Information Theory

Shanon Entropy

Information $\propto \frac{1}{Possibility}$

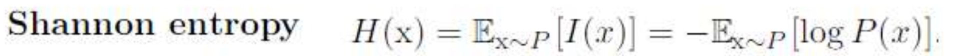
$$\text{Information} \propto \frac{1}{P(x)}$$

Adaditivity: For independent events $Z_1$and Z_2

$$\textbf{Inf}(Z_1 + Z_2) \propto \textbf{Inf}(Z_1) + \textbf{Inf}(Z_2)$$

$$I(X) = -\log P(x)$$

Covers two mwntioed properties

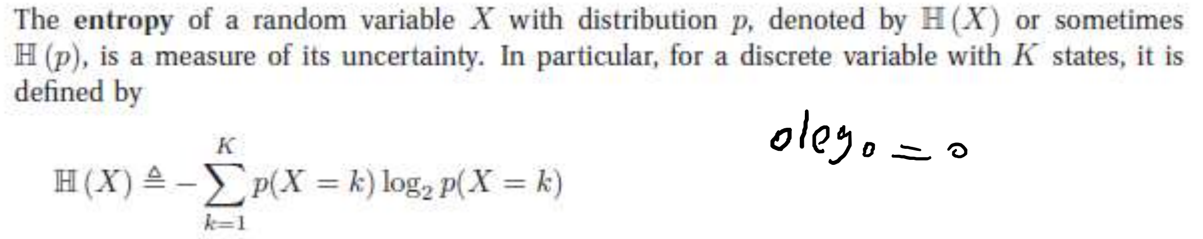


Shannon entropy $H(\mathrm{x}) = \mathbb{E}_{\mathrm{x}\sim P}[I(x)] = -\mathbb{E}_{\mathrm{x}\sim P}[\log P(x)].$

$\mathrm{inf}(X) \propto \frac{1}{p(x)}$, $\mathrm{inf}(x,y) = \mathrm{inf}(x) + \mathrm{inf}(y) \leftarrow \mathrm{inf}(x) = \log\left(\frac{1}{p(x)}\right) = -\log p(x)$

x, y indep

$X \rightarrow p(x)$ $-\log p(x)$



The **entropy** of a random variable X with distribution p, denoted by $\mathbb{H}(X)$ or sometimes $\mathbb{H}(p)$, is a measure of its uncertainty. In particular, for a discrete variable with K states, it is defined by

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

In order to co

$0\log 0 = 0$

ssage of length 3 bits.

$E(f(x)) = \sum f(x)p(x)$

$X = \{0, 1\}$

$$\mathrm{H}[x] = -8 \times \frac{1}{8}\log_2\frac{1}{8} = 3 \text{ bits.}$$

$H(x) = E[-\log p(x)] = -\sum p(x)\log p(x)$ Entropy

Event	$\{a, b, c, d, e, f, g, h\}$
Probability	$\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}$

$$\mathrm{H}[x] = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{16}\log_2\frac{1}{16} - \frac{4}{64}\log_2\frac{1}{64}$$

Coding

0, 10, 110, 1110, 111100, 111101, 111110, 111111

$$\text{average code length} = \frac{1}{2}\times 1 + \frac{1}{4}\times 2 + \frac{1}{8}\times 3 + \frac{1}{16}\times 4 + 4\times\frac{1}{64}$$

$P(x)$
$q(x)$

$d_1(P(x), q(x))$

$\underline{-\int p(x) \ln p(x)\, dx}$

$\varepsilon, 7$

Kullback-Leibler (KL)divergence

$\underline{-\int p(x) \ln q(x)\, dx}$

$q(x)$

Coding by inexact dist

$$\mathrm{KL}(p\|q) = \underline{-\int p(\mathrm{x}) \ln q(\mathrm{x})\, \mathrm{dx}} - \left(-\int p(\mathrm{x}) \ln p(\mathrm{x})\, \mathrm{dx}\right)$$

$$= -\int p(\mathrm{x}) \ln \left\{\frac{q(\mathrm{x})}{p(\mathrm{x})}\right\} \mathrm{dx}.$$

