Information theory

Some basic definitions: The case of continuous random variables

Georgios Ropokis

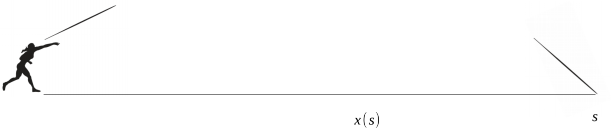
CentaraleSupélec, Campus Rennes

Table of contents

Continuous random variables

Definition of a continuous random variable

- We define a continuous random variable as a mapping between a random experiment and a sample space that is an infinite and uncountable subset of $\mathbb{R}$.
- Example: For a javelin throw experiment we can define the event that the javelin lands at a point s. The distance $x(s)$ covered by the javelin is then a continuous random variable.



Examples of continuous random variables

- Measurements obtained by an analogue sensor.
- The received signal strength for a wireless receiver, at a random position.
- Thermal noise at an electric circuit.
- The interference reaching a wireless communications receiver.

Characterizing a continuous random variable: Probability density function

- Assigning a non zero probability of occurrence to each one of the possible values of a continuous random variable X does not lead to a probability sum equal to 1.
- Instead, we assign non-zero probabilities for each one of the subintervals of the range of values of random variable X.
- We define the cumulative distribution function $F_X(x)$ as the function that gives us the probability that $X \leq x$, i.e., as:

$$F_x(x) = \Pr(X \leq x). \tag{1}$$

- Probability density function: If the derivative of the cumulative distribution function exists, then we call this derivative $f_X(\cdot)$, the probability density function of X. Using $f_X(x)$ we can calculate $\Pr\{a \leq x \leq b\}$ as:

$$\Pr\{a \leq x \leq b\} = \int_a^b f_X(x)\, dx. \tag{2}$$

- Both the probability density function and the cumulative distribution function uniquely characterize random variable X.

Properties of probability density functions

Property 1: Non negativity

$$f_X(x) \geq 0, \quad -\infty < x < \infty. \tag{3}$$

Property 2: Integration property

$$\int_{-\infty}^{\infty} f_X(x)\, dx = 1. \tag{4}$$

Some well known continuous distributions

Uniform distribution

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & a < x < b \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Exponential distribution

$$f_X(x) = \begin{cases} \lambda \exp\left(-\lambda x\right), & x > 0 \\ 0, & x \leq 0. \end{cases} \tag{6}$$

Gaussian distribution

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\left(x-\mu\right)^2}{2\sigma^2}\right), -\infty < x < \infty. \tag{7}$$

Implicit characterization of random variables

Expectation of a random variable

We define the expectation of a random variable as the integral:

$$\mathbb{E}\{X\} = \int_{-\infty}^{\infty} x f_X(x)\, dx. \tag{8}$$

Expectation of a function of a random variable

For a random variable $Y = g(X)$, we define the expectation of Y as:

$$\mathbb{E}\{Y\} = \mathbb{E}\{g(X)\} = \int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx. \tag{9}$$

Differential Entropy

Definition of Differential Entropy

Definition: Differential entropy

Let X be a continuous random variable having a non-zero probability density function $f_X(x)$ over a support set $\mathcal{D}$ and a cumulative distribution function $F_X(x)$. We define the differential entropy as:

$$h(X) = -\int_{\mathcal{D}} f_X(x) \log(f_X(x))\, dx.$$

Remark: Similar to the case of a discrete random variable, the differential entropy is determined by the probability density function of the considered random variable. For this reason, we can also use notation $h(f)$ instead of $h(X)$.

