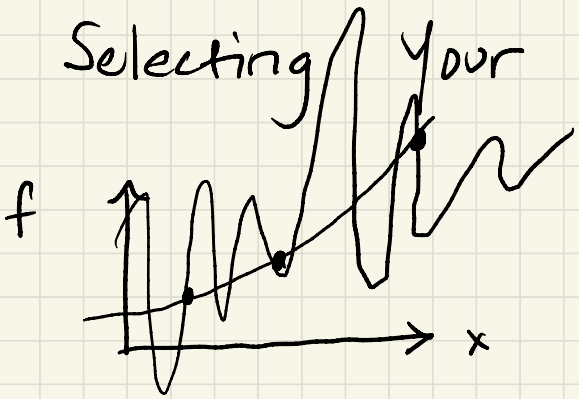
TAM 598 Lecture 11:

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: assign the same probability to all possible variable values

(2) principle of maximum entropy - pick the most "uncertain" distribution that is consistent with what you know

(3) principle of transformation groups - pick a distribution that is invariant under the problems' symmetries

(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*

if R.V. X can take N values, then we assign

$$p(x) = \frac{1}{N} \quad \text{for} \quad x \in \{1, 2, \dots, N\}$$

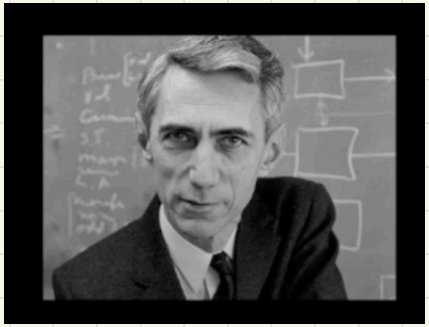
eg) throwing a die $\quad p(x) = \frac{1}{6} \quad x \in \{1, 2, \dots, 6\}$

(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 $\mathbb{V}[X]$.

choose a maximally uncertain distribution that is consistent with the available information.

The uncertainty of a distribution is measured by its information entropy (Claude Shannon, 1948)



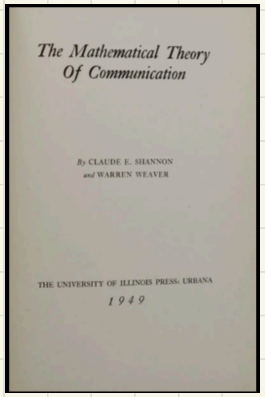
Shannon, Bell Labs

uncertainty $\mathbb{H}[p(X)]$

$$= -\sum_x p(x) \log_2 p(x) \quad \text{discrete}$$

$$= -\int p(x) \ln p(x) \, dx \quad \text{continuous}$$

Shannon Information Content

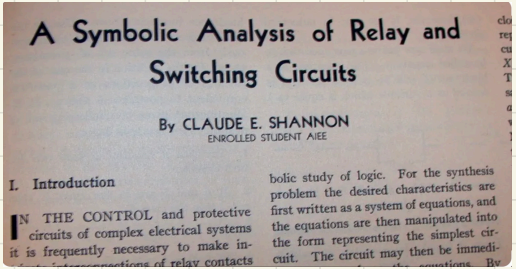


Consider a discrete random variable X. What is the information content associated with a particular measurement $X = x$?

proposal:

$$h(X=x) = \log_2 \frac{1}{P(X=x)} = -\log_2 P(X=x)$$

We can think of this as a measure of "surprise." We are more surprised when we measure an outcome that we thought was less probable


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.

