligned}}$$

$$= \exp\left(-\frac{1}{2} A \left(\theta - \frac{B}{A}\right)^2 + \text{Const}\right)$$

$$\sim \boxed{N\left(\frac{B}{A}, \frac{1}{A}\right)}$$

So the posterior is Normally Distributed:

$$p(\theta \mid D) \sim N\left(\frac{B}{A}, \frac{1}{A}\right)$$

with:

posterior mean: $$\mu_p = \frac{B}{A} = \frac{\sum^n \frac{x_i}{\sigma_1^2} + \frac{\mu_0}{\sigma_0^2}}{\frac{n}{\sigma_1^2} + \frac{1}{\sigma_0^2}}$$

weighted sum of data
weight $= \frac{1}{\sigma_1^2}$ which is the "inverse" noise

weighted prior
weight $= \frac{1}{\sigma_0^2}$

posterior variance: $$\sigma_p^2 = \frac{1}{A} = \left(\frac{n}{\sigma_1^2} + \frac{1}{\sigma_0^2}\right)^{-1}$$

Some observations:

- If $n = 0$ (no data): $\begin{cases} \mu_p = \mu_0 \\ \sigma_p^2 = \sigma_0^2 \end{cases}$
- If $n > 0$ (have data): $\begin{cases} \mu_p \text{ as stated above} \\ \sigma_p^2 \text{ as stated above} \end{cases}$
- If $n \to \infty$ (infinite data): $\begin{cases} \mu_p \approx \frac{\sum x_i}{n} \\ \sigma_p^2 \approx \frac{\sigma_1^2}{n} \end{cases}$

Estimate θ on μ_p and σ_p

In Bayesian, what is considered fixed and what is considered random variable ?

- Fixed: Observed dataset
  i.e.: $\{x_i\}_{i=1}^n$
  $\{x_i, y_i\}_{i=1}^n$
  ...
- Random variable:
  - Parameters: θ
  - Noise: σ_1 ε_1
  - Dataset before observed:
    i.e.: $x_i \mid \theta$
    $y_i \mid x_i, \theta$

Example: Bayesian Linear Regression

- Given data points $\{x_i, y_i\}_{i=1}^{n}$
- <u>Problem:</u> Find θ such that $y \approx x^T\theta$

$$y = x^T\theta = (x_1 \ 1)^T \begin{pmatrix}\theta \\ b\end{pmatrix} = x_1\theta + b$$

Solved with LLS and MLE

- Linear Least Square (LLS): $\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - x_i^T\theta)^2$

  or $\underset{\theta}{\operatorname{argmin}} \| y - X^T \theta \|_2$

  vector $\begin{pmatrix}\theta_1 \\ \vdots \\ b\end{pmatrix}$

  vector

  dataset matrix

$$\begin{array}{cc} & \text{features} \quad \text{bias} \\ \text{records} & \begin{bmatrix} \vdots & \vdots & \vdots & \vdots & 1 \\ & & & & 1 \\ & & & & 1 \end{bmatrix}\end{array}$$

- Maximum Likelihood Estimation (MLE): $\hat{\theta} = \underset{\theta}{\operatorname{argmax}} P(D \mid \theta)$

$\Rightarrow$ Both methods yield a deterministic value $\hat{\theta}$ (meaning fixed)

Solve with Bayesian Inference

- Bayesian method gives us an uncertainty estimation to see how accurate our estimation is.
- <u>Step 1:</u> Determine prior

  $P(\theta) \sim N(\mu_0, \sigma_0^2)$, where $\begin{cases} \mu_0 = 0 \\ \sigma_0^2 = \text{some large number} \end{cases}$

  default value of no prior knowledge

- <u>Step 2:</u> Determine likelihood function
  - Let D be data set $\{x_i, y_i\}_{i=1}^{n}$
  - There should be some noise in data observed:

    $y_i = x_i^T\theta + \sigma_1^2 \varepsilon_i$ where $\begin{cases} \varepsilon_1 \sim N(0,1) \\ \sigma_1^2 : \text{variance} \end{cases}$

    noise

  - From above assumption, the likelihood function for each data point is:

$$P(\{y_i, x_i\} \mid \theta) = P(y_i \mid x_i, \theta) \cdot P(x_i)$$

fixed, dont care

$$\propto P(y_i \mid x_i, \theta)$$

Because $y_i \mid x_i, \theta \sim N(x_i^T\theta, \sigma_1^2)$

$$= \boxed{\frac{1}{\sqrt{2\pi}\,\sigma_1} \exp\left(-\frac{(y_i - x_i^T\theta)^2}{2\sigma_1^2}\right)}$$

Because exponent part is more influential

$$\propto \exp\left(-\frac{(y_i - x_i^T\theta)^2}{2\sigma_1^2}\right)$$

- So the likelihood function for whole dataset is

$$P(D|\theta) = \prod_{i=1}^{n} P(y_i | x_i, \theta)$$

$$\propto \prod_{i=1}^{n} \exp\left(-\frac{(y_i - x_i^T\theta)^2}{2\sigma_1^2}\right)$$

<u>Step 4:</u> Determine posterior

$$P(\theta|D) \propto P(D|\theta) \cdot P(\theta)$$

$$\propto \left[\prod_{i=1}^{n} \exp\left(-\frac{(y_i - x_i^T\theta)^2}{2\sigma_1^2}\right)\right] \cdot \exp\left(-\frac{(\theta - \mu_0)^2}{2\sigma_0^2}\right)$$

$$= \exp\left[\sum_{i=1}^{n}\left(-\frac{(y_i - x_i^T\theta)^2}{2\sigma_1^2}\right) - \frac{(\theta - \mu_0)^2}{2\sigma_0^2}\right]$$

$$= \exp\left[-\frac{1}{2}\left(\theta^T A\,\theta - 2B^T\theta + \text{Const}\right)\right]$$

$$\sim N(A^{-1}B, A^{-1})$$

$$A = \sum^{n} \frac{x_i x_i^T}{\sigma_1^2} + \frac{I}{\sigma_0^2}$$

$$B = \sum^{n} \frac{y_i x_i}{\sigma_1^2} + \frac{\mu_0}{\sigma_2^2}$$

So the posterior is Normally Distributed

$$P(\theta|D) \sim N(A^{-1}B, A^{-1})$$

with:

- posterior mean: $\mu_p = \left(\sum^{n} \frac{x_i x_i^T}{\sigma_1^2} + \frac{I}{\sigma_0^2}\right)^{-1}\left(\sum^{n} \frac{y_i x_i}{\sigma_1^2} + \frac{\mu_0}{\sigma_0^2}\right)$
- posterior variance: $\sigma_p^2 = \left(\sum^{n} \frac{x_i x_i^T}{\sigma_1^2} + \frac{I}{\sigma_0^2}\right)^{-1}$

Some observations:

- If $n = 0$ (no data): $\begin{cases} \mu_p = \mu_0 \\ \sigma_p^2 = \sigma_0^2 \end{cases}$
- If $n > 0$ (have data): as stated above
- If $n \to \infty$ (infinite data): $\begin{cases} \mu_p = \left(\sum^{\infty} \frac{x_i x_i^T}{\sigma_1^2}\right)^{-1}\left(\sum^{\infty} \frac{y_i x_i}{\sigma_1^2}\right) \\ \sigma_p^2 = \left(\sum^{\infty} \frac{x_i x_i^T}{\sigma_1^2}\right)^{-1} \end{cases}$

Important notes:

The distribution of prior, likelihood and posterior can be different.

In total, there are 3 distributions:

- Dataset distribution (likelihood)
- Parameter distribution (prior)
- Parameter | Dataset distribution (posterior)