1 Introduction

Information Theory was originally developed by Claude Shannon for the purpose of an efficient information transmission through a communication channel, i.e. sending shorter messages without losing information. The Huffman Code is an example of a code that preserves all the information, it is the most efficient prefix code.

In this lesson we will discuss only codes that preserve all the information.

2 Efficient Coding

Let $X$ be a random variable of words, and let $l(x)$ be the length of the code of the word $x$.

The cost of the coding of $x$ is:

$$cost = \mathbb{E}[l(x)] = \sum_x P(x)l(x)$$

We would like the cost to be minimal.

A prefix code can be described by a binary tree and therefore the following inequality must hold:

$$\sum_x 2^{-l(x)} \leq 1$$

If every leaf in the tree represents a coded word then an equality will hold.

In an efficient code this term will be as close as possible to 1.

Let’s look at the following minimization problem:

$$\min_{l(x)} \sum_x P(x)l(x)$$

subject to:

$$\sum_x 2^{-l(x)} = 1$$
This problem can be solved using the Lagrangian:

$$\mathcal{L} = \sum_x P(x)l(x) - \lambda \sum_x s^{-l(x)} - 1$$

$$0 = \frac{\partial \mathcal{L}}{\partial l(x)} = P(x_0) + \lambda \ln(2) \cdot 2^{-l(x_0)}$$

$$2^{-l(x_0)} = \frac{-P(x_0)}{\lambda \ln(2)}$$

$$-l(x_0) = \log_2\left(\frac{-P(x_0)}{\lambda \ln(2)}\right)$$

$$l(x_0) = -\log_2(P(x_0)) + \log_2(-\lambda \ln(2))$$

The only way to fulfil the constraint \(\sum_x 2^{-l(x)} = 1\) is when \(l(x_0) = -\log_2(P(x))\)

and therefore: \(\log_2(-\lambda \ln(2)) = 0\) \(\Rightarrow\) \(\lambda \ln(2) = 1\)

and then we get that the condition for the most efficient code for \(x\) is:

$$l(x) = -\log_2 p(x) = \log_2 \frac{1}{p(x)}$$

### 3 Entropy

The entropy is defined as the average number of bits that are needed to code a word:

$$H_p(X) = \mathbb{E}[\log_2 \frac{1}{p(x)}] = \sum_x P(x) \log_2 \frac{1}{p(x)}$$

We have proved that the entropy is a lower limit for the compression ability using a prefix code.

The entropy is a real number and not an integer. If we define a code in which the length of each coded word is the rounded up entropy, the code of a large message will waste a lot of bits. One solution for this problem is to code \(k\) words together. The bigger \(k\) is, the closer the message length will be to its entropy.

#### 3.1 Entropy of a Binary Random Variable

Let \(X\) be a binary random variable that can get the value \(x_0\) and \(x_1\) while \(P(x_0) = p\) and \(P(x_1) = 1 - p\).

$$H(X) = H(p) = p \cdot \log_2 \frac{1}{p} + (1 - p) \cdot \log_2 \frac{1}{1 - p}$$
In Figure 1 we can see that when $X$ is deterministic the entropy is 0, and when there is a uniform distribution the entropy is maximal. In general: $0 \leq H(x) \leq \log_2(x)$

### 3.2 System Entropy

We can write the probability as a function of the variable’s energy:

$$p(x) = \frac{1}{z(T)} \cdot e^{-\frac{E(x)}{kT}}$$

while $z(T)$ is a constant partition function.

In this case the system’s entropy is:

$$H(X) = \sum_x (p(x)E(x)\frac{\log_2(e)}{kT}) + \log_2(z(T))$$

The system’s entropy can tell us how much the system is disordered.

### 3.3 Joint Entropy

The entropy of two random variables is:

$$H(X, Y) = \sum_x \sum_y P(x, y) \log_2 \frac{1}{P(x, y)}$$

In a similar way we can define the entropy of any number of variables.
If $X$ and $Y$ are independents then: $H(X, Y) - H(X) = H(Y)$
If they are dependent then: $H(X, Y) - H(X) < H(X)$
In the general case:

$$H(X, Y) - H(X) = -\sum_x \sum_y P(x, y) \log_2 P(x, y) + \sum_x P(x) \log_2 P(x) =$$

$$= -\sum_x P(x) \cdot \left[ \sum_y \frac{P(x, y)}{P(x)} \log_2 \frac{P(x, y)}{P(x)} \right] =$$

$$= \sum_x P(x) \cdot \left[ -\sum_y P(y|x) \log_2 P(y|x) \right] =$$

$$= \sum_x P(x) H_{P(Y|x)}(Y) = \sum_x P(x) H(Y|x) = H(Y|X)$$

$$\downarrow$$

$$H(X, Y) - H(X) = H(Y|X)$$

$H(Y|X)$ is the average number of bits needed to code $Y$ when we know the value of $X$, while $H(Y)$ is the average number of bits needed to code $Y$ without any other information.