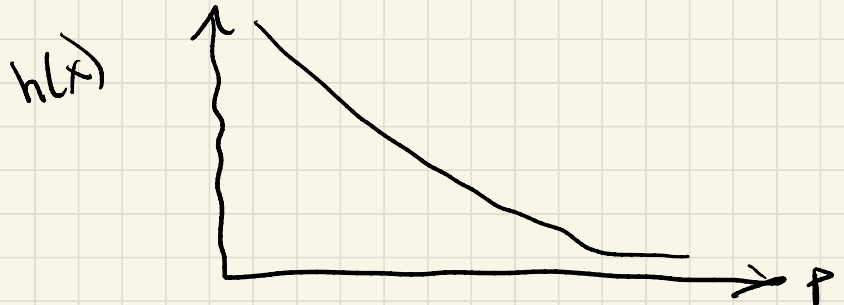
say $P(X=x) = 1$

Then $h(X=x) = \log_2 \frac{1}{P(X=x)} = \log_2 1 = 0$

(2) Information content increases with decreasing probability.

If $p(X=x) < p(X=x')$ then $h(X=x) > h(X=x')$

consider $\frac{d}{dp} h = \frac{d}{dp} \log_2 \frac{1}{p} = -\frac{1}{p \ln p} < 0$ for $p > 0$



Why is this a good proposal for information content?

(3) Information content is additive for independent RVs

if $P(X=x, Y=y) = P(X=x)\,P(Y=y)$

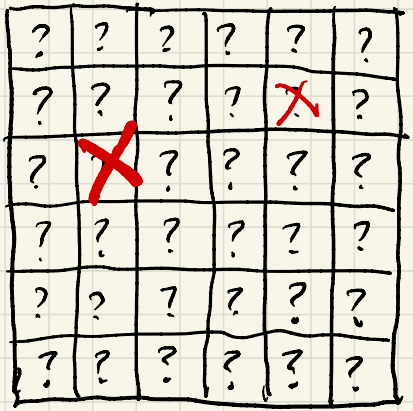
then $h(X=x, Y=y) = h(X=x) + h(Y=y)$

$$h(X=x, Y=y) = \log_2 \frac{1}{P(X=x)\,P(Y=y)}$$

$$= \log_2 \frac{1}{P(X=x)} + \log_2 \frac{1}{P(Y=y)}$$

$$= h(X=x) + h(Y=y)$$

Example Submarine (D. MacKay) – gain information as we do measurements



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

(1) First measurement is a miss

$$h(X=x_1) = \log_2 \frac{1}{p(x_1)} = \log_2 \left(\frac{36}{35}\right) = 0.0406 \text{ bits}$$

(2) Second measurement is a miss

$$h(X=x_1) + h(X=x_2) = \log_2 \left(\frac{36}{35}\right) + \log_2 \left(\frac{35}{34}\right) = 0.0824 \text{ bits}$$

(3) First 18 measurements are misses

$$\log_2 \left(\frac{36}{35}\right) + \log_2 \left(\frac{35}{34}\right) + \ldots + \log_2 \left(\frac{19}{18}\right) = \log_2 \left(\frac{36}{18}\right) = 1$$

(4) Find the submarine on our 25th measurement

$$\log_2 \left(\frac{36}{35}\right) + \log_2 \left(\frac{35}{34}\right) + \ldots \log_2 \left(\frac{12}{11}\right) + \log_2 \left(\frac{11}{1}\right) = \log_2 36 = 5.17 \text{ bits}$$

