Information and Entropy

When we resolve disorder, we gain information.

Shannon Information: measure of surprise / uncertainty for an event x with probability $P(x)$, (continuous, additive and symmetric),

$$h(x) = -\log_2 P(x) \quad [\text{bits}]$$

Entropy: measure of average level of disorder / information for a random variable. Given $X = \{x_1, x_2, ..., x_n\}$, with probability distribution $P(X)$,

$$H(X) = \sum_{i=1}^{n} P(x_i) \log_2 \frac{1}{P(x_i)} = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i).$$

Discussions,

- For a Bernoulli RV. X with $P(X) = \begin{cases} p & \text{when } X = 0, \\ 1-p & \text{when } X = 1 \end{cases}$, the binary entropy is defined as,

$$H_2(p) \equiv H(X) = -p \log_2 p - (1-p) \log_2 (1-p).$$

- The entropy is 0 if the distribution is deterministic, i.e. taking a single value with probability 1.
- The entropy is higher for equiprobable distributions since they are more unpredictable.
- Maximal Entropy achieved when $p = 0.5$, i.e., $H_2(0.5) = 1$.
    - For a general case, differentiate the Lagrange function $\mathcal{L}$ from $H(X)$ and set $\frac{\partial\, \mathcal{L}(H(p_1,p_2,...,p_n),\lambda)}{\partial\, p_i} = 0$, with constraint $\sum_{i=1}^{n} p_i = 1$, to find the maximum entropy.

Noiseless channels

The Source Coding Theorem

N i.i.d. random variables each with Entropy $H(X)$ can be compressed into more than $NH(X)$ bits with a negligible risk of loss as N tends to infinity. Conversely, if you compress to fewer than $NH(X)$ bits, you are almost guaranteed to lose information.

Symbol codes C

Binary symbol code C for an ensemble X is $\mathcal{A}_X = \{x_1, x_2, ..., x_n\} \to \{0,1\}^+$. The extended code C^+ is $\mathcal{A}_X^+ \to \{0,1\}^+$.

$$c^+(x_1, x_2, \ldots, x_N) = c(x_1)c(x_2)\ldots c(x_N).$$

The symbol code C **expected (encoded character) length** for an ensemble X is,

$$L(C, X) = \sum_{x \in \mathcal{A}_X} P(x)\ l(x) = \sum_{i=1}^{n} P(x_i)\ l_i.$$

Symbol code **source coding theorem** for an ensemble X,

there exists an encoding C such that the expected encoded character length $L(C, X)$ satisfies,

$$H(X) \leq L(C, X) < H(X) + 1.$$

The minimal expected length only if the the code lengths are equal to the Shannon information contents $l_i = -\log_2 P(x_i)$.

Unique decodability (Prefix codes)

$$\forall x, y \in \mathcal{A}_X, x \neq y \implies c^+(x) \neq c^+(y).$$

The uniquely decodable codeword, with length $l_1, l_2, ..., l_n$, over the binary alphabet $\{0, 1\}$ must satisfy **the Kraft inequality**,

$$\sum_{i=1}^{n} 2^{-l_i} \leq 1.$$

```
 0 : 0
/
\    0 : 10
 1 /
   \    0 : 110
    1 /
      \ ...
```

Huffman coding

Build a binary tree from the leaves to the root,

1. Take the two least probable symbols in the alphabet. They will be given the longest codewords, which will have equal length, and differ only in the last digit.
2. Combine these two symbols into a single symbol, and repeat.

Limitations: assume a const data distribution, thus fixed coding; the extra bit is problematic when $H(X) \approx 1$.

Stream codes

Live data stream, adaptive coding.

Arithmetic coding

Output a single floating point number with a high precision in the range $[0, 1)$, which represents the entire message.

Lempel-Ziv coding

Given a string of symbols str, **Lempel-Ziv complexity** $c(str)$ is the number of longest consecutive substrings that are not repeated from the beginning, e.g. $cstr = \mathrm{A} \mid \mathrm{T} \mid \mathrm{G} \mid \mathrm{T\ G} \implies c(str) = 4$.