Differential entropy for some known distribution functions

Uniform distribution

Let us consider a random variable uniformly distributed in the interval (a, b). We can then calculate its differential entropy as:

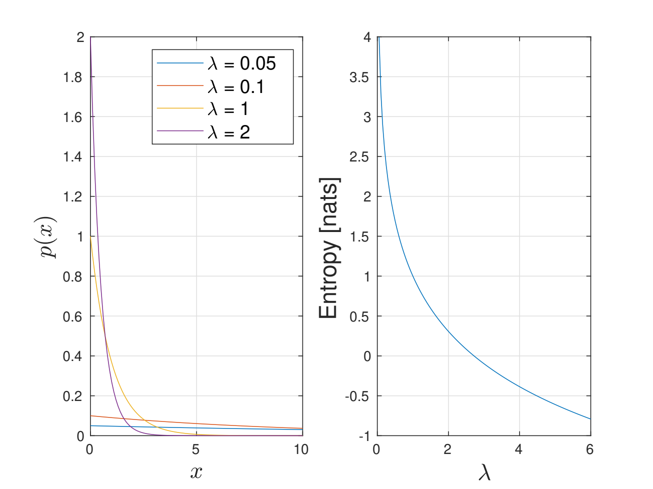
$$h(X) = -\mathbb{E}\left\{\log\left(f_X(X)\right)\right\} = \int_a^b \frac{1}{b-a}\log(b-a)\,dx = \log(b-a)\,. \tag{10}$$

Remark: Assuming $b - a < 1$, the differential entropy becomes negative. As a result, unlike the case of the entropy of a discrete random variable positivity is not necessary for the differential entropy.

Exponential distribution (Entropy in nats)

$$\begin{aligned} h(X) &= -\int_0^\infty \lambda\exp(-\lambda x)\ln\left(\lambda\exp(-\lambda x)\right)dx \\ &= -\ln\lambda + \lambda\int_0^\infty x\lambda\exp(-\lambda x)\,dx = 1 - \ln\lambda \end{aligned} \tag{11}$$

Differential entropy of exponential distribution



As λ increases X becomes more and more limited (with respect to the values taken with significant probability) and the entropy decreases. Differential entropy is still a measure of uncertainty!

Differential entropy for some known distribution functions

Gaussian distribution (Entropy in nats)

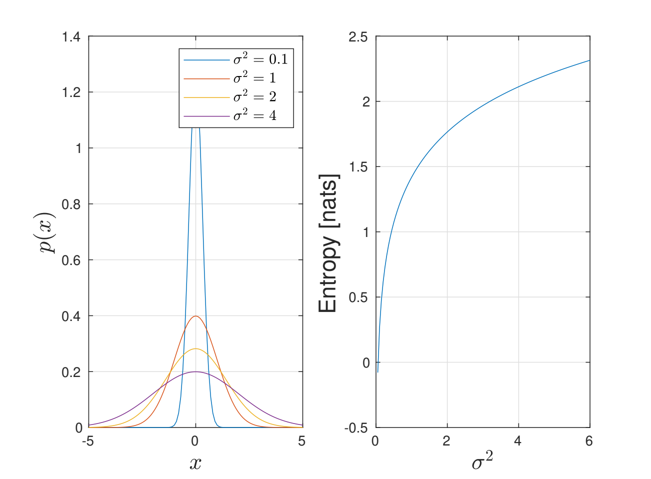
We consider a random variable X following a normal distribution of the form:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right). \tag{12}$$

The differential entropy (in nats) is then given as:

$$\begin{aligned}
h(X) &= -\int_{-\infty}^{+\infty} \frac{\exp\left(-\frac{x^2}{2\sigma^2}\right)}{\sqrt{2\pi\sigma^2}} \ln\left(\frac{\exp\left(-\frac{x^2}{2\sigma^2}\right)}{\sqrt{2\pi\sigma^2}}\right) dx \\
&= \frac{\ln 2\pi\sigma^2}{2} + \frac{1}{2\sigma^2}\int_{-\infty}^{+\infty} \frac{x^2 \exp\left(-\frac{x^2}{2\sigma^2}\right)}{\sqrt{2\pi\sigma^2}} \\
&= \frac{\ln 2\pi\sigma^2}{2} + \frac{\mathbb{E}\left\{X^2\right\}}{2\sigma^2} = \frac{1}{2}\ln 2\pi e\sigma^2
\end{aligned}$$

Differential entropy of a normal random variable



Regardless of the distribution (e.g. Gaussian or exponential), low entropy indicates that the random variable is confined to a small set of values, while high entropy indicates that the random variable is more dispersed.

