$\boxed{\text{Bayesian Inference}}$:

Recall MLE:

$$\hat{\theta} = \underset{\theta}{\arg\max}\left[p(\overset{\text{data}}{D} \mid \overset{\text{unknown parameter/variable}}{\theta}) \right]$$

Here, parameter θ is unknown but deterministic (frequentist view).

<u>Bayesian</u>:

- θ is viewed as a random variable (even when it is actually deterministic).
- use Bayes' Rule to calculate posterior distribution.

$$\underbrace{p(\theta \mid D)}_{\text{posterior distribution}} = \frac{\overbrace{p(D \mid \theta)}^{\text{likelihood}} \overset{\text{prior}}{p(\theta)}}{p(D)} \propto p(D \mid \theta)\, p(\theta).$$

$$p(D) = \int p(D \mid \theta)\, p(\theta)\, d\theta.$$

<u>Proof of Baye's Rule</u>:

Multiplication Rule:

$$p(\theta \cap D) = p(\theta)\, p(D \mid \theta)$$
$$p(D \cap \theta) = p(D)\, p(\theta \mid D)$$

same / same.

$$\Rightarrow p(\theta)\, p(D \mid \theta) = p(D)\, p(\theta \mid D)$$

$$\Rightarrow p(\theta \mid D) = \frac{p(D \mid \theta)\, p(\theta)}{p(D)}.$$

Example: Did the sun just explode?

- Suppose we have a device that detects if the sun explodes with high accuracy:

$$P(X=\theta \mid \theta) = 1-\alpha \qquad , \qquad \alpha = \text{error, known \& fixed.}$$
$$P(X=1-\theta \mid \theta) = \alpha \qquad \qquad \uparrow \text{small}$$

where

$\theta \in \{0,1\}$ = did the sun exploded. //binary

$X \in \{0,1\}$ = did the device alarms. //binary.

- If alarm fires $(X=1)$, should we believe the sun exploded or not?

MLE:

$$\hat{\theta} = \underset{\theta \in \{0,1\}}{\arg\max} \left[P(X=1 \mid \theta) \right] \qquad = \begin{cases} \alpha, & \text{for } \theta = 0 \\ 1-\alpha, & \text{for } \theta = 1 \end{cases}$$

$$= 1$$

Bayes:

Step 1: Find the prior.

$$P(\theta) = \begin{cases} \beta \text{ (very very small)} & \text{for } \theta = 1, \\ 1-\beta & \text{for } \theta = 0. \end{cases}$$

Step 2: Set up equation for posterior:

$$P(\theta \mid X=1) = \frac{P(X=1 \mid \theta)\, P(\theta)}{P(X=1)} \propto P(X=1 \mid \theta)\, P(\theta)$$

$$= \begin{cases} \underbrace{(1-\alpha)}_{P(x=1\mid\theta=1)} \underbrace{\beta}_{P(\theta=1)}, & \text{if } \theta = 1 \\ \underbrace{\alpha}_{P(x=1\mid\theta=0)} \underbrace{(1-\beta)}_{P(\theta=0)}, & \text{if } \theta = 0, \end{cases}$$

If $(1-\alpha)\beta > \alpha(1-\beta)$: predict $\theta = 1$,

If $(1-\alpha)\beta < \alpha(1-\beta)$: predict $\theta = 0$.

Equivalently:

predict $\theta = 1$, if $\dfrac{\beta}{1-\beta} > \dfrac{\alpha}{1-\alpha}$

predict $\theta = 0$, if $\dfrac{\beta}{1-\beta} < \dfrac{\alpha}{1-\alpha}$

$$\boxed{\text{Posterior} \propto \text{Likelihood} * \text{Prior}.}$$

Example: Predicting Commute Time:

- You moved to new apartment. Friend said commute time is 30 ± 10 mins.
- You drove a few times, and time $= \{25, 45, 30, 50\}$
- How should you predict commute time?

