Information Theory

Stephen Huan[1]

[1]Georgia Institute of Technology

January 20, 2023

Table of Contents

Entropy, Intuitively

- Let $h(x)$ be some measure of the "uncertainty" or "surprise" of event x. What are its properties?
- If $p(x) = 1$, then there's no uncertainty — $h(x) = 0$
- If $p(x) < p(y)$, then $h(x) > h(y)$ — if x is rarer than y, then it is more surprising, so it has higher information
- Finally, if x and y are independent, their information should just add — $h(xy) = h(x) + h(y)$
- $h(x)$ looks like $\log x$

Proof of the Form of $h(x)$

$$
\begin{aligned}
h(xy) &= h(x) + h(y) && \text{definition} \\
yh'(xy) &= h'(x) && \text{taking } \frac{\partial}{\partial x} \\
xyh''(xy) + h'(xy) &= 0 && \text{taking } \frac{\partial}{\partial y} \\
uh''(u) + h'(u) &= 0 && \text{letting } u = xy \\
uf'(u) + f(u) &= 0 && \text{letting } f(u) = h'(u)
\end{aligned}
$$

Solving the Differential Equation

$$
\begin{aligned}
u\frac{du}{df} + f &= 0 \\
u\, df &= -f\, du \\
\frac{1}{f} df &= -\frac{1}{u} du \\
\frac{1}{f}\frac{df}{du} &= -\frac{1}{u} \\
\int \frac{1}{f}\left(\frac{df}{du}\right) du &= -\int \frac{1}{u}\, du \\
\ln|f| &= -\ln|u| + k \\
f(u) &= k\frac{1}{u}
\end{aligned}
$$

Finishing up $h(x)$

$$f(x) = k\frac{1}{x} = h'(x)$$

$$h(x) = k\int \frac{1}{x}\,dx = k\ln x + C$$

but we know $h(xy) = h(x) + h(y)$, so

$$C = 0$$

and $x < y$ implies $h(x) > h(y)$ so

$$k < 0$$

$$\boxed{h(x) = -k\ln x}$$

- k corresponds to the *base* of the logarithm
- We'll take base e (natural log) for simplicity
- The units are known as "nats", if base 2 is used, "bits"

Expected value of $h(x)$ — Entropy!

- Suppose X is a random variable
- The *expected* amount of information it takes to encode X will be simply the expected value of $h(x)$:

$$\boxed{\mathrm{H}[X] = \mathrm{E}[-\ln(p(X))]}$$
$$= -\sum_{x \in X} p(x) \ln(p(x))$$

- $\mathrm{H}[X]$ is known as the *entropy* of X. This seems like a natural way to encode the "uncertainty" of a random variable, but how precise is this definition?

Heuristical Justification of Entropy

- The derivation provided earlier seems a bit . . . handwavy
- Luckily, we can sometimes ignore *how* we got to something and simply *prove* that it makes sense
  - Induction
  - Solving differential and recurrence relations by guessing
  - etc.
- Valid strategy in math when intuition hard to justify *a priori*

Justification of Expectation

- Expectation: $\mathrm{E}[X] = \sum_{x \in X} x\, p(x)$
  - Frequentist justification: discrete random variable with values $x_1 \ldots x_n$, probabilities $p_1 \ldots p_n$.
  - Run $\lim_{N \to \infty}$ trials, see $p_i N$ of x_i
  - Average value $\frac{1}{N} \sum_{i=1}^{n} x_i (p_i N)$
  - $\sum_{i=1}^{n} x_i p_i = \sum_{x} x\, p(x)$
- Doesn't really work with continuous!
- Made rigorous with *strong law of large numbers*

Strong Justification of Entropy

- Entropy H[X] *lower bound* on the (average) number of bits to encode a random variable X [*noiseless coding theorem*]
- This is quite a strong claim!
- Proved by Shannon, along with many other information-theoretic concepts

Table of Contents

Entropy Coding Example

- Discrete uniform distribution on 8 states
- $f(x) = \frac{1}{8}$, $\mathrm{H}[X] = -8(\frac{1}{8}\log_2\frac{1}{8}) = 3$
- Non-uniform distribution $\left\{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right\}$

$$\mathrm{H}[X] = -\sum_i p(x_i)\log_2 p(x_i)$$

$$= -\left(\frac{1}{2}\log\frac{1}{2} + \frac{1}{4}\log\frac{1}{4} + \frac{1}{8}\log\frac{1}{8} + \frac{1}{16}\log\frac{1}{16} + \frac{1}{64}\log\frac{1}{64} + \frac{1}{64}\log\frac{1}{64} + \frac{1}{64}\log\frac{1}{64} + \frac{1}{64}\log\frac{1}{64}\right)$$

$$= 2$$

Better than Uniform with Optimal Coding

- $\left\{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right\}$
- Use coding with length "proportional" to probability
- 0, 10, 110, 1110, 111100, 111101, 111110, 111111
- Valid code: no code is a prefix of any other (need to be able to uniquely distinguish in a concatenated sequence)
- Length of each code is precisely $\log_2 p(x_i)$
- Expected length also 2 bits!
- In general, *Huffman coding* to generate optimal codes

