purpose of the lecture

to introduce

Information, and its negative Entropy

Entropy is a measure of our ignorance

Maximizing it is a wise (humble) prior!

Links from https://github.com/MPOcanes/MPO624-2020/blob/master/CALENDAR.md

- Brief glimpse of information theory: entropy or *missing information*, "one of the most fundamental discoveries of human thought", which familiar distributions (uniform, exponential, normal) maximize for specified width, mean, and variance respectively. (Codes to compute mutual information, a non-independence generalization to "correlatedness", is in libraries in Python, Matlab.

Entropy = “missing information”

- entropy H = -(information) = H(a PDF)

- Call probability density *likelihood*
- H = *- the expected value of log-likelihood*

- Units are *bits* when log is base 2: clearest
- 1 bit = the answer to one coin flip, or *one Y/N question with 50-50 prior odds*

Entropy = “missing information”

- H is *- the expected value of log-likelihood*
- Expected value E[ ] is the probability-weighted sum or integral – like moments in HW
- for a random variable X, with p(x) its PDF,

$$\mathrm{E}[X] = \sum_{i=1}^{n} x_i\, p_i = x_1 p_1 + x_2 p_2 + \cdots + x_k p_k.$$

Since the sum of all probabilities p_i is 1 ($p_1 + p_2 + \cdots + p_k = 1$), the expected value is the weighted average of the x_i values, with the p_i values being the weights.

- the mean

Entropy = "missing information"

- H is *- the expected value of log-likelihood*
- Expected value is the probability-weighted sum or integral – like moments, in HW

Mean [edit]

Main article: Mean

The first raw moment is the mean, usually denoted $\mu \equiv \mathrm{E}[X]$.

Variance [edit]

Main article: Variance

The second central moment is the variance. The positive square root of the variance is the standard deviation $\sigma \equiv \left(\mathrm{E}\left[(x-\mu)^2\right]\right)^{\frac{1}{2}}$.

Entropy = “missing information”

- H is *the expected value of log-likelihood*

Given a random variable X, with possible outcomes x_i, each with probability $P_X(x_i)$, the entropy $H(X)$ of X is as follows:

$$H(X) = -\sum_i P_X(x_i) \log_b P_X(x_i) = \sum_i P_X(x_i) I_X(x_i) = \mathrm{E}[I_X]$$

- Expected value of log-likelihood of each value of x

Example: missing information

- "A stone is in one of eight boxes."

1	2	3	4	5	6	7	8

- How much *information is missing* in this probabilistic (rather than detailed) description of reality?

- Think in bits: *what is the fewest number of yes/no questions could you ask to find it?*

Example: missing information

- "<u>A stone is in one of eight boxes</u>."

1	2	3	4	5	6	7	8

- Or just use the formula. The likelihood in each box is equal (uniform: a maximum entropy prior distribution assumption!)

Given a random variable X, with possible outcomes x_i, each with probability $P_X(x_i)$, the entropy $H(X)$ of X is as follows:

$$H(X) = -\sum_i P_X(x_i) \log_b P_X(x_i) = \sum_i P_X(x_i) I_X(x_i) = \mathrm{E}[I_X]$$

Example: missing information

- "A stone is in one of eight boxes."

1	2	3	4	5	6	7	8

$\log_2(1/8) = -3$ because $2^3 = 8$ and $2^{-3} = 1/8$

- What is the expected value among 8 equally probable instances of a constant, -(-3)?

Given a random variable X, with possible outcomes x_i, each with probability $P_X(x_i)$, the entropy $H(X)$ of X is as follows:

$$H(X) = -\sum_i P_X(x_i) \log_b P_X(x_i) = \sum_i P_X(x_i) I_X(x_i) = \mathrm{E}[I_X]$$

Example: missing information

- "A stone is in one of eight boxes."

