TAM 598 Lecture 11:

Bayesian Inference - Selecting Your Prior

Announcements:

- HW 3 covers lectures 8-12; due on Mar 12

I. Selecting Prior Information

prior should reflect your beliefs about the variable before you see any data

approaches to selection:

(1) principle of insufficient reason/ principle of indifference

(2) principle of maximum entropy

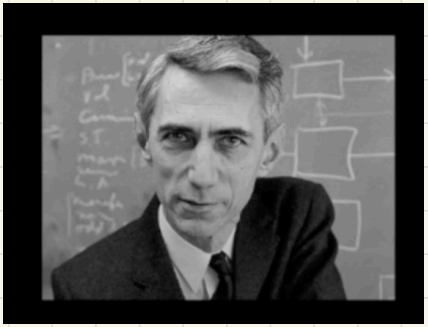
(3) principle of transformation groups

(1) principle of insufficient reason/ principle of indifference (Laplace)

> 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*

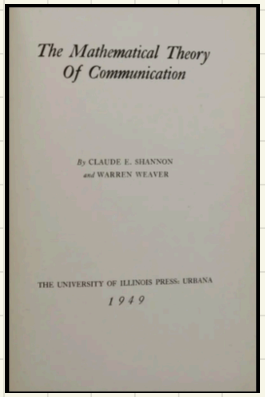
(2) principle of maximum entropy - pick most "uncertain" distribution

If you have some information about R.V. X, such as its expectation $\mathbb{E}[X]$ or variance $\mathbb{V}[X]$.

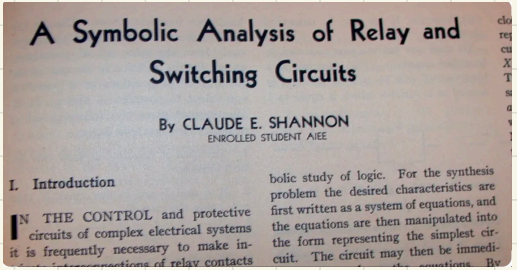


Shannon, Bell Labs

Shannon Information Content



Consider a discrete random variable X. What is the information content associated with a particular measurement $X = x$?


A Symbolic Analysis of Relay and Switching Circuits

By CLAUDE E. SHANNON
ENROLLED STUDENT AIEE

I. Introduction

IN THE CONTROL and protective circuits of complex electrical systems it is frequently necessary to make in-

bolic study of logic. For the synthesis problem the desired characteristics are first written as a system of equations, and the equations are then manipulated into the form representing the simplest circuit. The circuit may then be immedi-

Why is this a good proposal for information content?

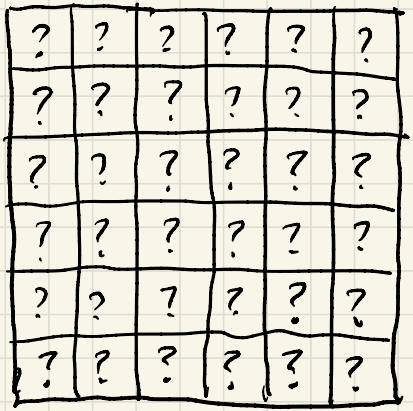
(1) Deterministic outcomes contain no information.

(2) Information content increases with decreasing probability.

Why is this a good proposal for information content?

(3) Information content is additive for independent RVs

Example Submarine (D. MacKay)



Start: $p(X = x) = \frac{1}{36}$ for $x = 1, 2, \ldots, 36$

Now we can think of the expected information content associated with a given random variable

$$H[X] = \sum_x p(X=x)\, h(X=x)$$

Example: Information entropy of a binary distribution.

We know that there are two possible outcomes, 0 and 1.
We know nothing else. What $P(x)$ should we choose?

Given some known information about a distribution, how do we select $p(x)$?

Given some known information about a distribution, how do we select $p(x)$?

Our problem now is to maximize $H[p(x)] = -\sum_i p_i \log p_i$

Subject to constraints:

$$\sum_i p_i = 1 \qquad \text{normalization}$$

$$\mathbb{E}[f_k(x)] = F_k \qquad \text{information} \quad k = 1 \ldots K$$

The general solution (Karush-Kuhn-Tucker) conditions

Example: Brandeis dice problem (E.T. Jaynes, 1962)

You toss a die N times. You measure $\mathbb{E}[x] = 4.5$ instead of the "fair dice" expected value of 3.5. Given this info and nothing else, what $p(x)$ should we assign to the next toss?

To find λ we need to solve:

Examples: discrete maximum entropy distributions:

- X takes N different values $\Rightarrow$ Categorical $\left(\frac{1}{N}, \frac{1}{N}, \ldots, \frac{1}{N}\right)$
- X is 0 with prob $(1-\theta)$, is 1 with prob θ
  $\Rightarrow$ Bernoulli (θ)
- X takes values $0, 1, \ldots, n$ with known expectation $\mathbb{E}[X] = \theta n$
  $\Rightarrow$ Binomial $B(\theta, n)$
- X takes values $0, 1, 2, \ldots, \infty$ with known $\mathbb{E}[X] = \lambda$
  $\Rightarrow$ Poisson (λ)
- thermodynamic ensembles: max entropy of

  NVT canonical – states of a quantum system w/ known expected energy

  μVT grand canonical – states of a quant. system w/ varying # of particles, known expected #, Energy

for continuous distributions, to a large extent:

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

- X lies in $[a,b] \Rightarrow \mathcal{U}([a,b])$
- X has $\mathbb{E}[X] = \mu$, $\mathbb{V}[X] = \sigma^2 \Rightarrow N(\mu, \sigma^2)$
- $\underset{\sim}{X}$ has $\mathbb{E}[\underset{\sim}{X}] = \underset{\sim}{\mu}$, covariance $\mathbb{C}[\underset{\sim}{X}, \underset{\sim}{X}] = \underline{\underline{\Sigma}}$
  $\Rightarrow N(\underset{\sim}{\mu}, \underline{\underline{\Sigma}})$
- X lies in $[0,\infty)$ with $\mathbb{E}[X] = \lambda^{-1}$
  $\Rightarrow \mathrm{Exp}(\lambda)$

Finding a continuous distribution that maximizes the Shannon entropy is a little trickier. A common approach is the method of moments

For the discrete case we had:

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

where $f_k(x_i)$ is some function of x_i and $\mathbb{E}[f_k(x_i)] = F_k$

For the continuous case: we match expectations of moments $n = 1, \ldots, M$