Table of Contents

Statistical Mechanics Perspective of Entropy

- N objects placed into bins, the ith bin can have n_i objects ($\sum_i n_i = N$). I think physicists call these "microstates".
- The number of ways W to do this is just combinatorial
- Place your objects in a line. Take the first n_1 as bin 1, next n_2 as bin 2, and so on. $N!$ ways to order N objects, but we don't care about order within a bin so divide by each $n_i!$
- $W = \frac{N!}{\prod_i n_i!}$. I think physicists call this the "macrostate".
- Amount of entropy $H = \frac{1}{N} W$ (normalized uncertainty)
- Take $\lim_{N\to\infty} H$

Statistical Mechanics, Continued

$$W = \frac{N!}{\prod_i n_i!} \qquad \text{definition of } W$$

$$H = \frac{1}{N} \ln W \qquad \text{definition of } H$$

$$= \frac{1}{N}[\ln(N!) - \sum_{i=1}^{N} \ln(n_i!)] \qquad \text{expanding}$$

From Stirling's approximation $n! \sim n \ln n - n$

$$= \frac{1}{N}[N \ln N - N - \sum_{i=1}^{N} (n_i \ln n_i - n_i)]$$

Statistical Mechanics, Continued

$$H = \frac{1}{N}[N \ln N - N - \sum_{i=1}^{N}(n_i \ln n_i - n_i)]$$

$$= \frac{1}{N}[N \ln N - \sum_{i=1}^{N}(n_i \ln n_i)] \qquad \text{from } \sum_i n_i = N$$

$$= -\sum_{i=1}^{N} \frac{1}{N}(n_i \ln n_i - n_i \ln N) \qquad \text{from } \sum_i n_i = N$$

$$= -\sum_{i=1}^{N}(\frac{n_i}{N}) \ln\left(\frac{n_i}{N}\right)$$

$$= -\sum_{i=1}^{N} p_i \ln p_i$$

Statistical Mechanics, Conclusion

$$H = -\sum_{i=1}^{N} p_i \ln p_i$$
$$= \mathrm{H}[X]$$

- Thermodynamic entropy $S = k \ln W$ equivalent to information-theoretic entropy!

Table of Contents

Differential Entropy

- Want to generalize entropy to continuous random variable
- Recall: $\mathrm{H}[X] = \mathrm{E}[-\ln(p(X))] = -\sum_{x \in X} p(X) \ln p(X)$
- Why not $\mathrm{H}[X] = \mathrm{E}[-\ln(p(X))] = -\int p(X) \ln p(X)$
- Yes, this quantity is known as *differential entropy*
- But there is a very important *caveat* we're missing by being cavalier about replacing sums with integrals
- We'll have to actually work through the derivation!

Differential Entropy Derivation

- Divide X into bins of width Δ
- Need to assign every element that falls into bin i to x_i
- Find x_i such that $p(x_i)$ equals probability of bin i
- Mean value theorem guarantees there exists x_i such that

$$\int_{i\Delta}^{(i+1)\Delta} p(x)\,dx = p(x_i)\Delta$$

- Given these x_i, have discrete distribution with values x_i and corresponding probabilities $p(x_i)\Delta$
- Compute entropy of this discrete distribution as $\lim_{\Delta \to 0}$

Differential Entropy, Continued

$$H_\Delta = -\sum_i p(x_i)\Delta \ln(p(x_i)\Delta)$$

expanding from ln, using $\sum_i p(x_i) = 1$

$$= -\sum_i p(x_i)\Delta \ln p(x_i) - \ln \Delta$$

$$\lim_{\Delta\to 0} H_\Delta = \lim_{\Delta \to 0}\left[-\sum_i p(x_i)\ln p(x_i)\Delta - \ln\Delta\right]$$

$$\underbrace{H_\Delta}_{\text{discretized entropy}} = \underbrace{-\int p(x)\ln p(x)\,dx}_{\text{differential entropy}} + \underbrace{(-\ln\Delta)}_{\text{infinite bits}}$$