1	2	3	4	5	6	7	8

$\log_2(1/8) = -3$ because $2^3 = 8$ and $2^{-3} = 1/8$

- What is the expected value among 8 equally probable instances of a constant, -(-3)? 3

Given a random variable X, with possible outcomes x_i, each with probability $P_X(x_i)$, the entropy $H(X)$ of X is as follows:

$$H(X) = -\sum_i P_X(x_i)\log_b P_X(x_i) = \sum_i P_X(x_i) I_X(x_i) = \mathrm{E}[I_X]$$

Works in multiple dimensions too

- http://pillowlab.princeton.edu/teaching/statneuro2018/slides/notes08_infotheory.pdf

Definition 8.2 (Conditional entropy) *The conditional entropy of a random variable is the entropy of one random variable conditioned on knowledge of another random variable, on average.*

Alternative interpretations: the average number of yes/no questions needed to identify X given knowledge of Y, on average; or How uncertain you are about X if you know Y, on average?

$$
\begin{aligned}
H(X \mid Y) &= \sum_Y P(Y)[H(P(X \mid Y))] = \sum_Y P(Y)\Big[-\sum_X P(X \mid Y)\log P(X \mid Y)\Big] \\
&= = -\sum_{X,Y} P(X,Y)\log P(X \mid Y) \\
&= -\mathbb{E}_{X,Y}[\log P(X \mid Y)] \qquad (8.2)
\end{aligned}
$$

Definition 8.3 (Joint entropy)

$$H(X,Y) = -\sum_{X,Y} P(X,Y)\log P(X,Y) = -\mathbb{E}_{X,Y}[\log P(X,Y)] \qquad (8.3)$$

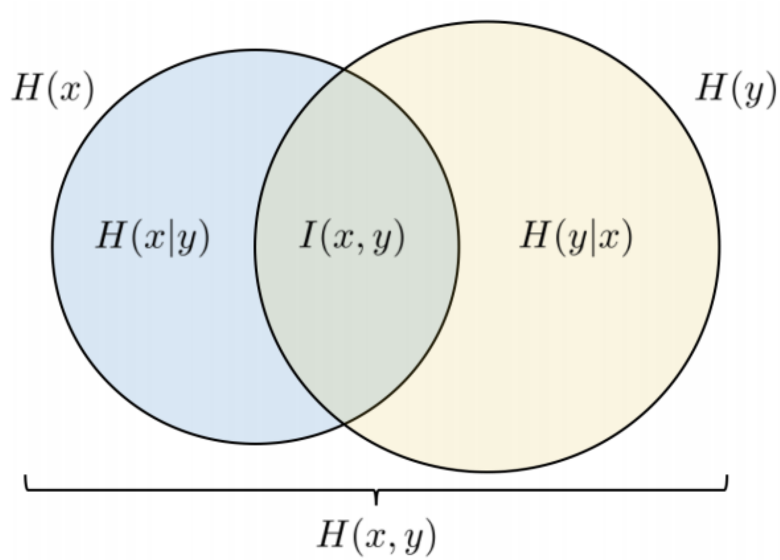
- **Bayes' rule for entropy**

$$H(X_1 \mid X_2) = H(X_2 \mid X_1) + H(X_1) - H(X_2) \tag{8.4}$$

- **Chain rule of entropies**

$$H(X_n, X_{n-1}, ...X_1) = \sum_{i=1}^{n} H(X_n \mid X_{n-1}, ...X_1) \tag{8.5}$$

It can be useful to think about these interrelated concepts with a so-called information diagram. These aid intuition, but are somewhat of a disservice to the mathematics behind them. Think of the area of each circle as the information needed to describe it, and any overlap would imply the "same information" (sorry.) describes both processes.



The entropy of X is the entire blue circle. Knowledge of Y removes the green slice. The joint entropy is the union of both circles. How do we describe their intersection, the green slice?

Works in multiple dimensions too

- http://pillowlab.princeton.edu/teaching/statneuro2018/slides/notes08_infotheory.pdf

Definition 8.4 (Mutual information) *The mutual information between two random variables is the "amount of information" describing one random variable obtained through the other (mutual dependence); alternate interpretations: how much is your uncertainty about X reduced from knowing Y, how much does X inform Y?*

$$\begin{aligned} I(X,Y) &= \sum_{X,Y} P(X,Y) \log \frac{P(X,Y)}{P(X)P(Y)} \\ &= H(X) - H(X \mid Y) \\ &= H(Y) - H(Y \mid X) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned} \tag{8.6}$$

Note that $I(X,Y) = I(Y,X) \geq 0$, with equality if and only if X and Y are independent.

Works in multiple dimensions too

- http://pillowlab.princeton.edu/teaching/statneuro2018/slides/notes08_infotheory.pdf

Lecture 8: Information Theory and Maximum Entropy 8-3

8.1.1 KL Divergence

From Bayes' rule, we can rewrite the joint distribution $P(X,Y) = P(X \mid Y)P(Y)$ and rewrite the mutual information as

$$I(X,Y) = \sum_Y P(Y) \sum_X P(X \mid Y) \log \frac{P(X \mid Y)}{P(X)} = \mathbb{E}_Y \Big[D_{KL}\big(P(X \mid Y) \,\|\, P(X)\big) \Big] \tag{8.7}$$

which we introduce as the Kullback-Leibler, or KL, divergence from $P(X)$ to $P(X \mid Y)$. Definition first, then intuition.

