Likelihood: observation: patient has yellow eyes

Explanation A: Patient has hepatitis B (HB) 乙肝

Explanation B: Patient has COVID

Explanation A is more reasonable because

$P(\text{yellow eyes} \mid HB) >> P(\text{yellow eyes} \mid \text{COVID})$

observation E, Explanation H

$P(E|H)$ is a measure of how well H explains E.

Fixed change

$L(H|E) = P(E|H)$ likelihood of H

function of H

Data : D (observations)

Model : m (explanation)

$$L(m|D) = P(D|m)$$

likelihood of model given Data

measures how well model explains data

fits

$p(m)$ — prior

$p(m|D)$ — posterior

$L(m|D) = p(D|m)$ — likelihood

$$p(m|D) = \frac{p(m, D)}{p(D)} = \frac{p(D|m)\, p(m)}{p(D)}$$

$p(D)$ → Evidence

partition function

$$\propto p(m)\, p(D|m)$$

posterior $\propto$ prior $\times$ likelihood

Bayes' Theorem

parameter estimation

$\theta = P(X = H)$

↑ result of thumb tack toss

$D = \{H, T, H, H, T\}$

θ:

A: D

B: 0.5

C: 0.6 ✓ — Explains / fits the data the best

D: 0.7

maximizes $L(\theta|D) = P(D|\theta)$

Maximum likelihood Estimation (MLE)

$D = \{H, T, H, H, T\}$

d_1, d_2, d_3, d_4, d_5

$\theta = P(X = H)$

$L(\theta | D) = P(D | \theta)$

$= P(d_1, d_2, \ldots, d_5 | \theta)$

independence assumption

$= P(d_1|\theta)\, P(d_2|\theta)\, P(d_3|\theta)\, P(d_4|\theta)\, P(d_5|\theta)$

$= \theta \quad (1-\theta) \quad \theta \quad \theta \quad (1-\theta)$

$= \theta^3 (1-\theta)^2$

m_h m_t

identically distributed (i.i.D)

$D' = \{H, H, H, T, T\}$

$L(\theta | D') = \theta^3 (1-\theta)^2$

sufficient statistics

$$\max_{\theta} \quad L(\theta \mid D) = \theta^3 (1-\theta)^2$$

$$\ell(\theta \mid D) = \log L(\theta \mid D)$$

$$\uparrow \qquad = 3 \log \theta + 2 \log (1-\theta)$$

log likelihood : Math simplicity

$$\frac{d\ell}{d\theta} = 0 \quad \Rightarrow \quad \theta = \frac{3}{3+2} = 0.6$$

$$\text{MLE:} \quad \theta = \frac{m_h}{m_h + m_t}$$

Bayesian Estimation

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

$$\propto \theta^{\alpha_h - 1} (1-\theta)^{\alpha_t - 1} \theta^{m_h} (1-\theta)^{m_t}$$

$$= \theta^{\alpha_h + m_h - 1} (1-\theta)^{\alpha_t + m_t - 1}$$

$$\Rightarrow P(\alpha_{m+1} = H | D) = \frac{\alpha_h + m_h}{\alpha_h + m_h + \alpha_t + m_t}$$

α_h, α_t : Virtual Counts

$\alpha_h = \alpha_t = 100$

Beta Distribution

200 : equivalent sample

$$P(d_{m+1} = H \mid D) = \frac{\alpha_h + m_h}{\alpha_h + m_h + \alpha_t + m_t}$$

Thumbtack: $\alpha_h = \alpha_t = 0$

$m_h = 3, \quad m_t = 7$: $\quad P(d_{m+1} = H \mid D) = \frac{3}{10} = 0.3$

Coin: $\quad \alpha_h = \alpha_t = 1{,}000$

① $m_h = 3, \quad m_t = 7 \quad P(d_{m+1} = H \mid D) = \frac{1{,}000 + 3}{1{,}000 + 3 + 1{,}000 + 7} \approx 0.5$

② $m_h = 30{,}000, \quad m_t = 70{,}000 \quad P(d_{m+1} = H \mid D) = \frac{1{,}000 + 30{,}000}{1{,}000 + 30{,}000 + 1{,}000 + 70{,}000} \approx 0.3$