Connection between differential entropy and discrete entropy

Quantization and entropy

- Let X be a continuous random variable having a probability density function $f_X(x)$
- Let us also consider dividing the range of values of X in n bins, each one of length $\Delta = \frac{1}{2^n}$.
- From the Mean Value Theorem, we have that for the interval $[i\Delta, (i+1)\Delta)$, there exists an $x_i \in [i\Delta, (i+1)\Delta)$ such that $f_X(x_i)\,\Delta = \int_{i\Delta}^{(i+1)\Delta} f_X(x)\,dx$
- Question: By introducing the discrete random variable:

$$X_Q = x_i, \;\; \text{if} \;\; i\Delta \leq X < (i+1)\,\Delta,$$

How is the entropy of X_Q connected to $h(X)$?

Connection between differential entropy and discrete entropy

Quantization and entropy

We start by introducing the probabilities $p_i = \Pr\left(i\Delta \leq X < (i+1)\Delta\right) = \int_{i\Delta}^{(i+1)\Delta} f_X(x)\,dx = f_X(x_i)\,\Delta$. We can then calculate the entropy of X_Q as:

$$\begin{aligned} H(X_Q) &= \sum_{i=-\infty}^{+\infty} p_i \log p_i = -\sum_{i=-\infty}^{+\infty} f_X(x_i)\,\Delta \log\left(f_X(x_i)\,\Delta\right) \\ &= -\sum_{i=-\infty}^{+\infty} f_X(x_i)\,\Delta \log\left(f_X(x_i)\right) - \sum_{i=-\infty}^{+\infty} f_X(x_i)\,\Delta \log(\Delta) \qquad (13) \\ &= -\sum_{i=-\infty}^{+\infty} f_X(x_i)\,\Delta \log\left(f_X(x_i)\right) - \log\Delta. \end{aligned}$$

Theorem: If the density $f(x)$ of X is Riemann integrable, then:

$$H(X_Q) + \log\Delta \to h(X), \quad \text{as} \quad \Delta \to 0.$$

Result: For a continuous random variable X, the number of bits to describe X with n-bit accuracy is equal to $h(X) + n$.

Joint and conditional differential entropy

Continuous random vectors

We characterize a random vector $\mathbf{X} = [X_1, \ldots, X_N]$ by means of its joint distribution function $F_{\mathbf{X}}(x_1, \ldots, x_N)$ defined as:

$$F_{\mathbf{X}}(x_1, \ldots, x_N) = \Pr(X_1 \leq x_1, \ldots, X_n \leq x_N). \tag{14}$$

We also define the joint probability distribution function as:

$$f_{\mathbf{X}}(x_1, \ldots, x_N) = \frac{\partial^N F_{\mathbf{X}}(x_1, \ldots, x_N)}{\partial x_1 \cdots \partial x_N} \tag{15}$$

Expectation operators for random vectors

Definition: Expectation of a function of a random vector

Let $g(\mathbf{X}) : \mathbb{R}^N \to \mathbb{R}$ be a multivariate function of X. Let $\mathcal{D}_\mathbf{X}$ be the domain of support for the joint probability density function of $\mathbf{X}$. We then define its expectation as the following N-dimensional integral:

$$\mathbb{E}\left\{g(\mathbf{X})\right\} = \int_{\mathcal{D}_\mathbf{X}} g(x_1, \ldots, x_N)\, f_\mathbf{X}(x_1, \ldots, x_N)\, dx_1 \ldots x_N \tag{16}$$

Some important moments

Expectation of a vector: $\mathbb{E}\left\{\mathbf{X}\right\} = \left[\mathbb{E}\left\{X_1\right\}, \ldots, \mathbb{E}\left\{X_N\right\}\right]^T$

Covariance matrix: $\mathbf{\Sigma} = \mathbb{E}\left\{(\mathbf{X} - \mathbb{E}\left\{\mathbf{X}\right\})(\mathbf{X} - \mathbb{E}\left\{\mathbf{X}\right\})^T\right\}$

Correlation matrix: $\mathbf{C} = \mathbb{E}\left\{\mathbf{X}\mathbf{X}^T\right\}$.

