1

Statistical Estimation: MLE, MAP, Full Bayesian

MLE

If India & Australia plays tomorrow, will India win or lose?

Past data: Win, Loss, Loss, Win, Win

So, $\boxed{P(\text{India Wins}) = \frac{3}{5}}$

Now we'll fit a statistical model & estimate MLE, and see whether we can obtain the same answer.

There can be two outcomes: {Win, Loss}

View this as a random variable with 2 states:

$$Y = \{0, 1\}$$

Let the hypothesis class is:

$$Y \sim \text{Bernoulli}(\theta)$$

$$Y = \begin{cases} 1, & \text{prob. } \theta \\ 0, & \text{prob. } (1-\theta) \end{cases}$$

↓

the estimation problem is to find θ.

The dataset is: $D = \{1, 0, 0, 1, 1\}$

Some trivial choices: $\left.\begin{array}{l}\hat{\theta} = 0 \\ \hat{\theta} = 1\end{array}\right\}$ bad estimators.

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \underbrace{P(D|\theta)}_{\text{Likelihood}}$$

[Likelihood is a function of the parameters θ]

- Let $D = \{1\} \Rightarrow L(\theta) = P(Y=1|\theta) = \theta$

[I.I.D. Samples] (independency)

- $D = \{1, 0\} \Rightarrow L(\theta) = P(Y=1|\theta)\, P(Y=0|\theta)$
$= \theta(1-\theta)$
- $D = \{1, 1, 1\} \Rightarrow L(\theta) = \theta^3$

$\alpha_H \rightarrow$ # of wins

$\alpha_T \rightarrow$ # of losses

In general: $L(\theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$

$$\hat{\theta}_{MLE} = \arg\max_{\theta} L(\theta) = \arg\max_{\theta} \text{Log}\, L(\theta)$$

[monotonic fn]

$$\Rightarrow \hat{\theta}_{MLE} = \arg\max_{\theta} [\alpha_H \log\theta + \alpha_T \log(1-\theta)]$$

$$\frac{d}{d\theta}[\alpha_H \log\theta + \alpha_T \log(1-\theta)] = 0$$

$$\Rightarrow \frac{\alpha_H}{\theta} - \frac{\alpha_T}{1-\theta} = 0 \qquad (\text{assume } \theta \neq 1, \neq 0)$$

$$\Rightarrow \hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

So in our dataset, $\alpha_H = 3$, $\alpha_T = 2$

Hence, $\boxed{\hat{\theta}_{MLE} = \frac{3}{5}}$ (matches with our previous estimation)

Why Max-Likelihood?

- Leads to "natural" estimators
- MLE is optimal if model-Class (hypothesis) is correct.

Let y is a discrete random variable:

$$Y = y \in \{1 \dots k\}$$

Consider log-Likelihood (for some model) average:

$$\frac{1}{N} LL(\theta) = \frac{1}{N}\sum_{i=1}^{N} \log p(y_i | \theta)$$

$$= \frac{1}{N}[\#(Y=1)\log p(Y=1|\theta) + \#(Y=2)\log p(Y=2|\theta) + \dots]$$

With ∞ data,

$$\lim_{N\to\infty} \frac{\#(Y=1)}{N} = \underbrace{p(Y=1|\theta^*)}_{p^*(y)}$$

under the "true" model

Let $p(Y=y|\theta) \Rightarrow P_\theta(y)$

$$= p^*(1)\log P_\theta(1) + \dots$$

$$= \sum_{y=1}^{k} p^*(y)\log P_\theta(y) = \sum_y p^*(y)\log\left[P_\theta(y)\frac{p^*(y)}{p^*(y)}\right]$$

$$= \underbrace{\sum_y p^*(y)\log p^*(y)}_{-\text{ve entropy (doesn't depend on } \theta)} - \underbrace{\sum_y p^*(y)\log\left[\frac{p^*(y)}{P_\theta(y)}\right]}_{KL(p^* \| P_\theta)}$$

So, $\boxed{\arg\max_\theta \frac{1}{N} LL(\theta) = \arg\min KL(p^* \| P_\theta)}$

So when we do MLE, we are trying to get closer to the reality, if we have infinite data.

⑤

MLE is a frequentist estimation.

Our answer was 3/5 as there were 3 wins & 2 losses. What if # win = 30, # loss = 20? → We'll still have the same answer. MLE estimator can't convey/confident on how much data you have seen.

$\boxed{\text{MAP}}$ We can put a belif on anything (e.g. I know from my heart that India will win). That's Bayesian.

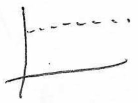
In frequentist, θ is constant, it's not random.

