Introduction to information theory

September 4, 2022

What is information theory

- Wikipedia – "*Information theory is the scientific study of the quantification, storage, and communication of digital information*".
- Developed in fields of communication and signal processing.
- A prominent figure is Claude Shannon.
- In statistics and data science, it concerns the quantification of the "amount of information in data" for certain tasks (estimation, inference, etc.).
  - For example, **Fisher Information** measures the amount of information the data carry for an unknown parameter to be estimated. Higher Fisher Information means the likelihood surface is sharp near the MLE so that the estimate is more precise (with smaller standard error).
- Information theories are nowadays widely used in AI and machine learning.

Information content

- Also known as “self-information”, “Shannon information”, “surprisal”, etc.
- Quantify the level of “surprise” of a particular outcome.
- Definition. Assume a *discrete* random variable X takes values in a set $\mathcal{X}$ with probability mass function $p(x)$. For observing a particular event $x_c \in \mathcal{X}$, the information content $I(x_c) = -\log p(x_c)$.
- Smaller $p(x_c)$ leads to larger $I(x_c)$: rare events carry more “information”.

Note: there are different definitions of using $\log$ or $\log 2$, but it’s not important.

Entropy

Measures the information of a random variable.

Definition:
Given a *discrete* random variable X taking values in a set $\mathcal{X}$ with probability mass function $p(x)$, the entropy (also known as "**Shannon entropy**".) $H(X)$ is defined as

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x)$$

For a *continuous* random variable X with probability density function $p(x)$, we have the "**differential entropy**" or "**continuous entropy**":

