Lecture 11: Selecting prior information

Professor Ilias Bilionis

The principle of insufficient reason

Example: Random variable with two possible values

- You have a random variable X with two possible values, say 0 and 1.
- If that's all you know, what probability do you assign to each value?

$$p(X=0) = p(X=1) = \frac{1}{2}$$

Example: Random variable with six possible values

- You have a random variable X with six possible values, say 1, 2, 3, 4, 5, 6.

- If that's all you know, what probability do you assign to each value?

$$p(X=1) = p(X=2) = \cdots = p(X=6) = \frac{1}{6}$$

Laplace considered this obvious

"The theory of chance consists in reducing all the events of the same kind to a certain number of cases equally possible, that is to say, to such as we may be equally undecided about in regard to their existence, and in determining the number of cases favorable to the event whose probability is sought. The ratio of this number to that of all the cases possible is the measure of this probability, which is thus simply a fraction whose numerator is the number of favorable cases and whose denominator is the number of all the cases possible…"
—Pierre-Simon Laplace

https://en.wikipedia.org/wiki/Pierre-Simon_Laplace#/media/File:Laplace,_Pierre-Simon,_marquis_de.jpg

Principle of insufficient reason

- Let X be a discrete random variable taking N different values $x_1, \ldots, x_N$.
- If that all we know, then the principle of insufficient reason states that we should assign:

$$p(X = x_1) = \cdots = p(X = x_N) = \frac{1}{N}$$

Generalizations of the principle of insufficient reason

Equivalent states of knowledge should be assigned equivalent epistemic probabilities
—E. T. Jaynes

https://en.wikipedia.org/wiki/Edwin_Thompson_Jaynes#/media/File:ETJaynes1.jpg

Principle of transformation groups (advanced)

Principle of maximum entropy

Lecture 11: Selecting prior information

Professor Ilias Bilionis

Information entropy

Prequel to the principle of maximum entropy

- You have a discrete random variable X.
- You know what values it takes, say $x_1, \ldots, x_N$.
- You also have some information about it, e.g., the expectation of X is 0.5, the variance 0.1, etc.
- What probability distribution do you assign to X?

Prequel to the principle of maximum entropy

https://en.wikipedia.org/wiki/Edwin_Thompson_Jaynes#/media/File:ETJaynes1.jpg

The knowledge of average values does give a reason for preferring some possibilities to others, but we would like [...] to assign a probability distribution which is as uniform as it can be while agreeing with the available information.
—E. T. Jaynes

The uniform is the most “uncertain” distribution.

We need to assign the distribution that has the maximum uncertainty while being consistent with the data.

PREDICTIVE
SCIENCE LABORATORY

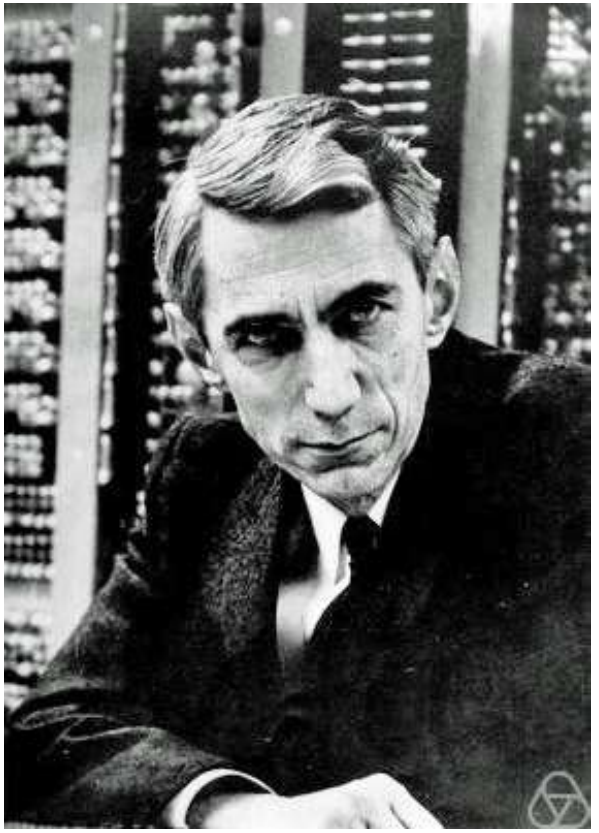
Measure of uncertainty

- You can think of the probability mass function of X as a vector $p = (p_1, \ldots, p_N)$.
- We are looking for a function $\mathbb{H}(p_1, \ldots, p_N)$ that tells how much uncertainty there is in this probability distribution.
- In 1948, Claude Shannon posed and answer this problem in the paper “A Mathematical Theory of Communication.”
- The function he came up with is called “information entropy.”

What did Shannon do?