According to the intuition: $H(Y|X) \leq H(Y)$ and when $X$ and $Y$ are independent: $H(Y|X) = H(Y)$ because $X$ doesn’t give us any information about $Y$.

### 3.4 Chain rule of Entropy

$$H(X_1, ..., X_n) = \sum_i H(X_i|X_1, ..., X_{i-1})$$

### 4 Information

The information of $X$ and $Y$ is:

$$I(X : Y) = H(Y) - H(Y|X) = \sum_x \sum_y P(x, y) \log_2 \frac{P(x, y)}{P(x)P(y)}$$

i.e. this is the number of bits we could save in the coding of $Y$ if we knew $X$. 

4
4.1 Information Properties

1. symmetry: \( I(X : Y) = I(Y : X) \)

2. \( X \perp Y \Rightarrow I(X : Y) = 0 \)

3. positivity: \( I(X : Y) \geq 0 \)

**Jensen’s Inequality** If \( f \) is concave then

\[
E[f(x)] \leq f(E[x])
\]

\[
\downarrow
\]

\[-I(X : Y) = E[\log_2 \frac{p(x)p(y)}{p(x,y)}] \leq \log_2 E[\frac{p(x)p(y)}{p(x,y)}] = \log_2 \sum_x \sum_y P(x,y) \frac{P(x)P(y)}{p(x,y)} = \]

\[
= \log_2 \sum_x \sum_y P(x)P(y) = \log_2 1 = 0
\]

\[
\downarrow
\]

\( I(X;Y) \geq 0 \)

4.2 Chain rule for Information

\[
I(X_1, \ldots, X_n; Y) = \sum_i I(X_i; Y|X_1, \ldots, X_{i-1})
\]

And for \( n=2 \):

\[
I(X_1, X_2; Y) = I(X_1; Y) + I(X_2; Y|X_1)
\]

5 What is between Entropy and Information?

5.1 An upper bound for Information

\[
I(X; Y) \leq H(X)
\]

\[
I(X; Y) \leq H(Y)
\]
5.2 Mutual Entropy

\[ H(X,Y) = H(Y) - H(Y|X) = H(Y) - I(X,Y) + H(X) \]

Visualization through Venn diagram:

5.3 Conditional Information of two random variables given the third

\[ I(X,Y|Z) = H(Y|Z) - H(Y|Z,X) = \sum_{x,y,z} P(x,y,z) \log_2 \frac{P(x,y|z)}{P(x|z)P(y|z)} \]

6 Using Entropy to represent Motif LOGOs

Motifs can be represented by giving an Information value for each position in the Motif:
X-axis represents the position in the Motif.
The height of each letter in position \( i \) is calculated the following way:

\[
y = H_{p_0}(x) - H_{p_i}(x)
\]

When \( H_{p_0}(x) \) is the Entropy of \( x \) according to background distribution and \( H_{p_i}(x) \) is the Entropy of \( x \) according to the distribution in position \( i \) of the motif.

The total height of all letters in a certain position is the Information in that position.

7 KL divergence (or relative entropy)

Given a real distribution \( p \) and an erroneous distribution \( q \) our coding inefficiency is measured as:

\[
D_{KL}(p||q) = \sum_x p(x) \log_2 \frac{1}{q(x)} - \sum_x p(x) \log_2 \frac{1}{p(x)} = \sum_x p(x) \log_2 \frac{p(x)}{q(x)}
\]
Coding inefficiency means using a code that is best for distribution \( q \) (was constructed for that distribution) and coding with it sequences from distribution \( p \).

It always holds that

\[
D_{KL}(p||q) \geq 0
\]

The equality holds when

\[
p(x) = q(x)
\]

1. \( D_{KL} \) has no upper bound.
2. \( D_{KL} \) is not symmetric.
3. \( D_{KL} \) doesn’t follow the triangle inequality.

### 7.1 Definition of Information using \( D_{KL} \):

\[
I(X;Y) = D_{KL}(P(X,Y)\|P(X)P(Y))
\]

### 7.2 Using \( D_{KL} \) to measure distance between empirical distribution of random variable \( X \) (\( \hat{P} \)) and the real distribution of \( X \) according to parameter \( \theta \):

\[
\frac{1}{M} l(\theta) = \frac{1}{M} \sum_m \log_2 p(x[m] : \theta)
\]

Empirical probability of \( X \):

\[
\hat{P}(x) = \frac{1}{M} \sum_m 1\{X[m] = x\}
\]

\[
\frac{1}{M} l(\theta) = \sum_x \hat{P}(X) \log_2 p(X : \theta)
\]

\[
D_{KL}(\hat{P}||P(X : \theta)) = -H_{\hat{P}}(X) - \frac{1}{M} l(\theta)
\]

Where the first addend is independent of \( \theta \) and the second addend minimizes the expression when the likelihood is maximized.