Normalized compression distance is a measure of the similarity between two strings x and y based on their Lempel-Ziv complexity.

$$\frac{C(xy) - \min(C(x), C(y))}{\max(C(x), C(y))}$$

Noisy discrete channels

Error Correcting Codes (ECCs)

Repetition code R_z, where z is the number of bits to repeat.

Block Codes (N, k), where $N > k$.

Hamming Codes $(7, 4)$, detecting and correcting 1-bit errors efficiently.

The 7-bit code-word c for 4-bit data word d is defined by the generator matrix G,

$$c^{7\times n} = G^{7\times 4} d^{4\times n} \mod 2 = \begin{bmatrix} 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \\ d_3 \\ d_4 \end{bmatrix} \mod 2.$$

The relationship is,

$$\begin{aligned} c_1 &= d_1 \oplus d_2 \oplus d_4 \\ c_2 &= d_1 \oplus d_3 \oplus d_4 \\ c_4 &= d_2 \oplus d_3 \oplus d_4 \end{aligned}$$

The parity check matrix H is defined by the generator matrix G, and c is valid if all bits in p are 0.

$$p^{3\times n} = H^{3\times 7} c^{7\times n} \mod 2 = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \\ c_6 \\ c_7 \end{bmatrix} \mod 2.$$

The relationship is,

$$
\begin{aligned}
p_1 &= c_1 \oplus c_3 \oplus c_5 \oplus c_7 = 2(d_1 \oplus d_2 \oplus d_4) \\
p_2 &= c_2 \oplus c_3 \oplus c_6 \oplus c_7 = 2(d_1 \oplus d_3 \oplus d_4) \\
p_3 &= c_4 \oplus c_5 \oplus c_6 \oplus c_7 = 2(d_2 \oplus d_3 \oplus d_4)
\end{aligned}
$$

Bayes' rule

Bayes' rule for conditional probability,

$$
P(X|Y) = \frac{P(X,Y)}{P(Y)} = \frac{P(Y|X)P(X)}{\sum_x P(Y|X)P(X)}
$$

for conditional entropy states,

$$
H(X|Y) = H(Y|X) + H(X) - H(Y).
$$

Conditional entropy

measure of uncertainty in random variable Y given event $X = x$,

$$
h(Y|X=x) = -\sum_y P(y|x) \log_2 P(y|x)
$$

conditional entropy: Y given X, i.e, weighted averaging $H(Y|X=x)$ over all values x of X.

$$
\begin{aligned}
H(Y|X) = \sum_x P(x) \cdot h(Y|X=x) &= -\sum_x \sum_y P(x,y) \log_2 P(y|x) \\
&= -\sum_x \sum_y P(x,y) \log_2 \frac{P(x,y)}{P(x)}
\end{aligned}
$$

Discussions,

- when $H(Y|X) = 0$, iff. $Y \subseteq X$, i.e. Y is completely determined by X and $I(X;Y) = H(Y)$.
- when $H(Y|X) = H(Y)$, $H(X,Y) = H(X) + H(Y)$, i.e. $I(X;Y) = 0$, iff. X and Y are independent RVs.
- Y is conditionally independent of Z given X: $P(Y|X,Z) = P(Y|X)$, $H(Y|X,Z) = H(Y|X)$.

Joint entropy

measure of uncertainty in two random variables X and Y.

$$
\begin{aligned}
H(X,Y) &= -\sum_x \sum_y P(x,y) \log_2 P(x,y) \\
&= -\sum_x \sum_y P(x,y) \log_2 P(y|x) \cdot P(x) \quad \text{by chain rule.} \\
&= -\sum_x \sum_y P(x,y) \log_2 P(y|x) - \sum_x (\sum_y P(x,y)) \log_2 P(x) \\
&= -\sum_x \sum_y P(x,y) \log_2 P(y|x) - \sum_x P(x) \log_2 P(x) \quad \text{by marginalization.} \\
&= H(Y|X) + H(X) = H(X|Y) + H(Y)
\end{aligned}
$$

Symmetric property of joint entropy: $H(X,Y) = H(Y,X)$.