Conditional distributions and statistics

Definition: Conditional distribution

Assuming knowledge of random variables $X_1, \ldots, X_k$, we define the conditional distribution of $x_{k+1}, \ldots, x_N$ as:

$$f_{\mathbf{X}}\left(x_{k+1}, \ldots, x_N | x_1, \ldots, x_k\right) = \frac{f_{\mathbf{X}}\left(x_1, \ldots, x_k, \ldots, x_N\right)}{f_{X_1, \ldots, X_k}\left(x_1, \ldots, x_k\right)}. \tag{17}$$

$$\text{where} \quad f_{X_1, \ldots, X_k}\left(x_1, \ldots, x_k\right) = \int_{\mathcal{D}_{k+1}} \cdots \int_{\mathcal{D}_N} f_{\mathbf{X}}\left(x_1, \ldots, x_N\right) dx_{k+1} \cdots dx_N. \tag{18}$$

and $\mathcal{D}_i$ is the domain of support of X_i.

Definition: Conditional expectation

Assuming knowledge of random variables $X_1, \ldots, X_k$, we define the conditional expectation of $g\left(x_1, \ldots, x_N\right)$ as

$$\begin{aligned} &\mathbb{E}_{|X_1, \ldots, X_k}\left\{g\left(x_1, \ldots, x_N\right)\right\} = \\ &\int_{\mathcal{D}_{k+1}} \cdots \int_{\mathcal{D}_N} g\left(x_1, \ldots, x_N\right) f_{\mathbf{X}}\left(x_{k+1}, \ldots, x_N | x_1, \ldots, x_k\right) dx_{k+1} \cdots dx_N. \end{aligned} \tag{19}$$

The Multivariate Gaussian Distribution and its properties

Definition: Jointly Gaussian random variables & Gaussian vectors

A collection $X_i, i \in I$ of random variables is said to have a joint Gaussian distribution if every linear combination of X_is is a Gaussian random variable.

A random vector $\mathbf{x}$ is said to be a Gaussian random vector if its elements are jointly Gaussian distributed.

Definition: The multivariate Gaussian distribution

Let $\mathbf{X}$ be a Gaussian random vector, having a mean value $\boldsymbol{\mu}$ and a non-singular covariance matrix $\mathbf{K}$. The joint distribution function of $\mathbf{X}$ is then defined as

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2} |\mathbf{K}|^{1/2}} \exp\left(-\frac{(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{K}^{-1} (\mathbf{x}-\boldsymbol{\mu})}{2}\right). \tag{20}$$

Definitions of joint and conditional differential entropy

Definition: Differential entropy

Definition: Let $X_1, \dots, X_n$ be a set of n random variables defined on a support set $\mathcal{D}_{\mathbf{X}} \subseteq \mathbb{R}^n$. We define the differential entropy of $X_1, \dots, X_n$ as:

$$h(X_1, \dots, X_n) = -\int_{\mathcal{D}_{\mathbf{X}}} f_{\mathbf{X}}(x_1, \dots, x_n) \log f_{\mathbf{X}}(x_1, \dots, x_n)\, dx_1 \cdots dx_n. \tag{21}$$

Definition: Conditional differential entropy

Let X and Y be random variables having a joint distribution $f_{X,Y}(x, y)$. We define the conditional differential entropy $h(X|Y)$ as:

$$h(X|Y) = -\int_{\mathcal{D}_X} \int_{\mathcal{D}_Y} f_{X,Y}(x, y) \log f_X(x|y)\, dxdy. \tag{22}$$

where $\mathcal{D}_X, \mathcal{D}_Y$ are the domains of support of X and Y respectively.