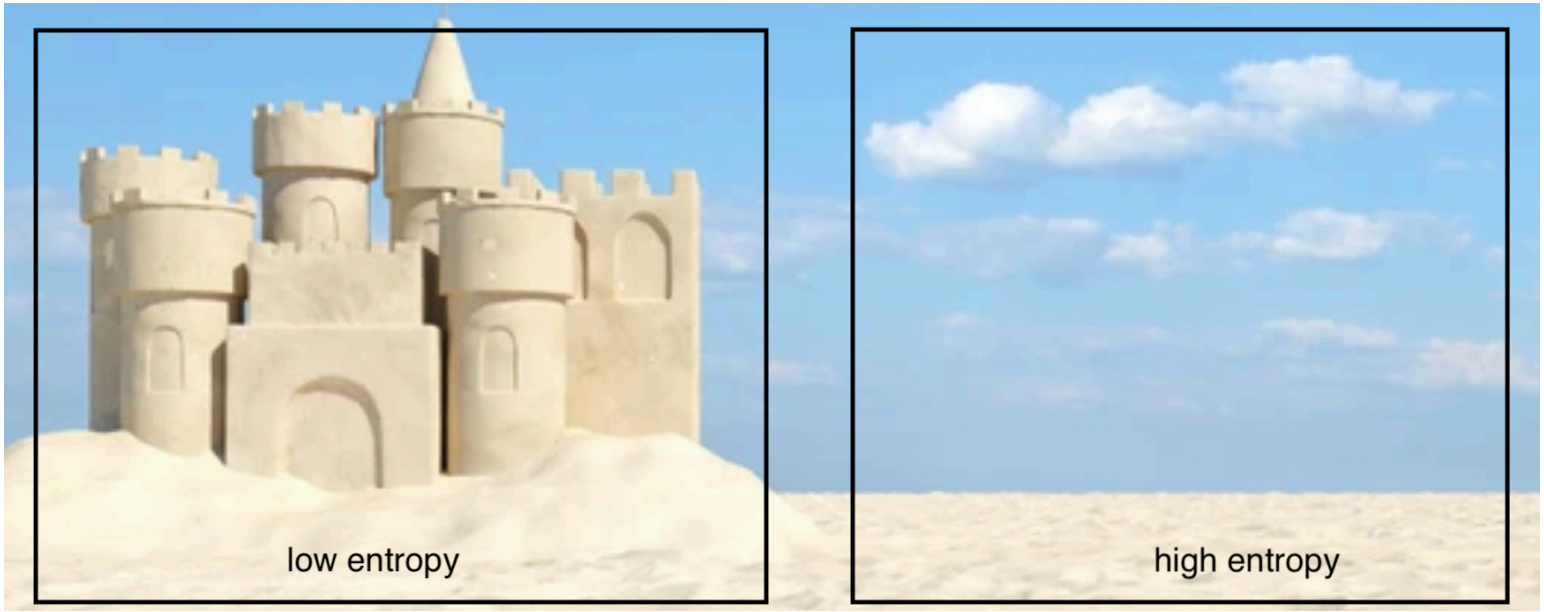
$$H(X) = -\int p(x) \log p(x) dx$$

Notes:

- Entropy is the expected values of the the information content for a random variable, i.e., $H(x) = E[I(x)]$.

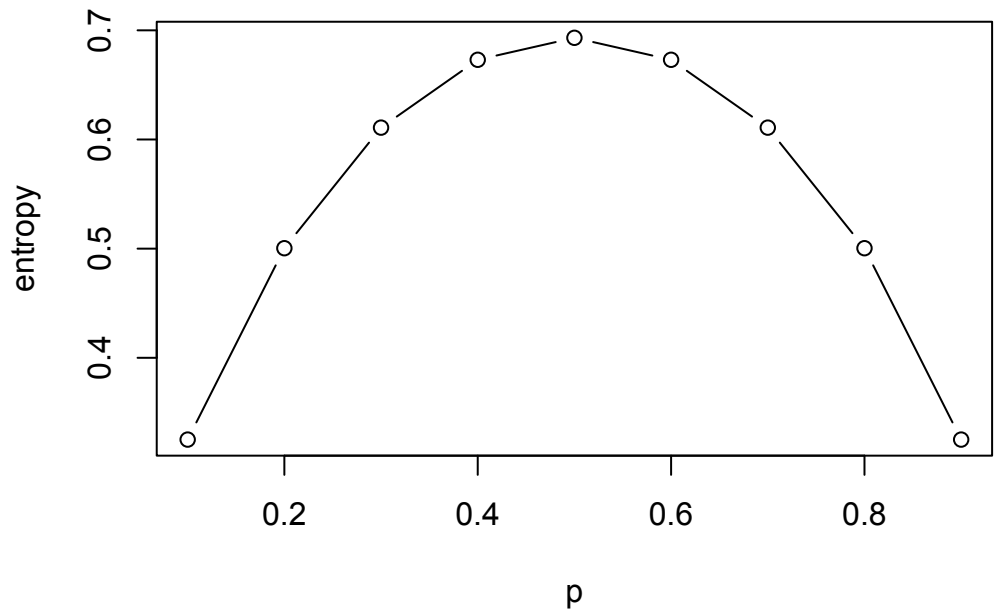
Entropy – basic properties

- Larger entropy indicates higher uncertainty (and more chaotic).
  - For example, the second law of thermodynamics states that in an isolated physical system, the entropy always increases. For example, the overall entropy of the universe always increases (even though locally it can decease, such as on earth).



- Shannon entropy (for discrete r.v.):
  - $H(X) \geq 0$. $H(X) = 0$ if and only if X is certain, i.e., it takes one value with probability 1.
  - $H(X)$ is maximal when all p_i's are equal, i.e., fair coin/dice produces the maximum entropy.

```
p = seq(0.1, 0.9, by=0.1)
entropy = - (p*log(p) + (1-p)*log(1-p))
plot(p, entropy, type="b")
```



- Differential entropy (for continous r.v.):
  - $H(X)$ can be negative, because the density can be greater than 1.
  - For r.v. x with a **fixed variance**, $H(X)$ is the maximum if x follow Gaussian distribution.
  - For **non-negative** r.v. x with a **fixed mean**, $H(X)$ is the maximum if x follow exponential distribution.
  - For **bounded** r.v. x (for example, beta distribution), $H(X)$ is the maximum if x follow uniform distribution.