Chain rule for multiple RVs probability distribution,

$$
\begin{aligned}
P(X_1, X_2, ..., X_n) &= P(X_1)P(X_2|X_1)...P(X_n|X_1, X_2, ..., X_{n-1}) \\
&= \prod_{i=1}^{n} P(X_i|X_1, X_2, ..., X_{i-1})
\end{aligned}
$$

Joint entropy extended for multiple random variables $X_1, X_2, ..., X_n$,

$$
\begin{aligned}
H(X_1, X_2, ..., X_n) &= H(X_1) + H(X_2|X_1) + ... + H(X_n|X_1, X_2, ..., X_{n-1}) \\
&= \sum_{i=1}^{n} H(X_i|X_1, X_2, ..., X_{i-1})
\end{aligned}
$$

Mutual information

measure the common information between two RVs, i.e., how much information one RV conveys about (propagates to) another. The information gained about Y when we know X.

$$
\begin{aligned}
I(X;Y) &= H(X) + H(Y) - H(X,Y) \\
&= H(X) - H(X|Y) = H(Y) - H(Y|X)
\end{aligned}
$$

Channel capacity: maximum mutual information achievable between input and output random variables of a channel.

$$
C = \max_{p(x)} I(X;Y) \quad [\text{bits/symbol}]
$$

(Symmetric) Channel capacity C with additive noise **independent** of the input X, i.e. $Y = X + \eta$,

$$
\begin{aligned}
C &= \max_{p(x)} I(X;Y) \\
&= \max_{p(x)} \{H(Y) - H(Y|X)\} \\
&= \max_{p(x)} \{H(Y)\} - H(\eta) \quad \text{since } H(Y|X) = H(\eta).
\end{aligned}
$$

Correlation coefficient: measure of the linear relationship strength between two random variables.

$$
\begin{aligned}
\operatorname{Corr}(X,Y) &= \frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)\cdot\operatorname{Var}(Y)}} = \frac{\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]}{\sqrt{\mathbb{E}[(X-\mathbb{E}[X])^2]\cdot\mathbb{E}[(Y-\mathbb{E}[Y])^2]}} \\
\mathbb{E}[X] &= \sum_x xP(x) \\
\mathbb{E}[X^2] &= \sum_x x^2P(x) \\
\operatorname{Var}(X) &= \mathbb{E}[(X-\mathbb{E}[X])^2] \\
\operatorname{Cov}(X,Y) &= \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]
\end{aligned}
$$

Continuous entropy for signals

Differential entropy $h(X)$ of a continuous random variable X with probability density function $p(x)$ is defined as,

$$h(X) = \mathbb{E}[-\log_2 p(X)] = -\int_{\mathcal{X}} p(x)\log_2 p(x)dx.$$

The differential entropy itself has no fundamental physical meaning, but it occurs often enough to have a name. The differential entropy is not the limiting case of the entropy; the entropy of a continuous distribution is infinite. It is not invariant under coordinate transformations.

$$
\begin{aligned}
H(X^\Delta) &= -\sum P_i \log_2 P_i \\
&= -\sum P(x_i)\Delta x \log_2(P(x_i)\Delta x) \\
&= -\sum P(x_i)\Delta x \log_2 P(x_i) - \sum P(x_i)\Delta x \log_2 \Delta x \\
&= -\sum P(x_i)\Delta x \log_2 P(x_i) + \sum \log_2 \frac{1}{\Delta x} \\
&=_{\Delta x\to 0} h(X) + \infty
\end{aligned}
$$

Continuous mutual information $I(X;Y)$ is defined as,

$$
\begin{aligned}
I(X;Y) &= H(X) + H(Y) - H(X,Y) = H(X) - H(X|Y) \\
&= H(Y) - H(Y|X) \\
&= h(Y) + \infty - h(X|Y) - \infty \\
&= h(Y) - h(X|Y)
\end{aligned}
$$

Maximum entropy via Lagrange multipliers and normal distribution,

$$L = \int P(x)\log_2 \frac{1}{P(x)}dx + \lambda\left(\int P(x)dx - 1\right) - \mu(\sigma^2 - \int (x-\bar{x})^2 dx),$$

where variance $\sigma^2 = \sum_i \frac{1}{N}(x_i-\bar{x})^2$ is a constraint. For a communication channel, the power of the signal $P \propto \text{signal}^2 \propto \sigma^2$. We assume that the transmitter is usually power-limited, i.e., the average power of the signal is limited to P.