In Bayesian,

$$P(\theta|D) = \frac{P(D|\theta)\,P(\theta)}{P(D)}$$

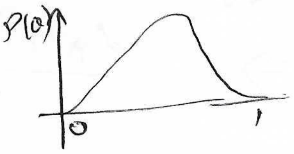
$$\hat{\theta}_{MAP} = \underset{\theta}{\arg\max}\; P(\theta|D)$$ (after you observe some data, what do you believe about this)

If $P(\theta)$ is constant, then MAP reduces to MLE:



$$\boxed{\hat{\theta}_{MAP} = \hat{\theta}_{MLE}}$$

Prior: For $P(\theta)$, we need some probability distribution or some density function between 0 & 1.



A useful prior:

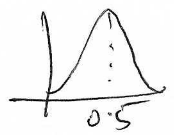
$$P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim \text{Beta}(\beta_H, \beta_T)$$

hyperparameters.

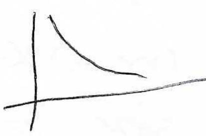
- $\beta_H = \beta_T = 1$

$\beta_H = 0.6$
$\beta_T = 0.6$

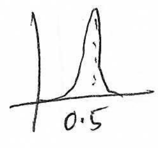
- $\beta_H = 2.5$
$\beta_T = 2.5$



$\beta_H = 0.6$
$\beta_T = 4.0$



- $\beta_H = 5.0$
$\beta_T = 5.0$



$\beta_H = 4.0$
$\beta_T = 0.6$

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta|D) \propto p(D|\theta)\, p(\theta)$$

$$\propto \theta^{\alpha_H}(1-\theta)^{\alpha_T} \frac{\theta^{\beta_H-1}(1-\theta)^{\beta_T-1}}{\text{constant}}$$

$$\propto \frac{\theta^{\alpha_H+\beta_H-1}(1-\theta)^{\alpha_T+\beta_T-1}}{\text{const.}}$$

$$\propto \text{Beta}(\alpha_H+\beta_H,\ \alpha_T+\beta_T)$$

$\boxed{\text{Conjugate prior}} \longrightarrow$

Now let's derive the MAP:

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta|D) = \theta^{\alpha_H}(1-\theta)^{\alpha_T} \frac{\theta^{\beta_H-1}(1-\theta)^{\beta_T-1}}{\text{constant}}$$

$$= \arg\max_{\theta} \log\left[\theta^{\alpha_H}(1-\theta)^{\alpha_T} \frac{\theta^{\beta_H-1}(1-\theta)^{\beta_T-1}}{\text{constant}}\right]$$

$$\frac{d}{d\theta}\log P(\theta|D) = \frac{d}{d\theta}\left[\alpha_H \log\theta + \alpha_T \log(1-\theta) + (\beta_H-1)\log\theta + (\beta_T-1)\log(1-\theta)\right] = 0$$

$$\Rightarrow \frac{d}{d\theta}\left[(\alpha_H+\beta_H-1)\log\theta + (\alpha_T+\beta_T-1)\log(1-\theta)\right] = 0$$

$$\Rightarrow \frac{\alpha_H+\beta_H-1}{\theta} - \frac{\alpha_T+\beta_T-1}{1-\theta} = 0$$

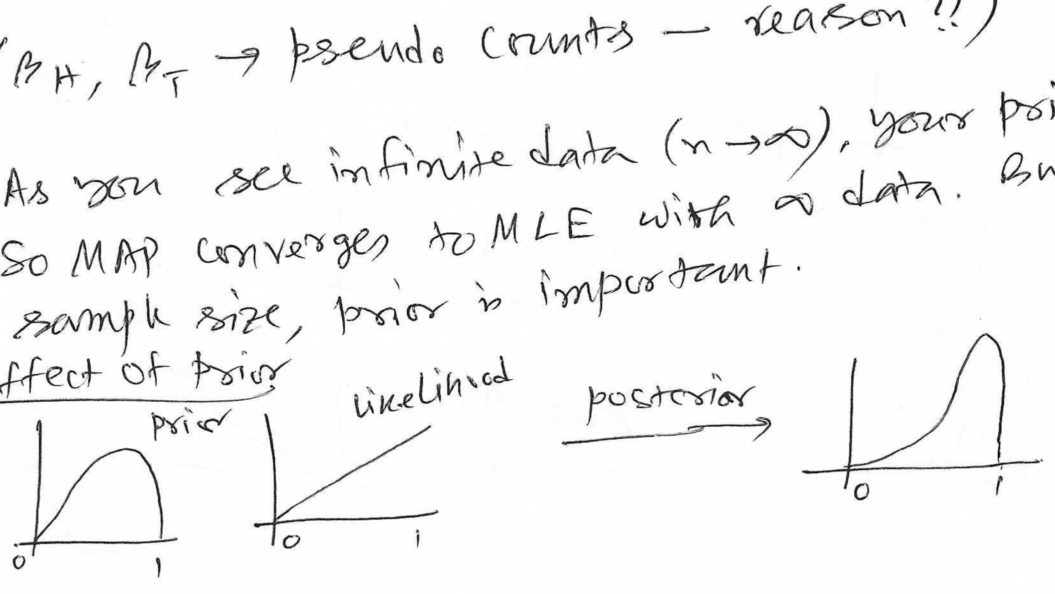
$$\Rightarrow \theta(\alpha_T+\beta_T+\alpha_H+\beta_H-2) = \alpha_H+\beta_H-1$$