Joint (total) entropy for multiple random variables

- The **total entropy** of two random variables X and Y is defined as $H(X, Y) = -\sum_{x,y} p(x, y) \log p(x, y)$.
- Properties:
  - Symmetry: $H(X, Y) = H(Y, X)$.
  - $H(X, Y) \geq \max[H(X), H(Y)]$.
  - $H(X, Y) \leq H(X) + H(Y)$, where the equality achieves when X any Y are independent. (proof it!)

  All above properties generalize to n random variable cases.

Conditional Entropy

- **Conditional entropy** of Y given $X = x$ is defined as $H(Y|X = x) = -\sum_y p(y|x) \log p(y|x)$.
- **Conditional entropy** of Y given X is defined as the expected value of $H(Y|X = x)$ (marginalized over X):

  $H(Y|X) = -\sum_x p(x) \sum_y p(y|x) \log p(y|x) = -\sum_{x,y} p(x, y) \log p(y|x)$
- Properties:
  - $H(Y|X) \leq H(Y)$. The quality achieves when Y and X are independent.
  - Chain rule: $H(Y|X) = H(X, Y) - H(X)$.
  - Bayes rule: $H(Y|X) = H(X|Y) - H(X) + H(Y)$.

Mutual information

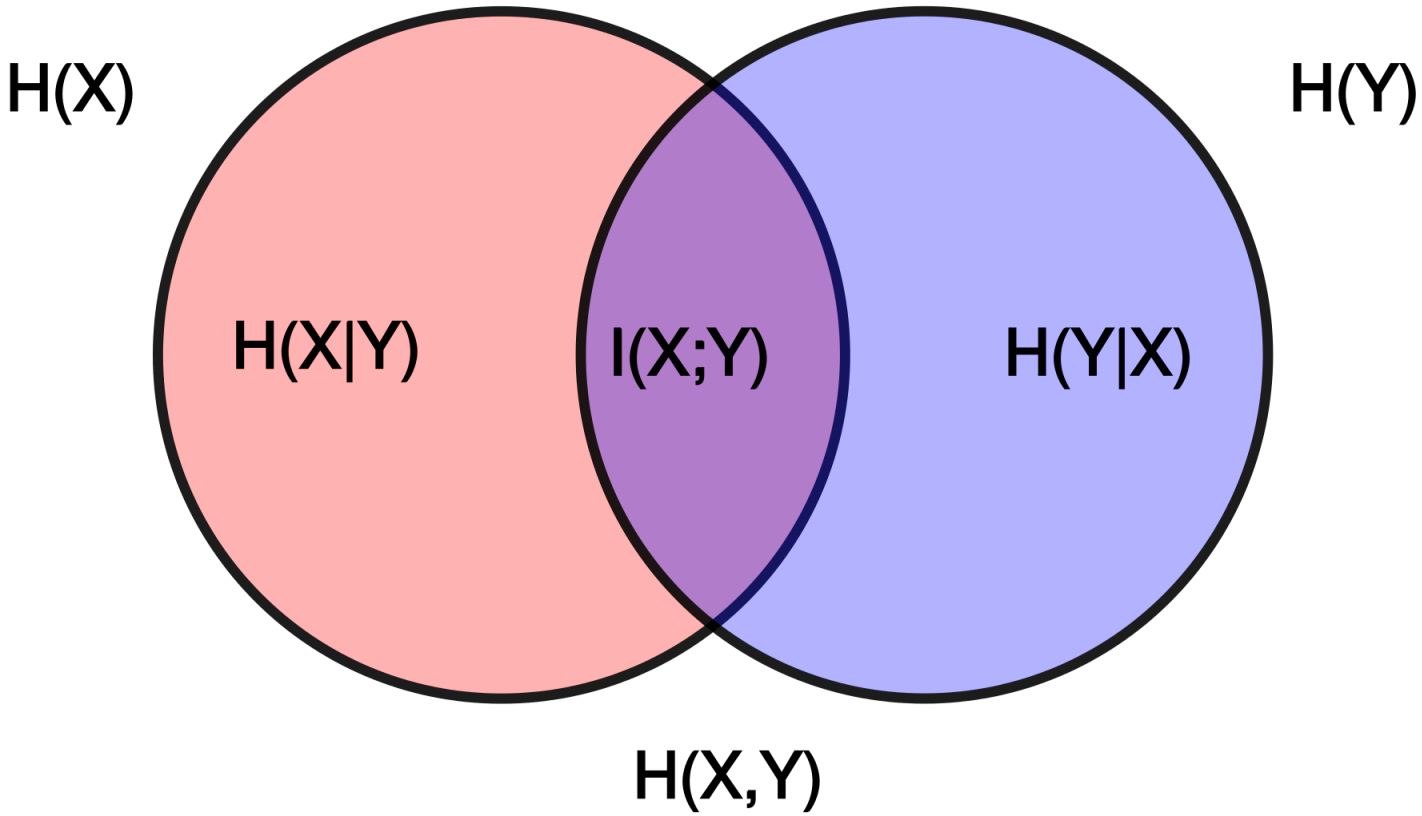
Given two random variables X and Y, the mutual information $I(X, Y)$ quantifies the amount of information one can obtain for X by observing Y, or vice versa.

