Probabilistic Inference

Bernoulli distribution

→ la distribución de Bernoulli es una distribución aplicada a una variable aleatoria discreta, la cual solo puede resultar en 2 sucesos posibles: "éxito" y "no éxito" (T, H)

$P_i(F_i = T) = \theta_i$

$P_i(F_i = T \mid \theta_i) = Ber(F_i = T \mid \theta_i) = \theta_i \qquad F_i \sim Ber(\theta_i)$

(θ_i: probabilidad)

Maximum Likelihood

1. The flips are qualitatively the same - identical distribution

$$P_i(F_i = f_i \mid \theta_i) = P_i(F_i = f_i \mid \theta)$$

2. The coins flips do not affect each other - independence

$$P(H, T, H, H \mid \theta) = P(F_1 = H \mid \theta) \cdot P(F_2 = T \mid \theta) \cdot P(F_3 = H \mid \theta) \cdot P(F_4 = H \mid \theta)$$

$$= \prod_{i=1}^{4} P(F_i = f_i \mid \theta)$$

$$f(\theta) := p(D \mid \theta)$$

$$\theta_{MLE} = \arg\max_{\theta \in [0,1]} f(\theta)$$

↑ is not a probability distribution

Monotonic functions preserve critical points

$$\arg\max_{\theta \in [0,1]} f(\theta) = \arg\max_{\theta \in [0,1]} \log f(\theta)$$

$f(x) = \log(x)$ ex.: $f(x) = 10 \log(x)$

$f'(x) = \frac{x'}{x} = \frac{1}{x}$ $\qquad f'(x) = \frac{10}{x}$

$$f(x) = \theta^t (1-\theta)^t = \log \theta^t (1-\theta)^h = \log \theta^t + \log (1-\theta)^h = t \log \theta + h \log(1-\theta)$$

$$f'(x) = \frac{t(1)}{\theta} + \frac{h(-1)}{1-\theta} = \frac{t}{\theta} - \frac{h}{1-\theta}$$

to find the max → $\frac{t}{\theta} - \frac{h}{1-\theta} = 0$; $\frac{t}{\theta} = \frac{h}{1-\theta}$; $(1-\theta)t = \theta h$

$(1-\theta)t = \theta h$; $h = \frac{(1-\theta)t}{\theta}$; $h = \frac{t - \theta t}{\theta}$; $t = \theta h + \theta t$; $t = \theta(h+t)$;

$$\theta = \frac{t}{h+t}$$

MLE is based on the results of the experiment, but if we have just a few instances, it may seem counter-intuitive.

We have prior beliefs that also counts

Bayesian inference

A prior distribution $p(\theta)$ represents our beliefs before we observe any data

1. It must not depend on the data
2. $p(\theta) \geq 0$ for all θ
3. $\int p(\theta)\, d\theta = 1$

} feasible values of θ, only values $\theta \in [0,1]$ makes sense

Bayes formula: $p(\theta|D) = \frac{p(D|\theta) \cdot p(\theta)}{p(D)}$ (likelihood · prior)

$p(D)$ → evidence → it acts as a normalizing constant that ensures that the posterior distribution integrates to 1.

$p(\theta|D)$ → posterior distribution → it encodes our beliefs in the value of θ after observing the data

posterior $\propto$ likelihood · prior $\Rightarrow p(\theta|D) \propto p(D|\theta) \cdot p(\theta)$

$$p(D) = \int p(D, \theta)\, d\theta = \int p(D|\theta)\, p(\theta)\, d\theta$$

if $p(\theta) = 0$ for some particular θ, posterior will always be zero for that particular θ regardless of the likelihood / data.

Maximum a posteriori estimation (MAP)

MLE ignores our prior beliefs and performs poorly if little data is available

$$\theta_{MAP} = \arg\max_{\theta} p(\theta \mid D) = \arg\max_{\theta} \frac{p(D \mid \theta)\, p(\theta)}{p(D)} = \arg\max_{\theta} p(D \mid \theta)\, p(\theta)$$

we can ignore $\frac{1}{p(D)}$ since it's a positive constant independent of θ

we choose Beta distribution for prior

$$\text{Beta}(\theta \mid a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}, \quad \theta \in [0,1]$$

where

- $a > 0, b > 0$ are the distribution parameters
- $\Gamma(n) = (n-1)!$ for $n \in \mathbb{N}$ is the gamma function

