Information Theory

Foundations and Applications

Note 1

Introduction to information theory

Shannon Theory

A INTRODUCTION

Information theory answers two fundamental questions:

- what is the ultimate data compression?
  Answer: Shannon entropy!
- what is the ultimate transmission rate of communication?
  Answer: The channel capacity!

These problems were solved by what is known today as Shannon coding theorems, that are the basis of information theory.

- statistical physics (thermodynamics)
- computer science (Kolmogorov complexity)
- statistical inference (Occam's razor)
- probability theory (hypotesis testing)

are some developments that follows from information theory.

Before proving the theorems, we need to introduce some fundamental concepts, starting with the very notion of information.

① What is information and how to measure it

One of the fundamental contributions of Shannon is the notion of a bit as a measure of information.

Physical bit
⇓
two states of a physical system (magnetic spin)

Shannon bit
⇓
measure of the surprise upon learning the outcome of a random binary experiment

If we toss a fair coin and look at the outcome, we learn one bit of information.

The outcome of a coin flip is the physical bit, but it is the information associated with the random nature of the physical bit that we want to measure.

Now that we have a unit to measure information, we need to define the measure.

* Let us assume that every physical system can be described as a random variable

$$X = \{ P_X(x), x \in \mathcal{X} \}$$

$\mathcal{X} \rightarrow$ alphabet

$x \rightarrow$ realization of the random variable

$P_X(x) \rightarrow$ probability distribution associated to x.

Shannon's notion of information contained in the occurrence of an event.

i) I must be a function only on the probability

ii) I must be a continuous function

iii) I must be additive for independent events.

There is only one function that respect these postulates.

Let us consider s independent occurrences of the event x. Then, I must be a function of the total probability $[P_X(x)]^s$.

$$I([P_X(x)]^s) \overset{\text{Ind.}}{=} I\left([P_X(x)]^{s-1}, P_X(x)\right)$$

$$\overset{\text{Add.}}{=} I([P_X(x)]^{s-1}) + I(P_X(x))$$

$$\overset{\text{Ind.}}{=} I\left([P_X(x)]^{s-2}, P_X(x)\right) + I(P_X(x))$$

$$\overset{\text{Add.}}{=} I([P_X(x)]^{s-2}) + 2\,I(P_X(x))$$

$$\vdots$$

$$= s\,I(P_X(x))$$

As a consequente, for any integer t we have

$$I([P_X(x)]^{1/t}) = \frac{t}{t} I([P_X(x)]^{1/t}) = \frac{1}{t} I([P_X(x)]^{t/t})$$

$$= \frac{1}{t} I(P_X(x))$$

Therefore, for any rational number

$$r = \frac{s}{t}$$

we must have

$$I\left([p_x(x)]^r\right) = r\, I\left(p_x(x)\right)$$

Now, any probability can be written as

$$p_x(x) = 2^{\log p_x(x)}$$

And any real number can be arbitrarily well approximated by a rational number. Then

$$I(p_x(x)) = I\left(2^{\log p_x(x)}\right) = \log p_x(x)\, I(2)$$

So, we choose $I(2) = -1$ to get

$$\boxed{I(p_x(x)) = -\log p_x(x)}$$

This is the amount of information contained in the event x. It is how much we learn from knowing the value of X.

I is the measure of the information contained in a single occurrence of the random variable. We are interested in the information contained in the physical system, which is the information source. Therefore, we define the average information

$$\boxed{H(X) = -\sum_{x \in \mathcal{X}} P_X(x) \log P_X(x) = \mathbb{E}\left[I(P_X(x)) \right]}$$

This is Shannon entropy, which measures the uncertaint we have about X, or how much information we gain when we learn the value of X.

For a fair coin we have

$$P_X(x) = \left\{ \frac{1}{2}, \frac{1}{2} \right\} \quad \Rightarrow \quad H(X) = -\frac{1}{2} \log \frac{1}{2} - \frac{1}{2} \log \frac{1}{2}$$

$$x = \{H, T\} \qquad = 1 \text{ bit}$$

Federal University of Goiás
Institute of Physics, 74.690-900, Goiânia, Goiás, Brazil

② Some properties of Shannon entropy

a) Entropy is non-negative

$$H(X) \geq 0$$

This follows because it is the average of a positive quantity.

b) The entropy is invariant with respect to the permutations of the realizations of X.