$$\Rightarrow \hat{\theta}_{MAP} = \frac{\alpha_H+\beta_H-1}{\alpha_T+\beta_T+\alpha_H+\beta_H-2}$$

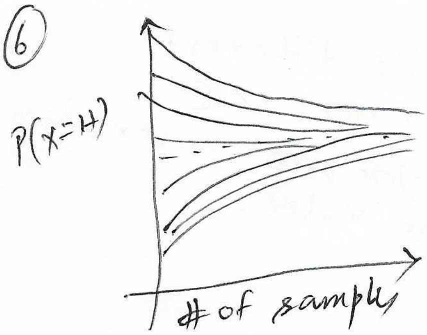
($\beta_H, \beta_T \rightarrow$ pseudo counts — reason?!)

As you see infinite data ($n \to \infty$), your prior is forgotten! So MAP converges to MLE with ∞ data. But for small sample size, prior is important.

Effect of Prior



6



You start with different priors, but as you see more data, you are converging to the same value.

$$\boxed{\begin{array}{l}\text{Frequentist tool} \Rightarrow \text{MLE} \\ \text{Bayesian tool} \Rightarrow \text{MAP}\end{array}}$$

MAP vs. full Bayesian

We are interested in $P(\theta|D)$

By Baye's rule: $P(\theta|D) = \dfrac{P(D|\theta)\,P(\theta)}{P(D)}$

Now $P(D) = \int P(D|\theta)\,P(\theta)$ ← marginal likelihood or evidence.
← This integral is the root of difficulty.

Consider every possible parameter setting θ, see how well it explains the dataset D, weight it by how plausible θ was a-priori, and add everything up.

So Bayes must account for all possible models, not just the best one.

Why MAP avoids the hard part?

$\theta_{MAP} = \arg\max_{\theta} P(D|\theta)\,P(\theta)$ → No integral, no normalization, no averaging. MAP doesn't care about $P(D)$ at all!!

In coin toss, θ is a number, $P(D|\theta)$ is simple, $P(\theta)$ is simple, & the integral has a closed form.

Now imagine complex ML model (Like Neural Network). So the integral becomes:

$$\int_{\mathbb{R}^{10^6}} P(D|\theta)\,P(\theta)\,d\theta$$ ← Integration over million dimensional space → no closed form sol^n.

MAP → point estimate

Full Bayesian → full posterior distribution.

MAP asks → what single parameter value is most plausible after seeing the data?

Full Bayesian inference → What is the entire distribution of plausible parameters?

Bayes does not ask for best θ; Bayes asks for the contribution of all θ. MAP ignores them.

Often Bayesian involves approximating the integral.

MLE/MAP example: Gaussian

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

Why Gaussian?

- Central limit theorem.
- They are easy
- Closely related to squarred loss
- Mixtures of Gaussians are sufficient to approximate many distributions

Some properties

- Affine transformation

$$X \sim \mathcal{N}(\mu,\sigma^2) \quad \Rightarrow \quad Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$$
$$Y = aX + b$$

- Sum of independent Gaussians

$$X \sim \mathcal{N}(\mu_X, \sigma_X^2)$$
$$Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2) \quad \Rightarrow \quad Z \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$
$$Z = X + Y$$

Learning a Gaussian

Collect bunch of data

↳ hopefully I.I.D. samples.

Learn parameters

- mean
- variance.