Prior: Assume $P(\theta) \sim N(\mu_0, \sigma_0^2)$, $\mu_0 = 30$, $\sigma_0 = 10$.

$$= \frac{1}{2\pi\sigma_0} \exp\left(\frac{-(\theta-\mu_0)^2}{2\sigma_0^2}\right) \propto \exp\left(\frac{-(\theta-\mu_0)^2}{2\sigma_0^2}\right)$$

Likelihood: Based on observation / data.

observe: $x_1, \dots, x_n$. Assume noise in observation.

$$x_i = \theta + \sigma_1 \xi_i, \quad \xi_i \sim N(0,1), \quad \sigma_1 = 5.$$

$$p(x_i \mid \theta) \sim N(\theta, \sigma_1^2)$$

ith observation

theoretical value

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

Posterior:

$$p(\theta \mid D) = \frac{P(D\mid\theta)\,P(\theta)}{P(D)} \propto P(D\mid\theta)\,P(\theta).$$

doesn't matter much; normalization constant.

$$= \underbrace{\prod_{i=1}^{n} p(x_i \mid \theta)}_{P(D\mid\theta) = \text{data.}} P(\theta)$$

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

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

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

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

$$= \exp\left[-\frac{1}{2}\left[\theta^2\overbrace{\left(\sum_{i=1}^{n}\left(\frac{1}{\sigma_1^2}\right)+\frac{1}{\sigma_0^2}\right)}^{A} - 2\theta\overbrace{\left(\sum_{i=1}^{n}\left(\frac{x_i}{\sigma_1^2}\right)+\frac{\mu_0}{\sigma_0^2}\right)}^{B} + \underbrace{\left(\sum_{n=1}^{n}\left(\frac{x_i^2}{\sigma_1^2}\right)+\frac{\mu_0^2}{\sigma_0^2}\right)}_{C}\right]\right]$$

$$= \exp\left(-\frac{1}{2}\left(A\theta^2 - 2B\theta + C\right)\right),$$ where

$$A = \sum_{i=1}^{n}\left(\frac{1}{\sigma_1^2}\right) + \frac{1}{\sigma_0^2} = \frac{n}{\sigma_1^2} + \frac{1}{\sigma_0^2}$$

$$B = \sum_{i=1}^{n}\left(\frac{x_i}{\sigma_1^2}\right) + \frac{\mu_0}{\sigma_0^2}$$

C = constant (bc θ not involved).

// Recall Gaussian distribution

$$\propto \exp\left[\left(-\frac{1}{2}\right)\frac{1}{\sigma^2}(x-\mu)^2\right].$$

$$= \exp\left(-\frac{1}{2}\overset{\frac{1}{\text{variance}}}{\overset{\downarrow}{A}}\left(\theta - \overset{\text{mean}}{\overset{\downarrow}{\frac{B}{A}}}\right)^2 + \text{const.}\right) \sim N\left(\frac{B}{A}, \frac{1}{A}\right)$$

$$\boxed{\begin{aligned} \mu_{\text{posterior}} &= \frac{B}{A} = \frac{\left(\sum_{i=1}^{n}\frac{x_i}{\sigma_1^2} + \frac{\mu_0}{\sigma_0^2}\right)}{\left(\frac{n}{\sigma_1^2} + \frac{1}{\sigma_0^2}\right)} \\ \sigma^2_{\text{posterior}} &= \frac{1}{A} = \left(\frac{n}{\sigma_1^2} + \frac{1}{\sigma_0^2}\right)^{-1} \end{aligned}}$$

If $n=0$ (no data), $\mu_p = \frac{\mu_0/\sigma_0^2}{1/\sigma_0^2} = \mu_0$.

$$\sigma_p^2 = \sigma_0^2.$$

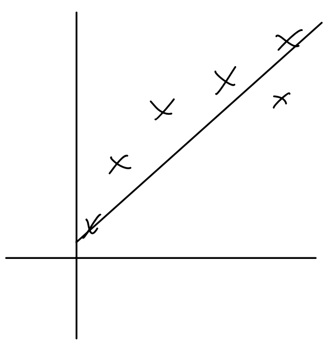
If $n = \infty$ (infinite data), $\mu_p \approx \dfrac{\sum_{i=1}^{n} \frac{x_i}{\sigma_1^2} + 0}{\frac{n}{\sigma_1^2} + 0} \approx \frac{1}{n} \sum_{i=1}^{n} x_i$

$$\sigma_p^2 \approx \left(\frac{n}{\sigma_1^2} \right)^{-1} = \frac{\sigma_1^2}{n}$$

If $n > 0$ (have data), μ_p = weighted sum of data and prior, where weight $= \dfrac{1}{\sigma_1^2}$

Bayesian Linear Regression: Goal is to use Bayesian Inference to solve Linear Regression.