This is because it depends only on the probabilities, not on the values of the realizations

c) $H(X) = 0$ for a deterministic variable

Let us consider a deterministic distribution

$$P_X(x) = \delta_{x,x_0}$$

$$\Rightarrow H(X) = 0$$

$\Leftarrow$ If $H(X) = 0$, then we have

$P_X(x) \log \frac{1}{P_X(x)} = 0$ for all $x \in \mathcal{X}$, which implies $P_X(x) = 0$ or $P_X(x) = 1$. Since $P_X(x)$ must be a probability distribution, we must have $P_X(x_0) = 1$ and $P_X(x) = 0$ for all others values of x.

This is intuitively expected from the meaning of entropy.

d) $H(X)$ is upper bounded

$$H(X) \leq \log|\mathcal{X}|$$

with $|\mathcal{X}|$ being the cardinality of $\mathcal{X}$.

First, let us consider a uniform random variable

$$P_X(x) = \frac{1}{|\mathcal{X}|} \quad \forall x$$

For this case we have

$$H(X) = \log|\mathcal{X}|$$

Let us now move to the inequality.

We consider a Lagrangian optimization[1] with the Lagrangian being defined as

$$\mathcal{L} = H(X) + \lambda \left(\sum_x p_X(x) - 1 \right)$$

$$\delta \mathcal{L} = \left\{ - \sum_x \left[\log(p_X(x)) - 1 \right] + \lambda \sum_x 1 \right\} \delta p_X(x) = 0$$

$$\Rightarrow -\log p_X - 1 + \lambda = 0 \Rightarrow p_X(x) = 2^{\lambda - 1}$$

Since λ is constant, the probability distribution that maximizes $H(X)$ is the uniform one.

Therefore we conclude that

$$\boxed{0 \leq H(X) \leq \log |\mathcal{X}|}$$

[1] We assume that the entropy is concave. We will prove this latter.

③ Other measures of information

Ⓐ Conditional entropy

If two random variables are correlated, by measuring one of them we obtain information about the other.

Let us define the conditional information content

$$i(x|y) = -\log\left(P_{X|Y}(x|y)\right)$$

The conditional entropy is defined as the expected conditional information content

$$H(X|Y) = \mathbb{E}_{X,Y}\{i(X|Y)\} = \sum_{y}{}' P_Y(y)\, H(X|Y=y)$$

$$= -\sum_{x,y}{}' P_{X,Y}(x,y) \log P_{X|Y}(x|y)$$

where we used $P_{X,Y} = P_Y(y)\, P_{X|Y}(x,y)$

$H(X|Y)$ is the amount of uncertainty about X when we know Y.

(B) Joint Entropy

It is the entropy of the joint random variable (X, Y)

$$H(X,Y) = \mathbb{E}_{X,Y}\{i(X,Y)\} = -\sum_{x,y} P_{X,Y}(x,y) \log P_{X,Y}(x,y)$$

$$H(X,Y) = -\sum_{x,y} P_{X,Y}(x,y)\left\{\log P_X(x) + \log P_{Y|X}(y|x)\right\}$$

$$= -\sum_{x,y} P_{X,Y}(x,y) \log P_X(x) - \sum_{x,y} P_{X,Y}(x,y) \log P_{Y|X}(y|x)$$

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

Entropy is subadditive

$$H(Y) \geq H(Y|X)$$

$$H(X,Y) = H(X) + H(Y|X) \leq H(X) + H(Y)$$

$$\Rightarrow \boxed{H(X,Y) \leq H(X) + H(Y)}$$

(c) The mutual information

It is a measure of the correlations between two random variables

$$I(X:Y) = H(X) - H(X|Y)$$

$$= \sum_{x,y} P_{X,Y}(x,y) \log\left(\frac{P_{X,Y}(x,y)}{P_X(x)\,P_Y(y)}\right)$$

Since $H(X) \geq H(X|Y) \Rightarrow \boxed{I(X:Y) \geq 0}$

The equality is achieved if and only if the random variables are independent

$$P_{X,Y}(x,y) = P_X(x)\,P_Y(y) \Rightarrow I(X:Y) = 0$$

D) Relative entropy

The relative entropy is a measure of how far one probability distribution $p_X(x)$ is from another one $q_X(x)$.