Differential Entropy, Commentary

- So, our differential entropy is the entropy of the binned discrete distribution as the discretization gets arbitrarily precise, minus infinite information
- Intuitively, this makes sense, because it takes infinite bits to specify an arbitrary real number
- The fact that we need to subtract infinite bits makes differential entropy less intuitive than its discrete counterpart, for example, it can be negative
- It still has some useful properties though — if one quantizes a continuous random variable X to n digits, the cost to encode it will be (approximately) $\mathrm{H}[X] + n$
- It's also useful to define the upcoming KL divergence, which will help quantify the difference between distributions

Table of Contents

Entropy of a Uniform Random Variable

- Now that we've defined both discrete and differential entropy, we can apply them to some simple distributions
- Discrete uniform on $0, 1, \ldots, N-1$ (location doesn't matter, entropy determined by distribution)
- Density $f(x) = \frac{1}{N}$

$$\begin{aligned} \mathrm{H}[X] &= -\sum_i p(x_i) \ln p(x_i) \\ &= -N\left(\frac{1}{N} \ln \frac{1}{N}\right) \\ &= \ln N \end{aligned}$$

- Recall N is the number of states, so this entropy is always positive ($N \geq 1 \implies \ln N \geq 0$)

Entropy of a Continuous Uniform

- Uniform on $[0, N]$, density $f(x) = \frac{1}{N}$

$$
\begin{aligned}
\mathrm{H}[X] &= -\int p(x) \ln p(x)\, dx \\
&= -\int_0^N \frac{1}{N} \ln \frac{1}{N}\, dx \\
&= \ln N
\end{aligned}
$$

- However, a continuous distribution can have $N < 1$
- So differential entropy can be negative!

Entropy of a Multivariate Gaussian

- Recall a multivariate Gaussian $X \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ has density $f(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} e^{-(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$

$$\begin{aligned}
\mathrm{H}[X] &= -\,\mathrm{E}[\ln p(X)] \\
&= -\,\mathrm{E}[\ln\left(\frac{1}{(2\pi)^{\frac{n}{2}}}\frac{1}{|\Sigma|^{\frac{1}{2}}}\right) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})] \\
&= \frac{n}{2}\ln(2\pi) + \frac{1}{2}\operatorname{logdet}\Sigma + \frac{1}{2}\,\mathrm{E}[(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})]
\end{aligned}$$

Entropy of a Multivariate Gaussian

Abusing the fact that the trace of a scalar is a scalar,

$$\mathrm{E}[(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})] = \mathrm{E}[\mathrm{trace}\left((\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)]$$

Using the cyclic property of trace,

$$= \mathrm{E}[\mathrm{trace}\left(\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top\right)]$$

Swapping expectation and trace by linearity of expectation,

$$= \mathrm{trace}(\mathrm{E}\left[\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top\right])$$

Bringing out Σ^{-1} since it is constant,

$$= \mathrm{trace}(\Sigma^{-1}\,\mathrm{E}\left[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top\right])$$

Here we recognize the covariance marix of X

$$= \mathrm{trace}(\Sigma^{-1}\Sigma) = \mathrm{trace}(I_n) = n$$

Entropy of Multivariate Gaussian

$$\begin{aligned} \mathrm{H}[X] &= \frac{1}{2}\left(n \ln(2\pi) + \operatorname{logdet}\Sigma + \mathrm{E}[(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})]\right) \\ &= \frac{1}{2}\left(n \ln(2\pi) + \operatorname{logdet}\Sigma + n\right) \end{aligned}$$

$$\boxed{\mathrm{H}[X] = \frac{n}{2}\ln\left(2\pi e|\Sigma|^{\frac{1}{n}}\right)}$$

- For the case of 1D, the entropy reduces to

$$\frac{1}{2}\ln\left(2\pi e \sigma^2\right)$$

- This will again be negative if $\sigma^2 < \frac{1}{2\pi e}$

Perplexity

- The *perplexity* of a random variable is $2^{\mathsf{H}[x]}$
- Entropy is measured in bits (base 2) here

Table of Contents

Cross-Entropy

- The *cross-entropy* between distributions p and q is the expected number of bits it takes to specify a sample from p given an (optimal) coding scheme from q
- Coding scheme: a way to encode a sequence as 1's and 0's
-

KL Divergence

- Kullback-Leibler Divergence (KL Divergence)

Conditional Entropy

Mutual Information

References

1. *Pattern Recognition and Machine Learning*, Bishop
2. *Probabilistic Machine Learning: An Introduction*, Murphy
3. ECE 587 / STA 563: Lecture 7 — Differential Entropy
4. *Near-Optimal Sensor Placements in Gaussian Processes: Theory, Efficient Algorithms and Empirical Studies*
5. Wikipedia on entropy