Putting everything together (likelihood, prior) because $p(D)$ is constant

$$p(\theta \mid D) = \frac{p(D \mid \theta) \cdot p(\theta)}{p(D)} \propto p(D \mid \theta) \cdot p(\theta)$$

$p(\theta \mid D)$: posterior

We know

$$p(D \mid \theta) = \theta^{|T|} (1-\theta)^{|H|},$$

$$p(\theta) \equiv p(\theta \mid a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1}$$

So:

$$p(\theta \mid D) \propto \theta^{|T|} (1-\theta)^{|H|} \cdot \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1}$$

$$\propto \theta^{|T|+a-1} (1-\theta)^{|H|+b-1}$$

We are looking for $\theta_{MAP} = \arg\max_{\theta} p(\theta \mid D) = \arg\max_{\theta} \theta^{|T|+a-1} (1-\theta)^{(H)+b-1} =$

$$\arg\max_{\theta} \log p(\theta \mid D) = \arg\max_{\theta} (|T|+a-1) \log\theta + (|H|+b-1) \log(1-\theta)$$

We obtain $\theta_{MAP} = \frac{|T|+a-1}{|H|+|T|+a+b-2}$

Estimating the posterior distribution

We need to consider the entire posterior distribution $p(\theta|D)$, not just its mode θ_{MAP}

$$p(\theta|D) \propto \theta^{|T|+a-1}(1-\theta)^{|H|+b-1}$$

Finding the true posterior $p(\theta|D)$ boils down to finding the normalization constant, such that the distribution integrates to 1.

Sol.: → Pattern matching

Unnormalized posterior → $p(\theta|D) \propto \theta^{|T|+a-1}(1-\theta)^{|H|+b-1}$

Beta distribution → $\text{Beta}(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$

Thus, we can conclude that the appropriate normalizing constant is → $\frac{\Gamma(|T|+a+|H|+b)}{\Gamma(|T|+a)\Gamma(|H|+b)}$

and the posterior is a Beta distribution $p(\theta|D) = \text{Beta}(\theta|a+|T|, b+|H|)$

The prior is also a beta distribution $p(\theta) = \text{Beta}(\theta|a,b)$

This is not a coincidence, this is an istance of a more general principle.

Beta distribution is a conjugate prior for the Bernoulli likelihood.

If a prior is conjugate for the given likelihood, then the posterior will be of the same family as the prior.

Normal distribution / Gaussian

$$\mathcal{N}(\mu|\underbrace{m}_{\text{mean}},\underbrace{\beta^{-1}}_{\text{precision}}) = \sqrt{\frac{\beta}{2\pi}}\exp\left(-\frac{\beta}{2}(\mu-m)^2\right) \propto \exp\left(-\frac{\beta}{2}\mu^2 + \beta m\mu\right)$$

mode of $p(\theta|D) = \text{Beta}(\alpha, \beta) = \frac{\alpha-1}{\alpha+\beta-2}$

Relation between MAP, MLE and posterior distribution

MAP is the mode of the posterior distribution

And if we choose a uniform prior ($a=b=1$) we obtain MLE

All this is due to we have chosen a conjugate prior. Had we chosen a non-conjugate prior, $p(\theta|D)$ and θ_{MAP} could not have a closed form

Predicting the next flip

We want to compute the probability that the next coin flip is (T), given observations D and prior belief a, b:

$$p(F = \text{T} \mid D, a, b)$$

This distribution is called the posterior predictive distribution

$$p(F = f \mid D, a, b) = p(f \mid D, a, b) = \int_0^1 p(f|\theta)\, p(\theta|D, a, b)\, d\theta \qquad f \in \{0, 1\}$$

Recall

$$p(f|\theta) = \text{Ber}(f|\theta) = \theta^f (1-\theta)^{1-f}$$

$$p(\theta|D, a, b) = \frac{\Gamma(|T|+a+|H|+b)}{\Gamma(|T|+a)\,\Gamma(|H|+b)}\, \theta^{|T|+a-1}\, (1-\theta)^{|H|+b-1}$$

$$= \frac{(|T|+a)^f\,(|H|+b)^{(1-f)}}{|T|+a+|H|+b} = \text{Ber}\left(f \;\middle|\; \frac{|T|+a}{|T|+a+|H|+b}\right)$$

↓ it does not contain θ

↓ fully Bayesian analysis