- He assumed that $\mathbb{H}(p_1, \ldots, p_N)$ is just a real number.
- He posed some obvious axioms for $\mathbb{H}(p_1, \ldots, p_N)$, e.g., it should be continuous, it should be maximized when given the uniform distribution.
- Then he did a little bit of math and proved that:

$$\mathbb{H}(p_1, \ldots, p_N) = -\sum_{i=1}^{N} p_i \log p_i \quad \text{(information entropy)}$$



https://en.wikipedia.org/wiki/Claude_Shannon#/media/File:ClaudeShannon_MFO3807.jpg

Notational convention for information entropy

X takes values in $\{x_1, x_2, \dots\}$

$$\mathrm{H}[p(X)] := -\sum_x p(x) \log p(x) = -\mathbb{E}[\log p(X)].$$

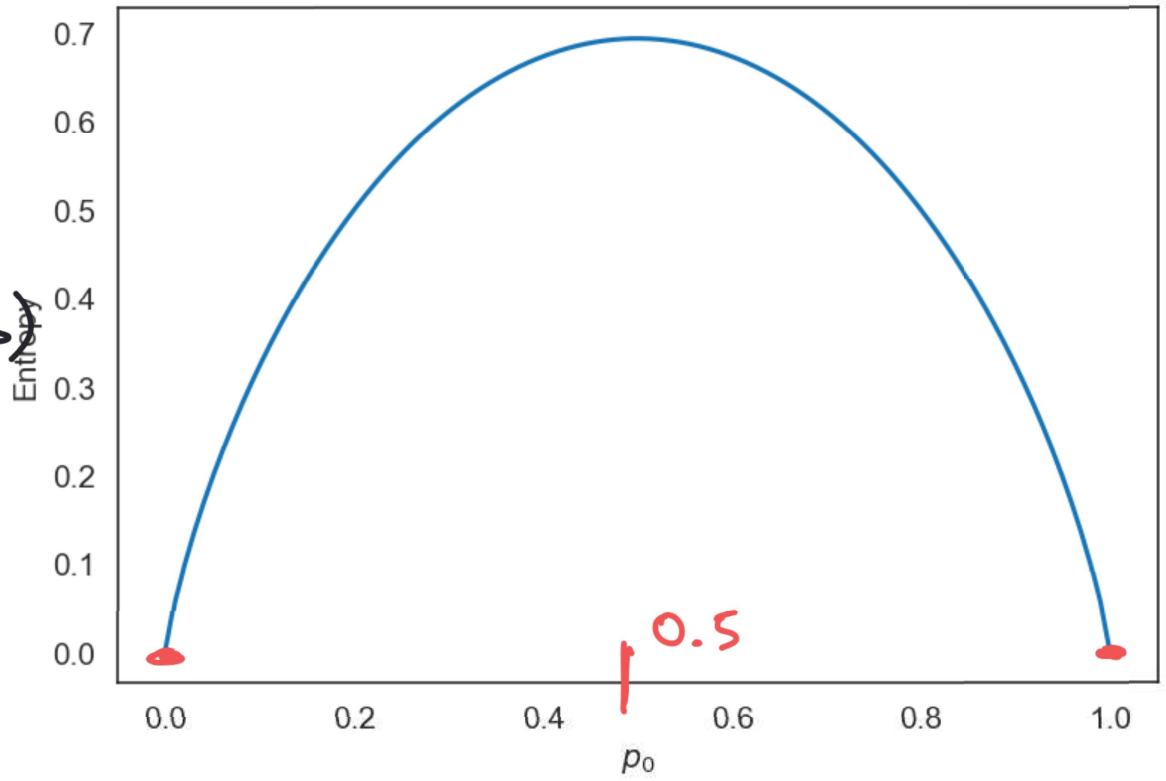
Information entropy of a distribution with two outcomes

$$X = \begin{cases} 0, & p_0 \\ 1, & p_1 = 1 - p_0 \end{cases}$$

$$H[p(X)] = -\sum_x p(x) \log p(x)$$

$$= -p_0 \log_2 p_0 - p_1 \log_2 p_1$$

$$= -p_0 \log_2 p_0 - (1 - p_0) \log_2 (1 - p_0)$$



Lecture 11: Selecting prior information

Professor Ilias Bilionis

The principle of maximum entropy for discrete random variables

Prequel to the principle of maximum entropy

- You have a discrete random variable X.
- You know what values it takes, say $x_1, \ldots, x_N$.
- You also have some **testable information** about it.
- The principle of maximum entropy states that we should assign to X the probability distribution that maximizes the entropy subject to the constraints imposed by the testable information.

Mathematical definition of testable information

$$\mathbb{F}[f_k(X)] = f_k$$

f_k: known function

f_k: known value

$k = 1, \ldots, K$

Is this definition broad enough?

I = “the expected value of X is μ”

$$\mathbb{E}[X] = \mu$$

$$K=1, \quad f_1(x) = x, \quad F_1 = \mu.$$

Is this definition broad enough?