Enough of that. What is *maximixing entropy?*

- A principle, not unlike Occam's razor

the most fundamental discoveries of human thought. In the MaxEnt method, we maximize the (relative) entropy of a system, subject to its constraints, to infer the state of the system. Depending on the philosophical perspective adopted by the user, this can be interpreted variously as:

- inferring the **least informative state** of the system (Jaynes 1957; Shore & Johnson 1980), or
- inferring the **most probable state** of the system (Boltzmann 1877; Planck 1901).

The power of the MaxEnt method lies in its ability to infer the (probabilistic) state of a system which is under-constrained, i.e. for which no closed-form, deterministic solution can be obtained. Mathematically, it enables the user to construct a probability distribution or probability density function over the state space of the system, enabling a substantial reduction in model order. In thermodynamics – the first and still one of the foremost applications of the MaxEnt method – this enables a tremendous reduction in model order compared to the underlying molecular dynamical system, of approximately 23 orders of magnitude !

Enough of that. What is *maximixing entropy?*

- A principle, not unlike Occam's razor
- Seems to turn our ignorance into power!
  - or at least prevents preconceptions from sneaking in under that ignorance
- Assumes that *physically independent* subsystems tend to drift into *unrelated* states, perhaps

Four Important Distributions used in hypothesis testing

are all based on maximum entropy distributions!

Uniform

Exponential

Normal

t = undersampled Normal

χ^2 is undersampled squared Normal

https://en.wikipedia.org/wiki/Maximum_entropy_probability_distribution#Uniform_and_piecewise_uniform_distributions

Uniform

The uniform distribution on the interval [a,b] is the maximum entropy distribution among all continuous distributions which are supported in the interval [a, b], and thus the probability density is 0 outside of the interval. This uniform density can be related to Laplace's principle of indifference, sometimes called the principle of insufficient reason. More generally, if we're given a subdivision $a=a_0 < a_1 < ... < a_k = b$ of the

Uniform prior (50-50 odds) in Ambaum's significance testing paper

(makes his critique seem rather academic and minor):

lation or is it a fluke? In other words, we try to calculate the probability that the relation is real, given that we measured a correlation r_0. If we assume that the observed correlation is larger than the threshold correlation r_p, then we see from the Table 1 that the probability that the relation is real is $60/(60 + 5) \approx 92\%$, where we

equal prior odds →

have employed equal prior odds on the time series being related or unrelated; this probability is different from the 95% that the significance test would have us believe.

Four Important Distributions used in hypothesis testing

are all based on maximum entropy distributions!

Uniform

Exponential

Normal

t = undersampled Normal

χ^2 is undersampled squared Normal

Four Important Distributions used in hypothesis testing

Positive and specified mean: the exponential distribution [edit]

The exponential distribution, for which the density function is

$$p(x|\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0, \end{cases}$$

is the maximum entropy distribution among all continuous distributions supported in [0,∞] that have a specified mean of 1/λ.

Uniform

Exponential

<u>Physical example</u>: the potential energy of a gas atmosphere is a fraction R/Cp of its internal energy (constant, for a given T). The mass therefore exponentially decays with height in a maximum-entropy configuration.

Four Important Distributions used in hypothesis testing

are all based on maximum entropy distributions!

Uniform

Exponential

Normal

t = undersampled Normal

χ^2 is undersampled squared Normal

Four Important Distributions used in hypothesis testing are all based on maximum entropy

Specified variance: the normal distribution [edit]

The normal distribution $N(\mu,\sigma^2)$, for which the density function is

$$p(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

has maximum entropy among all real-valued distributions supported on $(-\infty,\infty)$ with a specified variance σ^2 (a particular moment). Therefore, the assumption of normality imposes the minimal prior structural constraint beyond this moment. (See the differential entropy article for a derivation.)

Normal

Physical example: Kinetic energy in a gas is the given variance (energy). Velocity is therefore Normal (Boltzmann) distributed.

Four Important Distributions

used in hypothesis testing

(lectures 22-23 at https://www.ldeo.columbia.edu/users/menke/edawm/eda_lectures/)