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

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

$$= \sum_x p(X=x) \log_2 \frac{1}{p(X=x)}$$

$$= -\sum_x p(x) \log_2 p(x)$$

Shannon Entropy
OR
Information Entropy

measures how diffuse the range of outcomes is for a given R.V

eg) 5 possible outcomes

opt 1: 0.2, 0.2, 0.2, 0.2, 0.2

opt 2: 0.35, 0.1, 0.35, 0.1, 0.1

$$H[X] = -5(0.2)\log_2(0.2)$$
$$= 2.32$$

$$H[X] = 2.05$$
$$-2(0.35)\log_2(0.35)$$
$$-3(0.1)\log_2(0.1)$$

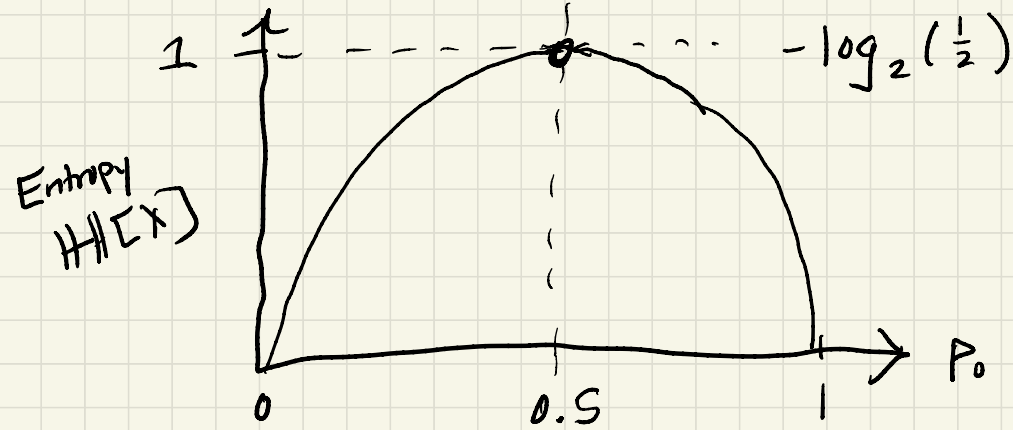
Example: Information entropy of a binary distribution.

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

let $P_0 = p(X=0)$

$P_1 = 1 - P_0 = p(X=1)$

$$H[X] = -P_0 \log_2 P_0 - (1-P_0) \log (1-P_0)$$



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

Our information about X comes in the form

$$\mathbb{E}[f_k(x)] = F_k$$

where f_k is a known function, and F_k are known constants

$k = 1, 2, \ldots, K$

eg) info I = "the expected value of X is μ"

$$\mathbb{E}[x] = \mu \qquad k=1,\ f_1(x) = x,\ F_1 = \mu$$

eg) I = "the expectation is μ; variance is σ^2"

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

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

since $\mathbb{V}[x] = \mathbb{E}[x^2] - (\mathbb{E}[x])^2$

$K=2,\ f_1(x) = x,\ f_2(x) = x^2,\ F_1 = \mu,\ F_2 = \sigma^2 + \mu^2$

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:

$$\begin{cases} \sum_i p_i = 1 & \text{normalization} \\ \mathbb{E}[f_k(x)] = F_k & \text{information} \quad k = 1 \dots K \end{cases}$$

The general solution (Karush-Kuhn-Tucker) conditions

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

λ_k = constants obtained by $F_k = \frac{\partial}{\partial \lambda_k} \log Z$

system of nonlinear equations $k = 1 \dots K$

Z = partition function, normalization $= \sum_i \exp\left\{ \sum_{k=1}^{K} \lambda_k f_k(x_i) \right\}$

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?

constraint: $$\mathbb{E}[X] = \sum_{x=1}^{6} x\, p(x) = 4.5$$

general solution: $$p(x) = \frac{\exp(\lambda x)}{Z(\lambda)}$$

where $$Z(\lambda) = \sum_{i=1}^{6} e^{\lambda i}$$

$$= e^{\lambda} + e^{2\lambda} + \dots e^{6\lambda}$$

To find λ we need to solve:

$$\frac{\partial}{\partial \lambda} \log Z = \frac{1}{Z} \left(\frac{\partial Z}{\partial \lambda} \right) = 4.5$$

$$\frac{1}{Z} \frac{\partial}{\partial \lambda} \left[(e^{\lambda})^1 + (e^{\lambda})^2 + \dots (e^{\lambda})^b \right] = 4.5$$

$$\frac{1}{Z} \left[e^{\lambda} + 2e^{2\lambda} + \dots + be^{b\lambda} \right] = 4.5$$

$$\boxed{\frac{1}{Z} \sum_{i=1}^{b} i e^{i\lambda} = 4.5}$$

Need the root of $f(\lambda) = \frac{1}{Z(\lambda)} \sum_{i=1}^{b} i e^{\lambda i} - 4.5$

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~~ mechanical 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 $n = 1, \ldots, M$