Information Theory

Machine learning, 2021
Mansoor Rezghi
Department of Computer science, TMU

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

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

Adaditivity: For independent events $Z_1$and Z_2

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

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

Covers two mwntioed properties

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

Entropy

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 communicate the value of x to a receiver, we would need to transmit a message of length 3 bits.

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

Event $\{a, b, c, d, e, f, g, h\}$

Probability $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}$

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

Coding

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

$$\text{average code length} = \frac{1}{2}\times 1 + \frac{1}{4}\times 2 + \frac{1}{8}\times 3 + \frac{1}{16}\times 4 + 4\times\frac{1}{64}\times 6 = 2 \text{ bits}$$

Kullback-Leibler (KL)divergence

Coding by inexact dist

$$\begin{aligned}\mathrm{KL}(p\|q) &= -\int p(\mathbf{x})\ln q(\mathbf{x})\,\mathrm{d}\mathbf{x} - \left(-\int p(\mathbf{x})\ln p(\mathbf{x})\,\mathrm{d}\mathbf{x}\right) \\ &= -\int p(\mathbf{x})\ln\left\{\frac{q(\mathbf{x})}{p(\mathbf{x})}\right\}\mathrm{d}\mathbf{x}.\end{aligned}$$

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)

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

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

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

cross-entropy

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

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

ML and Cross Entropy Equivalence

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

$$= argmin \;\; H(P_{\mathrm{data}}, Q)$$

P_{model}.

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

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.

Mutual Information

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

$$
\mathrm{I}[\mathbf{x}, \mathbf{y}] = \mathrm{H}[\mathbf{x}] - \mathrm{H}[\mathbf{x}|\mathbf{y}] = \mathrm{H}[\mathbf{y}] - \mathrm{H}[\mathbf{y}|\mathbf{x}]
$$

Conditional log-Likelihood

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

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

$$\Downarrow$$

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

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