⑧ $\boxed{\text{MLE}}$ Data: $D = \{y_1, \dots, y_N\}$

Model assumption: $Y \sim \mathcal{N}(\mu, \sigma^2)$

Density fn: $p(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right]$

$$(\hat{\mu}_{MLE}, \hat{\sigma}_{MLE}) = \underset{\mu,\sigma}{\arg\max}\ LL(\mu, \sigma)$$

Now, $LL(\mu,\sigma) = \log p(D \mid \mu, \sigma) = \log \prod_{i=1}^{N} p(y_i \mid \mu, \sigma)$ [I.I.D.]

$$= \sum_{i=1}^{N} \log p(y_i \mid \mu, \sigma)$$

$$= \sum_{i=1}^{N} \log\left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i-\mu)^2}{2\sigma^2}\right)\right]$$

$$= \sum_{i=1}^{N} \left[-\frac{(y_i-\mu)^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right]$$

$$= \sum_{i=1}^{N} -\frac{(y_i-\mu)^2}{2\sigma^2} - \frac{N}{2}\log(2\pi\sigma^2)$$

$$\frac{\partial LL}{\partial \mu} = -\sum_{i=1}^{N} \frac{2(y_i-\mu)(-1)}{2\sigma^2} = 0 \Rightarrow \sum_{i=1}^{N}(y_i - \mu) = 0 \quad \left(\text{assume } \sigma \neq 0\right)$$

$$\Rightarrow \boxed{\hat{\mu}_{MLE} = \frac{1}{N}\sum y_i}$$ ← sample mean is the estimator that maximizes the likelihood.

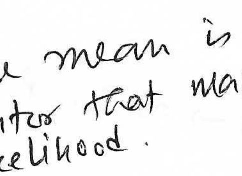
$$\frac{\partial LL}{\partial \sigma} = -\sum_{i=1}^{N} \frac{(y_i-\mu)^2}{2}\left(-\frac{2}{\sigma^3}\right) - \frac{N}{2}\frac{1}{2\pi\sigma^2} \cdot 2\pi(2\sigma) = 0$$

$$= \frac{\sum (y_i-\mu)^2}{\sigma^3} - \frac{N}{\sigma} = 0$$

$$\Rightarrow \boxed{\hat{\sigma}^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{\mu}_{MLE})^2}$$

$\boxed{\text{MAP}}$ Assume σ = (unknown) constant.

Let's put a prior distribution on μ:



$$\mu \sim \mathcal{N}(\eta, \lambda^2)$$

hyper-parameters.

$$p(\mu \mid \eta, \lambda) = \frac{1}{\sqrt{2\pi\lambda^2}} \exp\left[-\frac{(\mu-\eta)^2}{2\lambda^2}\right]$$

$$\hat{\mu}_{MAP} = \arg\max_{\mu} \log p(\mu|D)$$

$$= \arg\max_{\mu} \log \frac{p(D|\mu)\, p(\mu|\eta,\lambda^2)}{p(D)}$$

$$= \arg\max_{\mu} \underbrace{\log p(D|\mu)}_{\text{log likelihood}} + \underbrace{\log p(\mu|\eta,\lambda^2)}_{\text{log prior}} - \underbrace{\log p(D)}_{\text{constant w.r.t. } \mu}$$

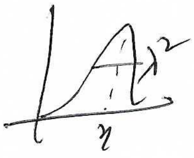
$$\frac{\partial}{\partial \mu}\left(\log p(\mu|D)\right) = \frac{\partial LL}{\partial \mu} + \frac{\partial}{\partial \mu} \log p(\mu|\eta,\lambda^2)$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{N}(y_i-\mu) + \frac{\partial}{\partial \mu}\left[-\frac{(\mu-\eta)^2}{2\lambda^2} - \frac{1}{2}\log(2\pi\sigma^2)\right]$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{N}(y_i-\mu) - \frac{\mu-\eta}{\lambda^2} = 0$$

$$\Rightarrow \quad \hat{\mu}_{MAP} = \frac{\sum \frac{y_i}{\sigma^2} + \frac{\eta}{\lambda^2}}{\frac{N}{\sigma^2} + \frac{1}{\lambda^2}}$$

$$p(\mu|\eta,\lambda^2) = \frac{1}{\sqrt{2\pi\lambda^2}} \exp\left(-\frac{(\mu-\eta)^2}{2\lambda^2}\right)$$



As $\lambda \to \infty$ in $p(\mu|\eta,\lambda^2)$

nearly uniform distribution

So, $$\hat{\mu}_{MAP} = \frac{\sum \frac{y_i}{\sigma^2} + (\to 0)}{\frac{N}{\sigma^2} + (\to 0)} \Rightarrow \hat{\mu}_{MLE}$$

As $\lambda \to 0$, $p(\mu|\eta,\lambda^2) \to$  (very strong prior) (delta fn)

$$\hat{\mu}_{MAP} = \frac{\lambda^2 \sum \frac{y_i}{N} + \eta}{\lambda^2 \frac{N}{\sigma^2} + 1}$$

$$\Rightarrow \hat{\mu}_{MAP} \to \eta \quad \text{(prior mode)}$$

Same can be obtained from variance as well.