Discrete version definition:

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

Basic properties:

- $I(X, Y) \geq 0$.
- $I(X, Y) = I(Y, X)$.
- $I(X, Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X, Y) = H(X, Y) - H(X|Y) - H(Y|X)$


H(X)
H(Y)
H(X|Y)
I(X;Y)
H(Y|X)
H(X,Y)

Some proofs of the properties

1. Positivity of the mutual information, i.e., $I(X, Y) \geq 0$.

   Proof: We use the inequality $\ln x \leq x - 1$ for all $x > 0$, where the equality achieves when $x = 1$:

$$-I(X, Y) = \sum_{x,y} P(x, y) \log \left(\frac{P(x)P(y)}{P(x, y)} \right) \leq \sum_{x,y} P(x, y) \left(\frac{P(x)P(y)}{P(x, y)} - 1 \right) = 0$$

   The equality achieves when $\frac{P(x)P(y)}{P(x,y)} = 1$, or when X and Y are independent. □

2. $I(X, Y) = H(X) - H(X|Y)$.

   Proof:

$$I(X, Y) = \sum_{x,y} P(x, y) \log \left(\frac{P(x, y)}{P(x)P(y)} \right) = \sum_{x,y} P(x, y) \log P(x|y) - \sum_{x,y} P(x, y) \log P(x)$$

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

3. Combine the above two, we have the inequality $H(X) \geq H(X|Y)$.

Do other proofs by yourself!

Relative entropy: Kullback-Leibler divergence

Definition: given two probability distribution P and Q:

- For discrete distributions,

$$D_{KL}(P \parallel Q) = \sum_x P(x) \log\left(\frac{P(x)}{Q(x)}\right).$$

- For continuous distributions:

$$D_{KL}(P \parallel Q) = \int_x p(x) \log\left(\frac{p(x)}{q(x)}\right) dx.$$

The KL divergence can also be written as $D_{KL}(P \parallel Q) = \int \log\left(\frac{dP}{dQ}\right) \frac{dP}{dQ}\, dQ$, which is the entropy of P relative to Q.

A simple interpretation: the expected information loss from using Q as a model when the real data distribution is P.