Coding by exact dist

Extra information for decoding

$$\mathrm{KL}(p\|q) \not\equiv \mathrm{KL}(q\|p).$$

$$\mathrm{KL}(p\|q) \geqslant 0$$

p(x): True unknown distribution

q(x): An approximation of p(x)

$\boxed{q(x)} \rightarrow \ln q(x)$

$P(x)$

P x_1 x_2 x_3 x_4

$\frac{1}{4}$ $\frac{1}{4}$ $\frac{1}{4}$ $\frac{1}{4}$

Q x_1 x_2 x_3 x_4

$\frac{1}{4}$ $\frac{3}{8}$ $\frac{1}{8}$, $\frac{1}{4}$

x_1 x_2 x_3 x_4

$\frac{1}{4}$ $\frac{1}{2}$ 0 $\frac{1}{4}$

$E(f(x)) = \underline{\sum} f(x) p(x)$

$-\log p(x)$

Kullback-Leibler (KL) divergence:The similarity of two distributions

$$D_{\mathrm{KL}}(P\|Q) = \mathbb{E}_{\mathrm{x}\sim P}\left[\log \frac{P(x)}{Q(x)}\right] = \underbrace{\mathbb{E}_{\mathrm{x}\sim P}[\log P(x)}_{-H(P)} - \log Q(x)]$$

$$D_{\mathrm{KL}}(P\|Q) \neq D_{\mathrm{KL}}(Q\|P)$$

cross-entropy

argmin Q

$$H(P,Q) = H(P) + D_{\mathrm{KL}}(P\|Q) = -\mathbb{E}_{\mathrm{x}\sim P}\log Q(x)$$

$$E_{x\sim P}\left(\underbrace{-\log p(x)}_{f(x)}\right)$$

$$argmin_Q H(P,Q) = argmin D_{KL}(P||Q)$$

$$\boxed{-\int p(x)\log p(x)} + \left(-\int p(x)\log q(x)\right)$$

17 17 17 17 18

$\frac{17+17+18+19}{4} = \frac{2\times17+18+19}{4} = 17\times\frac{1}{2} + 18\times\frac{1}{4} + 19\times\frac{1}{4}$

ML and Cross Entropy Equivalence

$x^{(i)}$

17 17 18 19

$\frac{1}{2}$ $\frac{1}{4}$ $\frac{1}{4}$

$$\theta_{\mathrm{ML}} = \arg\max_{\theta} \sum_{i=1}^{m} \log p_{\mathrm{model}}(\boldsymbol{x}^{(i)};\boldsymbol{\theta}). \Rightarrow \theta_{\mathrm{ML}} = \arg\max_{\theta} \mathbb{E}_{x\sim\hat{p}_{dat}} \log p_{\mathrm{model}}(\boldsymbol{x};\boldsymbol{\theta}) = argmin \; -\mathbb{E}_{\sim\hat{p}_{dat}} [\log p_{\mathrm{model}}(\boldsymbol{x})]$$

$x \sim P_{dat}$

$$= argmin \; H(P_{dat}, Q) \quad P_{model}$$

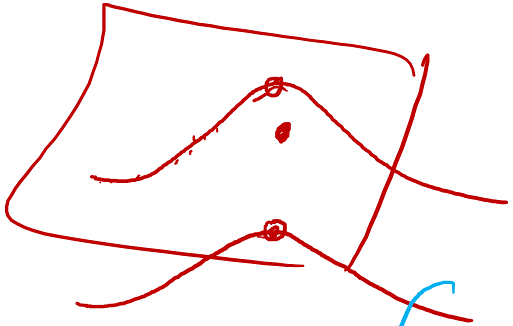
$$D_{\mathrm{KL}}(\hat{p}_{\mathrm{data}} \| p_{\mathrm{model}}) = \mathbb{E}_{\sim\hat{p}} [\log \hat{p}_{\mathrm{data}}(\boldsymbol{x}) - \log p_{\mathrm{model}}(\boldsymbol{x})]$$