f is concave

$x_1, x_2, \cdots, x_n$

$p_1, p_2, \cdots, p_n$

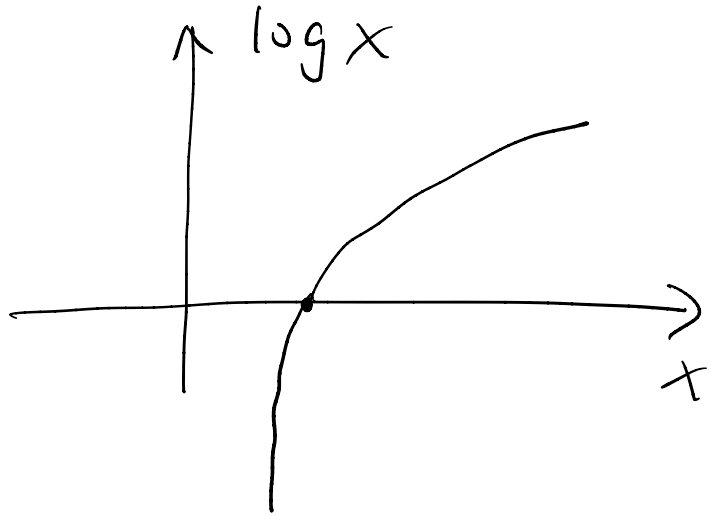
$p_i \geq 0, \quad \sum_i p_i = 1$

$$\sum_{i=1}^{n} p_i f(x_i) \leq f\left(\sum_{i=1}^{n} p_i x_i\right)$$

Jensen's inequality

$$E_p[f(x)] \leq f(E_p[x])$$

$\log x$ is concave



$$E_p[\log x] \leq \log E_p[x]$$

Example 2: Thumb tack $\theta = P(X = h)$

Guess X : result of Thumb tack toss

① $\theta = 0.2$ ② $\theta = 0.5$ ③ $\theta = 0.8$

$$H(X) = P(X=h) \log \frac{1}{P(X=h)} + P(X=t) \log \frac{1}{P(X=t)}$$

$$= \theta \log \frac{1}{\theta} + (1-\theta) \log \frac{1}{1-\theta}$$

$$= -\theta \log \theta - (1-\theta) \log (1-\theta)$$

$$H(x) = \sum_x p(x) \log \frac{1}{p(x)}$$

$x \in \{1, 2, \ldots 6\}$

① $H(x) \geq 0$

② $H(x) \leq \log |x|$

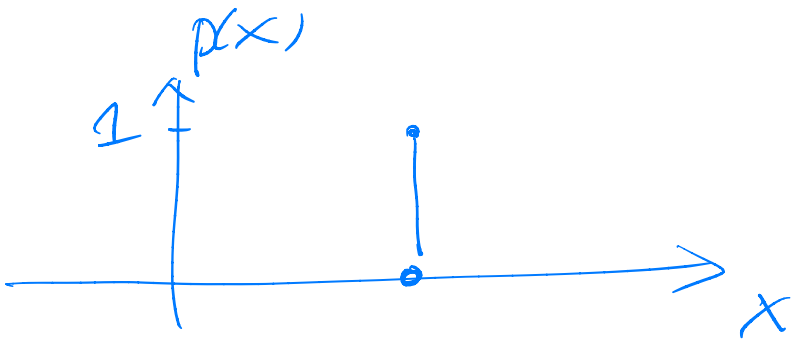
$$\leq \log \sum_x p(x) \frac{1}{p(x)}$$

$$= \log |x|$$

↖ cardinality of x

$H(x) = \log |x|$ when $p(x)$ is uniform

③ $H(x) = 0$ when $p(x=a) = 1$ for some a



Conditional Entropy

$H(Y) = \sum_y p(y) \log \frac{1}{p(y)}$

Y : initial dist $p(Y)$, $H(Y)$

Evidence $X=a$: $p(Y|X=a)$

$$H(Y|X=a) = \sum_y p(y|X=a) \log \frac{1}{p(y|X=a)}$$

How uncertainty about Y will change if we test on X?

$$H(Y|X) = \sum_a H(Y|X=a)\, p(X=a)$$

$$= \sum_a \sum_y p(y|X=a) \log \frac{1}{p(y|X=a)}\, p(X=a)$$

$$= \sum_{x,y} p(x,y) \log \frac{1}{p(y|x)} = -E[\log p(y|x)]$$

KL Divergence

$$KL(p \| q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

$$= -\sum_x p(x) \log \frac{q(x)}{p(x)}$$

(1) $KL(p \| q) \geq 0$

(2) $KL(p \| q) = 0$ when $p = q$

$$\geq -\log \sum_x p(x) \frac{q(x)}{p(x)}$$

(3) $KL(p \| q) \neq KL(q \| p)$

$$= -\log \sum_x q(x)$$

$$= -\log 1$$

$$= 0$$

$$0 \leq KL(p \| Q) = \sum_x p(x) \log \frac{p(x)}{Q(x)}$$

$$= \sum_x p(x) \log p(x) + \sum_x p(x) \log \frac{1}{Q(x)}$$

$$= -\sum_x p(x) \log \frac{1}{p(x)} + \sum_x p(x) \log \frac{1}{Q(x)}$$

cross entropy

$$= -H(p) + H(p, Q)$$

$$\min KL(p \| Q) \equiv \min H(p, Q)$$

$$H(p, Q) \geq H(p)$$

$$-E_p[\log Q(x)] \geq -E_p[\log p(x)]$$

$$E_p[\log Q(x)] \leq E_p[\log p(x)]$$

Gibbs Inequality

Mutual Information

$I(X;Y) = H(X) - H(X|Y)$

$H(X)$: Uncertainty about X

$H(X|Y)$: Expected uncertainty about X if test Y

Info Y has about X

0.9 0.2

$= KL(p(x,y) \,||\, p(x)p(y))$

≥ 0

$I(Y;X)$

$= 0$ if $p(x,y) = p(x)p(y)$

① $I(X;Y) = I(Y;X)$

② $I(X;Y) \geq 0$

③ $I(X;Y) = 0$ when X & Y independent

④ $H(X) \geq H(X|Y)$

Gibbs Inequality

$$E_p[\log Q(x)] \leq E_p[\log P(x)]$$

Thumbtack: Data m_h heads, m_t tails

$$\ell(\theta \mid D) = m_h \log \theta + m_t \log(1-\theta)$$

$$\frac{d\ell}{d\theta} = 0 \implies \theta^* = \frac{m_h}{m_h + m_t} \qquad m = m_h + m_t$$

$$\max_Q \; m_h \log Q(x=h) + m_t \log Q(x=t) \qquad \theta \quad Q(x=h)$$

$$\Leftrightarrow \max_Q \; \frac{m_h}{m} \log Q(x=h) + \frac{m_t}{m} \log Q(x=t)$$

$\frac{m_h}{m} = \tilde{P}(x=h)$ $\qquad$ $\frac{m_t}{m} = \tilde{P}(x=t)$ $\qquad$ empirical distribution

$$\Leftrightarrow Q = \tilde{P}, \quad Q(x=h) = \frac{m_h}{m}$$