- Given data points $\{x_i, y_i\}_{i=1}^{n}$, want to find θ, such that $y \approx x^T\theta$.



$y = x^T\theta + b \Rightarrow$ can incorporate b into x and θ using the following augmentation:

$$\underbrace{[x \quad 1]}_{\tilde{x}^T} \underbrace{\begin{bmatrix} \theta \\ b \end{bmatrix}}_{\tilde{\theta}} = \tilde{x}^T \tilde{\theta}$$

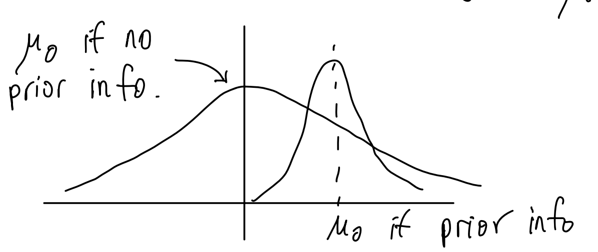
Least Square Method:

$$\hat{\theta} = \underset{\theta}{\arg\min} \sum_{i=1}^{n} (y_i - x_i^T\theta)^2$$

Bayesian Inference: Treat θ as a random variable

Prior: Assume Gaussian Distribution; $P(\theta) \sim N(\mu_0, \sigma_0^2)$

The prior depends on the user. If you have information on θ, then maybe you center μ_0 at that, and make σ small. If no prior information, then set $\mu_0 = 0$, $\sigma_2^2 =$ large number.



Likelihood: Assume $y_i = x_i^T \theta + \sigma_1 \xi_i$, where

σ_1 : variance,

$\xi \sim N(0, 1)$

$$\Rightarrow P(\{y_i, x_i\} \mid \theta) = \underbrace{P(y_i \mid x_i, \theta)}_{\text{Gaussian.}} \underbrace{p(x_i)}_{\substack{\text{constant} \\ \text{wrt } \theta.}}$$

// multiplication rule?

Posterior:

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

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

$$= \left[\prod_{i=1}^{n} P(y_i \mid x_i, \theta) \underbrace{P(x_i)}_{\text{does not depend on } \theta.} \right] P(\theta)$$

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

actual, predicted

prior: can be known or unknown

$$\propto \underbrace{\left[\prod_{i=1}^{n} \exp\left(- \frac{(y_i - x_i^T \theta)^2}{2\sigma_1^2} \right) \right]}_{P(D \mid \theta) = \prod_{i=1}^{n} P(y_i \mid x_i, \theta) \cancel{P(x_i)}} \underbrace{\exp\left[- \frac{(\theta - \mu_0)^2}{2\sigma_0^2} \right]}_{P(\theta)}$$

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

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

$$= \exp\left[-\frac{1}{2}\left(\underbrace{\theta^T A \theta}_{\propto\, \theta^2} - \underbrace{2B^T\theta}_{\propto\, \theta} + \text{const} \right) \right] \sim N\left(\underbrace{A^{-1} B}_{\mu_p} ,\ \underbrace{A^{-1}}_{\sigma_p^2} \right)$$

where $$A = \sum_{i=1}^{n} \frac{x_i x_i^T}{\sigma_1^2} + \frac{I}{\sigma_0^2}$$ Identity matrix.

$$B = \sum_{i=1}^{n} \frac{y_i x_i}{\sigma_1^2} + \frac{\mu_0}{\sigma_0^2}$$

$$\theta = \underbrace{\begin{bmatrix} \\ \\ \\ \end{bmatrix}}_{d \times 1} \qquad A = \underbrace{\begin{bmatrix} & & \\ & & \\ & & \\ \end{bmatrix}}_{d \times d}.$$