Remark: Given that $f_{X,Y}(x, y) = f_X(x|y)\, f_Y(y)$, we can equivalently right the conditional differential entropy as:

$$h(X|Y) = h(X, Y) - h(Y). \tag{23}$$

Entropy of Gaussian distribution (in nats)

Theorem: Entropy of a multivariate Gaussian vector (Entropy in nats)

The entropy of a multivariate Gaussian vector with a mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{K}$ is written as:

$$h(\mathbf{X}) = \frac{1}{2} \ln (2\pi e)^n |\mathbf{K}| .$$

Proof: Recall that:

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2} |\mathbf{K}|^{1/2}} \exp\left(-\frac{(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{K}^{-1} (\mathbf{x}-\boldsymbol{\mu})}{2}\right), \mathbf{x} \in \mathbb{R}^n. \tag{24}$$

We then have that:

$$h(\mathbf{X}) = -\int_{\mathbb{R}^n} f_{\mathbf{X}}(\mathbf{x}) \left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{K}^{-1} (\mathbf{x}-\boldsymbol{\mu}) - \ln\left((2\pi)^{n/2} |\mathbf{K}|^{1/2}\right)\right) d\mathbf{x} \tag{25}$$

Entropy of Gaussian distribution

Proof continued

$$
\begin{aligned}
h(\mathbf{X}) &= -\int_{\mathbb{R}^n} f_{\mathbf{X}}(\mathbf{x}) \left(-\frac{1}{2} (\mathbf{x}-\boldsymbol{\mu})^T \mathbf{K}^{-1} (\mathbf{x}-\boldsymbol{\mu}) \right) d\mathbf{x} + \frac{1}{2} \ln\left((2\pi)^n |\mathbf{K}|\right) \\
&= \frac{1}{2} \sum_{j=1}^{n} \sum_{i=1}^{n} [\mathbf{K}]_{j,i} \left[\mathbf{K}^{-1}\right]_{i,j} + \frac{1}{2} \ln\left((2\pi)^n |\mathbf{K}|\right) \\
&= \frac{1}{2} \sum_{j=1}^{n} \left(\mathbf{K}\mathbf{K}^{-1}\right)_{j,j} + \ln\left((2\pi)^n |\mathbf{K}|\right) \\
&= \frac{n}{2} + \ln\left((2\pi)^n |\mathbf{K}|\right) = \frac{1}{2} \ln\left((2\pi e)^n |K|\right)
\end{aligned}
\tag{26}
$$

Relative entropy and mutual information

Relative entropy

Definition:Relative entropy

Given two probability density functions $f_X(\cdot)$ and $z_X(\cdot)$, we define the relative entropy (also called Kullback-Leibler distance) $D(f_X||z_X)$ between the two densities as:

$$D(f_X||z_X) = \int_{\mathcal{D}_X} f_X(x) \log\left(\frac{f_X(x)}{z_X(x)}\right) dx. \tag{27}$$

where $\mathcal{D}_x$ is the domain of support for $f_X(\cdot)$.

Remark: If the domain of support of $f_X(\cdot)$ is not fully contained in the domain of support of $z_X(\cdot)$, then Kullback-Leibler distance is infinite.

Mutual information

Definition: Mutual information

Given two random variables X and Y with a joint density $f_{X,Y}(x,y)$, we define the mutual information as:

$$I(X;Y) = \int_{\mathcal{D}_X} \int_{\mathcal{D}_Y} f_{X,Y}(x,y) \log \left(\frac{f_{X,Y}(x,y)}{f_X(x) f_Y(y)} \right) dxdy. \tag{28}$$

Basic properties

1. $I(X;Y) = h(X) - h(X|Y)$
2. $I(X;Y) = h(Y) - h(Y|X)$
3. $I(X;Y) = h(X) + h(Y) - h(X,Y)$
4. $I(X;Y) = D(f_{X,Y}(x,y) || f_X(x) f_Y(y))$

Further properties

Properties of differential entropy/relative entropy/ mutual information

Theorem: Positivity of relative entropy

Given two densities $f_X(\cdot)$ and $z_X(\cdot)$ it holds that $D(f_X||z_X) \geq 0$ with equality if and only if $f_X = z_X$ almost everywhere.

