Information and Entropy

Shannon Information: measure of surprise / uncertainty. $h(x) = -\log_2 P(x)$ bit(s), for event x with probability $P(x)$. (Continuous, additive and symmetric).

Entropy: measure of disorder. When we resolve disorder, we gain information. Given a RV. (random variable) $X = \{x_1, x_2, ..., x_n\}$, with probability distribution $P(X)$,

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

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,

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

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 $\frac{\partial\ \mathcal{L}(H(p_1, p_2, \ldots, p_n), \lambda)}{\partial\ p_i} = 0$, 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 $\mathcal{A}_X = \{x_1, x_2, ..., x_n\} \rightarrow \{0, 1\}^+$. The extended code C^+ is $\mathcal{A}_X^+ \rightarrow \{0, 1\}^+$.

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

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 $H(Y|X) = H(Y)$, 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 another.

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

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

$$
C = \max_{p(x)} I(X;Y)
$$

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

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

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:

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

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