It is defined as

$$D(p\|q) = \sum_x p_X(x) \log\left(\frac{p_X(x)}{q_X(x)}\right)$$

If $\mathrm{supp}(p) \subseteq \mathrm{supp}(q)$

The mutual information can be written in terms of the relative entropy as

$$I(X:Y) = D\left(p_{X,Y}(x,y) \,\|\, p_X(x) \times p_Y(y)\right)$$

This tells us how far we are from independence!

We can use this result to prove that entropy is a concave function.

$$H(X) = -\sum_x{}' P_X(x) \log P_X(x) = -\sum_x P_x \log\left(\frac{P_x U_x}{U_x}\right)$$

$$= -\sum_x P_x \log \frac{P_x}{U_x} - \sum_x P_x \log U_x$$

Taking $U_x = \frac{1}{|X|}$ (uniform distribution)

$$H(X) = -D(P_X \| U_X) + \log|X|$$

$$\Rightarrow \log|X| - H(X) = D(P_X \| U_X)$$

Now, we have that $D(P_X \| q_X)$ is convex

$$D(\lambda p_1 + (1-\lambda)p_2 \| \lambda q_1 + (1-\lambda)q_2) = \sum_x{}' [\lambda p_1 + (1-\lambda)p_2] \log \frac{\lambda p_1 + (1-\lambda)p_2}{\lambda q_1 + (1-\lambda)q_2}$$

$$\leq \sum_x \left[\lambda p_1 \log \frac{\lambda p_1}{\lambda q_1} + (1-\lambda)p_2 \log \frac{(1-\lambda)p_2}{(1-\lambda)q_2}\right]$$

$$= \lambda D(p_1 \| q_1) + (1-\lambda) D(p_2 \| q_2) \quad \square$$

The second step follows from the fact that, for real positive numbers a_i and b_i

$$\left(\sum_{i=1}^{n} a_i\right) \log \frac{\sum_i a_i}{\sum_i b_i} \leq \sum_i a_i \log \frac{a_i}{b_i}$$

Therefore, for our special case

$$D(\lambda p_1 + (1-\lambda) p_2 \| \lambda u_1 + (1-\lambda) u_2) \leq \lambda D(p_1 \| u_1) + (1-\lambda) D(p_2 \| u_2)$$

$$\Rightarrow D(\lambda p_1 + (1-\lambda) p_2 \| u) \leq \lambda D(p_1 \| u) + (1-\lambda) D(p_2 \| u)$$

which implies

$$\log|X| - H(\lambda p_1 + (1-\lambda) p_2) \leq \lambda(\log|X| - H(p_1)) + (1-\lambda)(\log|X| - H(p_2))$$

$$= \log|X| - \lambda H(p_1) - (1-\lambda) H(p_2)$$

$$\Rightarrow H(\lambda p_1 + (1-\lambda) p_2) \geq \lambda H(p_1) + (1-\lambda) H(p_2)$$

$$\boxed{H(X) \text{ is concave}}$$

(4) Data processing inequality

Let p and q be two probability distributions, and let Δ be a classical channel. Then

$$\boxed{D(p\|q) \geq D(\Delta p\|\Delta q)}$$

Proof

If $\mathrm{supp}(p) \not\subseteq \mathrm{supp}(q)$, then $D(p\|q) = \infty$ and the inequality is trivial.