Proof: Using Jensen inequality and the concavity of the logarithm we have that:

$$\begin{aligned} -D(f_X||z_X) &= \int_{\mathcal{D}_X} f_X(x) \log\left(\frac{z_X(x)}{f_X(x)}\right) dx \leq \log\left(\int_{\mathcal{D}_X} f_X(x) \frac{z_X(x)}{f_X(x)} dx\right) \\ &= \log\left(\int_{\mathcal{D}_X} \log z_X(x)\, dx\right) \leq \log 1 = 0. \end{aligned} \tag{29}$$

where $\mathcal{D}_X$ is the domain of support of $f_X(\cdot)$. Moreover, we note that equality can only be satisfied if the Jensen inequality is satisfied with equality. This only occurs if $f(x) = g(x)$ almost everywhere.

What does the above theorem imply?

Consequences of the theorem

- $I(X;Y) \geq 0$ where equality holds if and only if X ad Y are independent.
- $h(X|Y) \leq h(X)$ where equality holds if and only if X and Y are independent.

Transformations of random variables

Translation and scaling of random variables

Theorem: Differential entropy and invariance to translations

$$h(X + c) = h(X) \tag{30}$$

Proof: This result is a direct consequence of the definition of differential entropy.

Theorem: Differential entropy and scaling

$$h(aX) = h(X) + \log|a| \tag{31}$$

Proof: Use the fact that $f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y}{a}\right)$, as well as the definition of differential entropy.

Remark: As $|a|$ increases, the uncertainty increases.

Gaussian random vectors and differential entropy

Maximizing the differential entropy

Theorem: Upper bound on the differential entropy

Let $\mathbf{X} \in \mathbb{R}^n$ be a zero mean random vector having a covariance matrix $\mathbf{K} = \mathbb{E}\left\{\mathbf{X}\mathbf{X}^T\right\}$. In then holds that $h(\mathbf{X}) \leq \frac{1}{2}\log\left((2\pi e)^n |K|\right)$ where equality holds if and only if $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$.

Proof: Let us start by using notation $\phi_{\mathbf{K}}(\cdot)$ for the probability density function of a vector distributed according to $\mathcal{N}(\mathbf{0}, \mathbf{K})$. It is easy then to see that the logarithm of $\phi_K(\mathbf{x})$ is a quadratic form and that $\int x_i x_j \phi_{\mathbf{K}}(\mathbf{x})\, d\mathbf{x} = [\mathbf{K}]_{i,j}$. We now use the non negativity property for the differential entropy between any distribution $g(\cdot)$ having a covariance matrix $\mathbf{K}$ and a Gaussian distribution having the same covariance matrix, and obtain that:

$$\begin{aligned} 0 \leq D(g||\phi_{\mathbf{K}}) &= \int_{\mathbb{R}^n} g(\mathbf{x}) \log\left(g(\mathbf{x}) / \phi_{\mathbf{K}}(\mathbf{x})\right) d\mathbf{x} \\ &= -h(g) - \int_{\mathbb{R}^n} g(\mathbf{x}) \log \phi_K(\mathbf{x})\, d\mathbf{x} \\ &= -h(g) - \int \phi_K(\mathbf{x}) \log \phi_K(\mathbf{x}) = -h(g) + h(\phi_{\mathbf{K}}) \end{aligned} \tag{32}$$

Note: In our proof we have used the fact that term $\int g \log \phi_K = \int \phi_K \log \phi_K$, since ϕ_K is a quadratic form, it is solely determined by the covariance matrix.