Gibbs inequality

Discrete version: Given two discrete probability distributions P and Q, where $P = \{p_1, \ldots, p_n\}$ and $Q = \{q_1, \ldots, q_n\}$, we have

$$\sum_{i=1}^{n} p_i \log p_i \geq \sum_{i=1}^{n} p_i \log q_i.$$

with equality if and only if $p_i = q_i$.

Proof:

We will prove in the natural log scale (ln). Since $\ln x \leq x - 1$ for all $x > 0$, with equality if and only if $x = 1$, we have:

$$\sum_i p_i \ln \frac{q_i}{p_i} \leq \sum_i p_i \left(\frac{q_i}{p_i} - 1 \right) = \sum_i q_i - \sum_i p_i = 0$$

□

The Gibbs inequality shows that the KL divergency is non-negative, i.e., $D_{\mathrm{KL}}(P \parallel Q) \geq 0$, with equality holds iff $P = Q$.

Cross entropy

Definition: for two distributions P and Q: $H(P, Q) = -\sum_x P(x) \log Q(x)$.

It's easy to see that $H(P, Q) = H(P) + D_{KL}(P, Q)$, or the cross entropy is the sum of entropy of P and the KL divergence of P and Q.

Relationships of entropy, relative entropy (KL divergence), and cross entropy: Assume we have data following a distribution P, and we have another (approximated, estimated, etc.) distribution Q:

- Entropy: the information in the data itself (self-information), if we know P.
- Relative entropy (KL divergence): how the approximated distribution Q differs from the true distribution P, based on the data.
- Cross entropy: the information in the data assuming (wrongly) the distribution is Q. This value will be greater than the entropy, and the difference is the KL divergence.

A simple example


```
x = sample(c(1,2), 100, replace=TRUE, prob=c(0.5,0.5))
Px = c(0.5, 0.5)
Qx = c(0.1, 0.9)

## entropy
> - sum(Px[x] * log(Px[x]))
[1] 34.65736

## cross entropy
> - sum(Px[x] * log(Qx[x]))
[1] 61.29725

## KL divergence
> r = Px / Qx
> sum(Px[x] * log(r[x]))
[1] 26.63989
```

Cross entropy loss (CEL) function

- Cross entropy is widely used as the loss function in machine/deep learning.
- Also known as logarithmic loss, log loss or logistic loss.
- Note, since the entropy doesn't involve the model (Q), minimizing the cross entropy is equivalent to minimizing the KL divergence. Because of its simplicity, it's often used as the loss for training.

For a prediction task with K classes, the cross entropy loss for one observation is defined as

$$l_{CE}(= -\sum_{i=1}^{K} t_i \log(p_i)$$

Here, t_i is the one-hot encoded true label, and p_i is the predicted (softmax) probability for the i-th class. Then the overall cross entropy loss is the sum over all observations.

Note: the predicted probability p_i depends on a model (such as a neural network) that we want to estimate.

Minimizing CEL vs MLE

Given iid Bernoulli random variable $x_i, i = i, \ldots, n$ from $Bernoulli(\theta)$.

Since $P(x_i) = \theta^{x_i}(1-\theta)^{1-x_i}$, the data log-likelihood is

$$L(\theta) = \sum_i x_i \log\theta + (1-x_i)\log(1-\theta)$$

This is exactly the negative cross entropy loss.

- Minimizing the CLE is equivalent to maximizing the likelihood in simple cases.
- For complex high-dimensional data using a complex prediction model (such as a large neural network with hidden layers), the log-likelihood is not necessarily the negative cross entropy loss. But who cares, since the data can't even be described by a distribution.

Application of KL divergence

- KL divergence measures the information loss in the fitted model relative to the true model.
- It is widely used in machine/statistical learning field. A famous application is the variational inference.
  - One wants to obtain an *intractable* target distribution P
  - Approximate P by an easier distribution Q.
  - Minimize the KL divergence between P and Q – this is an optimization problem that gives the parameters for Q.