[The standard electrical power $P = \frac{V^2}{R} = I^2R$.]

Discussions,

- What $P(X)$ gives the $P(Y)$ we want?
- What $P(Y)$ makes $h(Y)$ maximum?
- Answer: Gaussian distributions.

The Gaussian channel $X + \eta = Y$, modelling the relationship between transmitted signal X and received signal Y with additive white noise η, which has equal intensity at all frequencies. Usage: satellite, deep-space communication links and radio transmission.

$$P(X) \sim \mathcal{N}(\mu_x, \sigma_x^2), \quad P(\eta) \sim \mathcal{N}(\mu_\eta, \sigma_\eta^2), \quad P(Y) \sim \mathcal{N}(\mu_y, \sigma_y^2).$$

The differential entropy of a Gaussian distribution $P(z) \sim \mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(z-\mu)^2}{2\sigma^2})$ is given by $\boxed{h(Z) = \frac{1}{2}\log_2(2\pi e\sigma^2)}$. Proof,

$$\begin{aligned}
h(Z \sim \mathcal{N}(\mu, \sigma^2)) &= -\int P(z)\log_2 P(z) dz \\
&= -\int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z-\mu)^2}{2\sigma^2}} \log_2(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z-\mu)^2}{2\sigma^2}}) dz \\
&= -\int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z-\mu)^2}{2\sigma^2}} [\log_2(2\pi\sigma^2)^{-\frac{1}{2}} + \log_2 e^{-\frac{(z-\mu)^2}{2\sigma^2}}] dz \\
&= -\int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z-\mu)^2}{2\sigma^2}} [-\frac{1}{2}\log_2(2\pi\sigma^2) - \frac{(z-\mu)^2}{2\sigma^2}\log_2 e] dz \\
&= \frac{1}{2}\log_2(2\pi\sigma^2) \int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z-\mu)^2}{2\sigma^2}} dz + \frac{\log_2 e}{2\sigma^2} \int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z-\mu)^2}{2\sigma^2}} (z-\mu)^2 dz \\
&= \frac{1}{2}\log_2(2\pi\sigma^2) \times 1 + \frac{\log_2 e}{2\sigma^2} \cdot \sigma^2 \\
&= \frac{1}{2}\log_2(2\pi e\sigma^2).
\end{aligned}$$

The per-symbol capacity of the Gaussian channel is given by,

$$\begin{aligned}
C &= \max_{p(x)} \{h(Y) - h(Y|X)\} \\
&= \max_{p(x)} \{h(Y)\} - h(\eta) \\
&= \frac{1}{2}\log_2(2\pi e\sigma_y^2) - \frac{1}{2}\log_2(2\pi e\sigma_\eta^2) \\
&= \frac{1}{2}\log_2\left(\frac{\sigma_y^2}{\sigma_\eta^2}\right) = \frac{1}{2}\log_2(1 + \mathrm{SNR}) \quad [\text{bits/symbol}]
\end{aligned}$$

where N_0 is the noise power spectral density, and $\mathrm{SNR} = \frac{S}{N} = \frac{P}{N_0 B}$ is the signal-to-noise ratio (dB).

Bandwidth B (Hz, Hertz) of the channel is the difference between the upper and lower frequencies in a continuous band of frequencies, i.e. the frequency range available is $[f_{low}, f_{low} + B]$ Hz.

The **Nyquist theorem** states the Nyquist rate $f_s \geq 2B$ is the minimum sampling rate required to reconstruct signals, which is twice the maximum frequency of the signal.