I = "the expected value of X is μ and the variance of X is σ^2"

$\mathbb{E}[X] = \mu$; $\mathbb{V}[X] = \sigma^2$

$\sigma^2 = \mathbb{V}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \mathbb{E}[X^2] - \mu^2$

$\Rightarrow$ $\mathbb{E}[X^2] = \sigma^2 + \mu^2$

Mathematical statement of the principle of maximum entropy

You should assign to X the pmf $p(x)$ that

$$\max \mathbb{H}[p(X)] = \max -\sum_{i=1}^{N} p(x_i) \log p(x_i)$$

subject to

$$\mathbb{E}[f_k(X)] = f_k, \quad \text{for } k = 1, \ldots, K$$

$$\| \sum_{i=1}^{N} f_k(x_i) p(x_i)$$

and

$$\sum_{i=1}^{N} p(x_i) = 1$$

The general solution to the maximum entropy problem

$$p(X = x_i) = \frac{1}{Z} \exp\left\{ \sum_{k=1}^{K} \lambda_k f_k(x_i) \right\}$$

need to be det.

$$Z = \sum_{i=1}^{N} \exp\left\{ \sum_{k=1}^{K} \lambda_k f_k(x_i) \right\}$$

$$\boxed{F_k = \frac{\partial Z}{\partial \lambda_k}}$$

Example 1

- X takes N different values (no other constraints)

$$p(X = x_i) = \frac{1}{N}$$

Example 2

- X takes two values 0 and 1.
- $\mathbb{E}[X] = \theta$.

$X \sim \text{Bernoulli}(\theta)$

Example 3

- X takes values $0,1,2,\ldots,N$.
- $\mathbb{E}[X] = \mu$.
- X is the number of successful trials in N sequential experiments (potentially correlated)/

$$X \sim B\left(N, \frac{\mu}{N}\right)$$

Example 4

- X takes values $0,1,2,\ldots$
- $\mathbb{E}[X] = \mu$.
- X is the number of successful trials in an infinite number of sequential experiments (potentially correlated).

$$X \sim \text{Poisson}(\mu)$$

Lecture 11: Selecting prior information

Professor Ilias Bilionis

The principle of maximum entropy for continuous random variables

The naïve extension of information entropy to continuous distributions and why it doesn't always work

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

$$Y = T(X)$$

$$H[p(Y)] \neq H[p(X)]$$

The correct information entropy for continuous distributions

$$H[p(x)] := -\int p(x) \log \frac{p(x)}{q(x)} \, dx$$

density of maximal uncertainty

Mathematical statement of the principle of maximum entropy for continuous distributions

$$\max H[p(X)] = \max -\int p(x) \log \frac{p(x)}{q(x)} dx$$

subject to

$$\int p(x) dx = 1$$

and

$$\mathbb{E}[f_k(X)] = F_k, \quad k = 1, \ldots, K$$

The general solution to the maximum entropy problem for continuous distributions

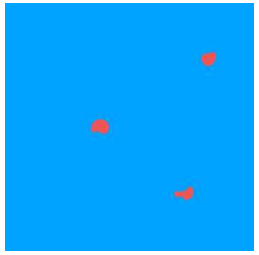
$$p(x) = \frac{q(x)}{Z} \exp\left\{ \sum_{k=1}^{K} \lambda_k f_k(x) \right\}$$

$$Z = \int q(x) \exp\left\{ \sum_{k=1}^{K} \lambda_k f_k(x) \right\} dx$$

$$\boxed{F_k = \frac{\partial Z}{\partial \lambda_k}, \quad k = 1, \ldots, K}$$

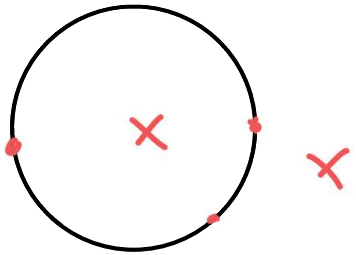
A few comments on the maximal uncertainty density $q(x)$

Example: Particle in a box:



$$q(x) = \begin{cases} 1, & \text{in box} \\ 0, & \text{otherwise} \end{cases} \propto \mathbb{1}_{\text{box}}(x)$$

Example: Particle restricted on circular guide:



$$q(x) \propto \mathbb{1}_{\text{guide}}(x)$$

Mathematical theory for finding the maximal uncertainty density $q(x)$

- Principle of transformation groups.
- Theory of Haar measures.

Example 1

- X takes values in $[a, b]$
- $q(x) = 1$

$X \sim U([a,b])$

Example 2

- X takes values in $\mathbb{R}$
- $q(x) = 1$
- $\mathbb{E}[X] = \mu$
- $\mathbb{V}[X] = \sigma^2$

$$X \sim N(\mu, \sigma^2)$$

Example 3

- X takes values in $[0,\infty)$
- $q(x) = 1$
- $\mathbb{E}[X] = \mu$

$X \sim \text{Exp}\left(\frac{1}{\mu}\right)$

A final note on the use of maximum entropy for finding priors

- The principle of maximum entropy is a great tool for assigning “objective” priors.
- However:
  - The cost of “theorizing” and “computing” for finding the ideal distributions should be taken into account. This was called “type-2 reasoning” by I. J. Good.
  - Sometimes, you have subjective information. You should not be afraid to use it.