09/07/2022 Wed

Entropy of coin shows heads

$$H(x) = -\sum_{k=1}^{K} p \log_2 p$$

- $P(x) = 0.5$

$$H(x) = -0.5 \underbrace{\log_2 0.5}_{\log_2 \frac{1}{2} \; = \; -1} - 0.5 \log_2 0.5$$

$$= -0.5 \cdot (-1) - 0.5 \cdot (-1)$$

$$= 1$$

- $P(x) = 1$ getting heads

$$H(x) = -1 \log_2 1 - 0 \log_2 0$$

$$= 0$$

KL divergence

$$KL(p \| q) \triangleq \sum_{k=1}^{K} p \log \frac{p}{q}$$

$\log \frac{p}{q} = \log p - \log q$

$$= \sum_k p \log p - \sum_k p \log q$$

$$= \underbrace{-H(p)}_{\text{entropy}} + \underbrace{H(p, q)}_{\text{cross-entropy}}$$

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

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

$X \sim Ber(\theta)$, $0 \le \theta \le 1$

Likelihood of D $\quad P(D|\theta) = \prod_{n=1}^{N} P(x_n|\theta)$

$$P(D|\theta) = \theta \cdot (1-\theta) \cdot \theta \cdot \theta$$
$$= \theta^3 \cdot (1-\theta)$$

$D = \{N_1 \text{ positive samples} ; N_2 \text{ negative samples}\}$

$$P(D|\theta) = \theta^{N_1}(1-\theta)^{N_2}$$

MLE : maximum Likelihood Estimate

$$\hat{\theta}_{MLE} = \arg\max_{\theta} P(D|\theta)$$
$$= \arg\max_{\theta} \theta^{N_1}(1-\theta)^{N_2}$$

$$\frac{\partial P(D|\theta)}{\partial \theta} = 0$$

$$\frac{\partial \theta^{N_1}(1-\theta)^{N_2}}{\partial \theta}$$

$$= N_1 \theta^{N_1-1}(1-\theta)^{N_2} - N_2\theta^{N_1}(1-\theta)^{N_2-1} = 0$$

$$N_1(1-\theta) = N_2\theta$$

$$N_1 - N_1\theta = N_2\theta$$

$$\hat{\theta}_{MLE} = \frac{N_1}{N_1+N_2}$$

Posterior $\propto$ likelihood $\times$ prior

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

$$\propto \theta^{N_1}(1-\theta)^{N_2}\theta^{a-1}(1-\theta)^{b-1}$$

$$\propto \theta^{N_1+a-1}(1-\theta)^{N_2+b-1}$$

$$\mathrm{Beta}(\theta|a,b) \propto \theta^{a-1}(1-\theta)^{b-1}$$

MAP : maximum a posterior estimate

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

$$= \frac{N_1+a-1}{N_1+N_2+a+b-2}$$

$$E(\theta|D) = \frac{N_1+a}{N_1+N_2+a+b}$$

Bayesian Averaging

$$p(x^*=1|D) = \int_0^1 p(x^*=1|\theta)\,p(\theta|D)\,d\theta$$

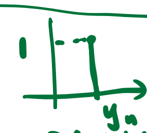
$$= \int_0^1 \theta \cdot \mathrm{Beta}(\theta|N_1+a, N_2+b)\,d\theta$$

$$= \frac{N_1+a}{N_1+N_2+a+b}$$

$$KL(p\|q) = \sum_y p(y) \log \frac{p(y)}{q(y)} = \underbrace{\sum_y p \log p}_{\text{entropy } P} - \underbrace{\sum_y p \log q}_{\text{cross-entropy}}$$

$p(y) = p_D(y)$ empirical distribution $\triangleq \frac{1}{N} \sum_{n=1}^{N} \delta(y - y_n)$

$q(y) = p(y|\theta)$ learned model



$$KL(p\|q) = \sum_y p_D(y) \log p_D(y) - \sum_y p_D(y) \log p(y|\theta)$$

$$= -H(p_D) - \frac{1}{N} \sum_{n=1}^{N} \log(y_n|\theta)$$

$$= \text{const} + NLL(\theta)$$

$$NLL(\theta) = -\log p(D|\theta)$$

$$= -\log \prod_y (y|\theta)$$

$$= -\sum_y \log(y|\theta)$$

$$\hat{h}_{MAP} = \operatorname*{argmax}_h \underbrace{p(D|h)}_{\text{likelihood}} \underbrace{p(h)}_{\text{prior}}$$

$$= \operatorname*{argmax}_h \left(\log p(D|h) + \log p(h)\right)$$