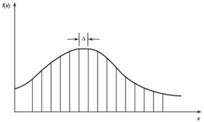
Introduction to Machine Learning

Information Theory
Differential Entropy



Learning goals

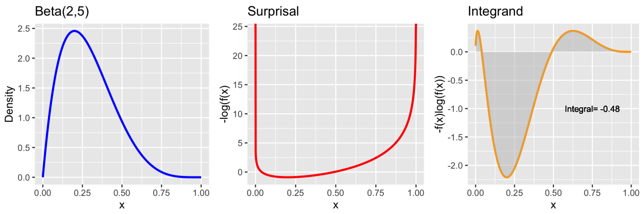
- Know that the entropy expresses expected information for continuous RVs
- Know the basic properties of the differential entropy

DIFFERENTIAL ENTROPY

- For a continuous random variable X with density function $f(x)$ and support $\mathcal{X}$, the analogue of entropy is **differential entropy**:

$$h(X) := h(f) := -\mathbb{E}[\log(f(x))] = -\int_{\mathcal{X}} f(x)\log(f(x))dx$$

- The base of the log is again somewhat arbitrary, and we could either use 2 (and measure in bits) or e (to measure in nats).
- The integral above does not necessarily exist for all densities.
- Differential entropy lacks the non-negativeness of discrete entropy: $h(X) < 0$ is possible as $f(x) > 1$ is possible:



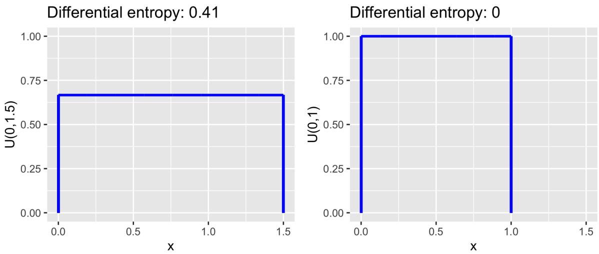
The diffent. is given by the integral: $h(X) = -0.48$.

DIFF. ENTROPY OF UNIFORM DISTRIBUTION

Let X be a uniform random variable on $[0, a]$.

$$\begin{aligned} h(X) &= -\int_0^a f(x)\log(f(x))dx \\ &= -\int_0^a \frac{1}{a}\log\left(\frac{1}{a}\right)dx = \log(a) \end{aligned}$$

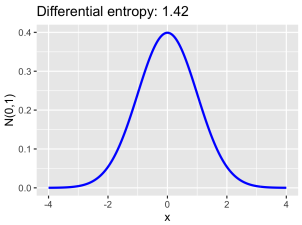
- For $a < 1$, $h(X) < 0$. a

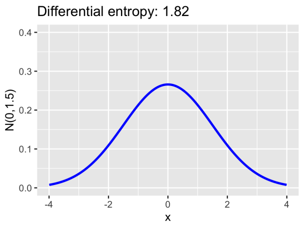


DIFF. ENTROPY OF GAUSSIAN

Let $X \sim \mathcal{N}(\mu, \sigma^2)$ and let us measure in nats:

$$
\begin{aligned}
h(X) &= -\int_{\mathbb{R}} f(x) \log(f(x)) dx = -\int_{\mathbb{R}} f(x) \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right) dx \\
&= -\int_{\mathbb{R}} f(x) \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) dx + \int_{\mathbb{R}} f(x) \frac{(x-\mu)^2}{2\sigma^2} dx \\
&= -\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) \underbrace{\int_{\mathbb{R}} f(x) dx}_{=1} + \frac{1}{2\sigma^2} \underbrace{\int_{\mathbb{R}} f(x)(x-\mu)^2 \, dx}_{=:\sigma^2} \\
&= \frac{1}{2} \log\left(2\pi\sigma^2\right) + \frac{1}{2} = \log(\sigma\sqrt{2\pi e})
\end{aligned}
$$

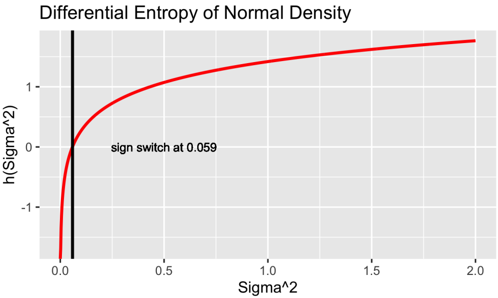




DIFF. ENTROPY OF GAUSSIAN

$$h(X) = -\int_{\mathbb{R}} f(x)\log(f(x))dx = \log(\sigma\sqrt{2\pi e})$$

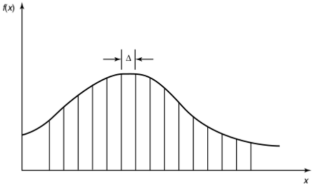
- $h(X)$ is not a function of μ (see translation invariance later).
- As σ^2 increases, the differential entropy also increases.
- For $\sigma^2 < \frac{1}{2\pi e} \approx 0.059$, it is negative.



DIFF. ENTROPY VS. DISCRETE

It is not so simple as to characterize $h(X)$ as a straightforward generalization of $H(X)$ of a limiting process. Consider the quantized random variable X^{Δ}, which is defined by

$$X^{\Delta} = x_i \qquad \text{if} \qquad i\Delta \leq X < (i+1)\Delta$$



If the density $f(x)$ of the random variable X is Riemann-integrable, then

$$H(X^{\Delta}) + \log(\Delta) \rightarrow h(X) \text{ as } \Delta \rightarrow 0.$$

Thus, the entropy of an n-bit quantization of a continuous random variable X is approximately $h(X) + n$.

JOINT DIFFERENTIAL ENTROPY

- For a continuous random vector X with density function $f(x)$ and support $\mathcal{X}$, differential entropy is also defined as:

$$h(X) = h(X_1, \ldots, X_n) = h(f) = -\int_{\mathcal{X}} f(x) \log(f(x)) dx$$

- Hence this also defines the joint differential entropy for a set of continuous RVs.

Entropy of a multivariate normal distribution: If $X \sim N(\mu, \Sigma)$ is multivariate Gaussian, then

$$h(X) = \frac{1}{2} \log(2\pi e)^n |\Sigma| \qquad \text{(nats)}$$

PROPERTIES OF DIFFERENTIAL ENTROPY

1. $h(f)$ can be negative.
2. $h(f)$ is additive for independent RVs.
3. $h(f)$ is maximized by the multivariate normal, if we restrict to all distributions with the same (co)variance, so $h(X) \leq \frac{1}{2}\log(2\pi e)^n|\Sigma|$.
4. $h(f)$ is maximized by the continuous uniform distribution for a random variable with a fixed range.
5. Translation-invariant, $h(X + a) = h(X)$.
6. $h(aX) = h(X) + \log|a|$.
7. $h(AX) = h(X) + \log|A|$ for random vectors and matrix A.

3) and 4) are slightly involved to prove, while the other properties are relatively straightforward to show