The **Shannon-Hartley theorem**, for bandwidth-limited channels, states that the theoretical tightest upper bound on the information rate of data (per second) that can be communicated at an arbitrarily low error rate using an average received signal power S through an analog communication channel subject to additive white Gaussian noise (AWGN) of power N,

$$C = B \log_2(1 + \mathrm{SNR}) \quad [\mathrm{bits/s}]$$

Limitations:

- Larger bandwidths brings more noise $N = BN_0$. The spectrum available is limited, and higher bandwidths require larger physical infrastructure.
- Diminishing returns as SNR increases.

Ultra-WideBand (UWB) communications system (bandwidths ~ GHz) can avoid interference with other non-UWB users of the same radio spectrum part, by using a different range of frequencies / transmitting below the ambient noise floor.

Prob. distributions comparison and ML

Entropy of distributions $p(x)$, i.e., the average number of bits needed to encode data with distribution $p(x)$ using a code optimised for $p(x)$.

Cross entropy of $p(x)$ and $q(x)$ measures the average number of bits needed if a code optimised for distribution $q(x)$ is used to encode data with distribution $p(x)$. It is defined as,

$$H(p, q) = \sum_x p(x) \log_2 \frac{1}{q(x)}.$$

Kullback-Leibler divergence / **relative entropy** of $p(x)$ from $q(x)$ tells how many average **additional** number of bits needed if a code optimised for distribution $q(x)$ is used to encode instead for data with distribution $p(x)$. It is defined as,

$$\begin{aligned} D_{KL}(p||q) &= \sum_x p(x) \log_2 \frac{p(x)}{q(x)} \\ &= \sum_x p(x) \log_2 \frac{1}{q(x)} - \sum_x p(x) \log_2 \frac{1}{p(x)} \\ &= H(p, q) - H(p) \end{aligned}$$

Its minimal is achieved when $p(x) = q(x)$, i.e., $D_{KL}(p||q) = 0$.

They both measure the divergence / inefficiency of using a predicted/approximated distribution $q(x)$ instead of the true distribution $p(x)$. In machine learning, they are used as loss functions to measure the difference between the predicted and true distributions or in the variational inference.

They are both asymmetric and thus not a distance. Instead, the entropy distance is defined as,

$$D_H(X, Y) \equiv H(X, Y) - I(X; Y) = H(X|Y) + H(Y|X).$$

Relationship between cross entropy and KL divergence:

$$\boxed{H(p, q) = D_{KL}(p||q) + H(p)}$$

$$\begin{aligned}
\text{LHS} = \sum_x p(x) \log_2 \frac{1}{q(x)} &= \sum_x p(x) \log_2 \frac{p(x)}{q(x)} \cdot \frac{1}{p(x)} \quad \text{by chain rule.} \\
&= \sum_x p(x) \log_2 \frac{p(x)}{q(x)} + \sum_x p(x) \log_2 \frac{1}{p(x)} \\
&= D_{KL}(p||q) + H(p) = \text{RHS.}
\end{aligned}$$

Applications

(a) **Classical thermodynamics** (“entropy”, developed by Clausius, etc.)

express the direction or outcome of spontaneous changes in the system, with an increase representing energy that becomes unavailable for work.

T is the uniform temperature of the system, and dQ is the heat energy transferred to the system. The entropy S change of the system is given by $dS = \frac{dQ}{T}$.

(b) **Statistical mechanics** (“statistical entropy”, developed by Boltzmann, etc.)

The macroscopic state of a system is described by the Gibbs probability distribution over its N microscopic states. p_i is the probability of the system being in state i, and S is the entropy of the system,

$$S = -k_B \sum_i p_i \log_2 p_i$$

When states are equiprobable, $p_i = \frac{1}{N}$, the entropy is maximized at $S = k_B \log_2 N$, and the system is in a state of maximum disorder. It expresses entropy as the logarithm of the number of accessible microstates.

(c) **Information theory** (“information entropy”, developed by Shannon)

measure of disorder in random variable X, with probability distribution $p_X(x)$. When we resolve disorder, we gain information.

$$H(X) = -\sum_x p_X(x) \log_2 p_X(x)$$