$H(P_{dat}, P_{model})$

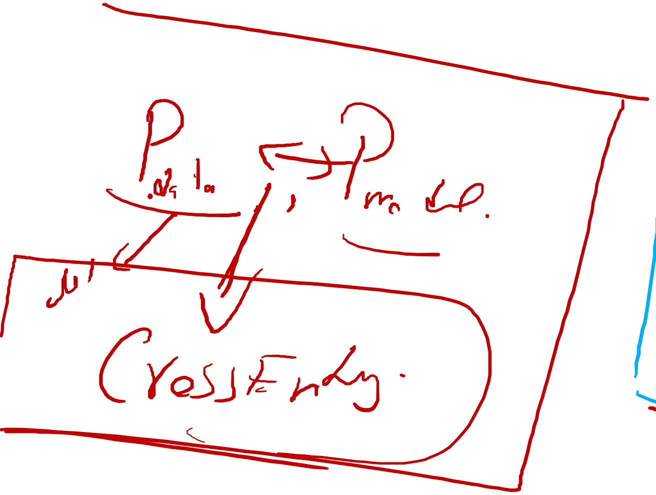
One way to interpret maximum likelihood estimation is to view it as minimizing the dissimilarity between the empirical distribution $\hat{p}_{\mathrm{data}}$ defined by the training set and the model distribution, with the degree of dissimilarity between the two measured by the KL divergence.

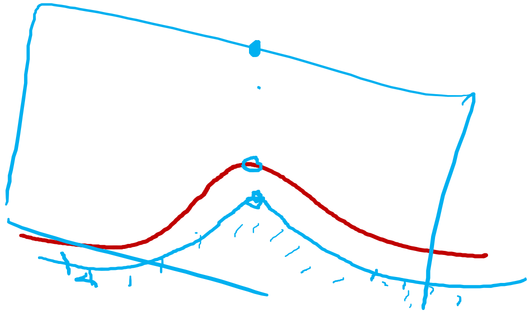
$$* \; E_{x\sim P_{dat}}[-\log Q(x)] = H(P,Q)$$

P, q   P, Q



P, Q





$$-\int P(x)\log P(x)$$

$$\int P(x)\log q(x)$$

(b)

Mutual Information

$$KL\left(P(x,y) \,\|\, P(x)P(y)\right)$$

$$\underline{P(x,y)}$$

$$\underline{Q(x,y)} = P(x)P(y)$$

$$\begin{aligned} I[x,y] &\equiv KL(p(x,y)\|p(x)p(y)) \\ &= -\iint p(x,y)\ln\left(\frac{p(x)p(y)}{p(x,y)}\right) dx\,dy \end{aligned}$$

$$-\iint P(x,y)\ln P(x,y) + \iint \overset{P(x,y)}{\boxed{\cancel{P(x)P(y)}}}\ln p(x)p(y)$$

$$I[x,y] = H[x] - H[x|y] = H[y] - H[y|x]$$

$$= \iint p(x,y)\ln\frac{P(x)P(y)}{P(x,y)}\,dx\,dy$$

Conditional log-Likelihood

If $\boldsymbol{X}$ represents all our inputs and $\boldsymbol{Y}$ all our observed targets, then the conditional maximum likelihood estimator is

$$\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} P(\boldsymbol{Y} \mid \boldsymbol{X}; \boldsymbol{\theta}).$$

If the examples are assumed to be i.i.d., then this can be decomposed into

$$\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{m} \log P(\boldsymbol{y}^{(i)} \mid \boldsymbol{x}^{(i)}; \boldsymbol{\theta}).$$