If $\mathrm{supp}(p) \subseteq \mathrm{supp}(q)$, we have

$$\mathrm{supp}(\Delta p) \subseteq \mathrm{supp}(\Delta q)$$

Let us start by rewritting the the quantities appearing in the inequality.

$$D(\Delta p\|\Delta q) = \sum_y (\Delta p)(y) \log \frac{(\Delta p)(y)}{(\Delta q)(y)}$$

$$D(\Delta p \| \Delta q) = \sum_{x,y} \Delta(y|x)\, p(x) \log \frac{(\Delta p)(y)}{(\Delta q)(x)}$$

$$= \sum_x p(x) \left[\sum_y \Delta(y|x) \log \frac{(\Delta p)(y)}{(\Delta q)(y)} \right]$$

$$= \sum_x p_x \log \exp \left[\sum_y \Delta(y|x) \log \frac{\Delta p(y)}{\Delta q(y)} \right]$$

$\Delta(y|x)$ is the conditional probability distribution defining the classical channel

$$X \xrightarrow{\Delta} Y$$

This implies that

$$D(p \| q) - D(\Delta p \| \Delta q) = D(p \| r)$$

$$r = q(x) \exp \left[\sum_y \Delta(y|x) \log \frac{\Delta p(y)}{\Delta q(y)} \right]$$

Now, note that

$$\sum_x r(x) = \sum_x q \exp\left[\sum_y \Delta(y|x) \log \frac{\Delta p(y)}{\Delta q(y)}\right]$$

$$\leq \sum_x q(x) \sum_y \Delta(y|x) \exp\left\{\log \frac{\Delta p(y)}{\Delta q(y)}\right\}$$

$$= \sum_x q(x) \sum_y \Delta(y|x) \frac{\Delta p(y)}{\Delta q(y)}$$

$$= \sum_y \left[\sum_x q(x) \Delta(y|x)\right] \frac{\Delta p(y)}{\Delta q(y)}$$

$$= \sum_y \Delta p(y) = 1$$

$$\Rightarrow \sum_x r(x) \leq 1$$

$$\Rightarrow D(p\|r) \geq 0$$

which proves data processing inequality!

(5) Fano's inequality

$$X \xrightarrow{P_{Y|X}(y|x)} Y$$

↳ noisy communication channel

Y is processed and the best estimation $\hat{X}$ of X is produced. The probability error is

$$P_e = \Pr\{\hat{X} \neq X\}$$

If the channel is noiseless, we have

$$P_{Y|X}(y|x) = \delta_{y,x} \Rightarrow H(X|Y) = 0$$

If the noise increases, $H(X|Y)$ increases

$H(X|Y)$ quantifies the amount of information lost in the channel.

Fano's inequality provides a quantitative relation between P_e and $H(X|Y)$

let us assume

$$X \longrightarrow Y \longrightarrow \hat{X}$$

Then

$$H(X|Y) \leq H(X|\hat{X}) \leq h_2(P_e) + P_e \log(|\mathcal{X}|-1)$$

with $h_2(p) = -p \log p - (1-p) \log(1-p)$

Note that

$$\lim_{P_e \to 0} \left(h_2(P_e) + P_e \log(|\mathcal{X}|-1) \right) = 0$$

$$\Rightarrow H(X|Y) = 0$$

as it should.

Proof:

Let E denote an error indicator

$$E = \begin{cases} 0 : X = \hat{X} \\ 1 : X \neq \hat{X} \end{cases}$$

Consider the entropy

$$H(EX|\hat{X}) = H(X|\hat{X}) + H(E|X\hat{X})$$

If we know both X and $\hat{X}$, there is no uncertainty about E. Therefore

$$H(E|X\hat{X}) = 0$$

And

$$H(EX|\hat{X}) = H(X|\hat{X}) \qquad (1)$$

Now, let us consider the following chain

$$X \rightarrow Y \rightarrow \hat{X}$$

then, we have

$$I(X;Y) \geq I(X:\hat{X}) \Rightarrow H(X|\hat{X}) \geq H(X|Y) \quad ②$$

↳ data processing inequality

Now, we have

$$H(EX|\hat{X}) = H(E|\hat{X}) + H(X|E\hat{X})$$

Conditioning reduces entropy

$$\leq H(E) + H(X|E\hat{X})$$

$$= h_2(p_e) + p_e H(X|\hat{X}, E=1) + (1-p_e) H(X|\hat{X}, E=0)$$

$$\leq h_2(p_e) + p_e \log(|X|-1) \quad ③$$

When there is no error $(E=0)$, there is no uncertainty about X. Also, the uncertainty about X, when $\hat{X}$ is available and we have an error $(E=1)$, is less than the uncertainty of a uniform distribution $1/(|X|-1)$. Fano's inequality follows from ①, ② and ③.