Lecture 11: Selecting prior information

Professor Ilias Bilionis

Information entropy

PREDICTIVE
SCIENCE LABORATORY

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?

PREDICTIVE
SCIENCE LABORATORY

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

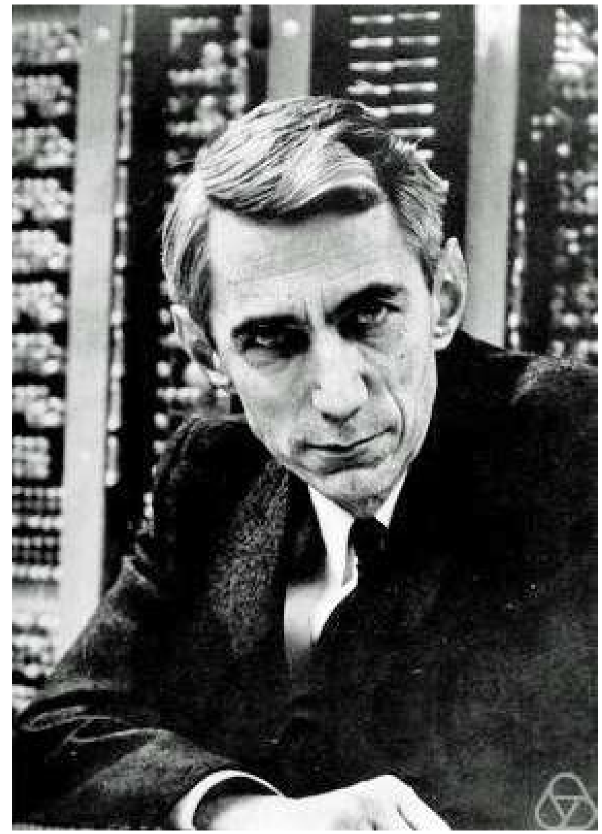
- You can think of the probability mass function of X as a vector $p = (p_1, \ldots, p_N)$.
- We are looking for a function $\mathbb{H}(p_1, \ldots, p_N)$ that tells how much uncertainty there is in this probability distribution.
- In 1948, Claude Shannon posed and answer this problem in the paper "A 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:



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

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

Notational convention for information entropy

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

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

Information entropy of a distribution with two outcomes

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

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

$$= -p_0 \log p_0 - p_1 \log p_1$$

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