Information theory

Tuesday, January 31, 2023 10:30 AM

- The reason we are studying information theory in this class is because we want to measure how to gather the best data from experiments

- specifically, think about the situation where we want to measure information recieved by making an observation about a random variable x

A rough approximation for this would be the probability $P(x)$

- Moreover, think of communicating perfectly over imperfect channels

  Shanon's interest : modem $\rightarrow$ phone line $\rightarrow$ modem

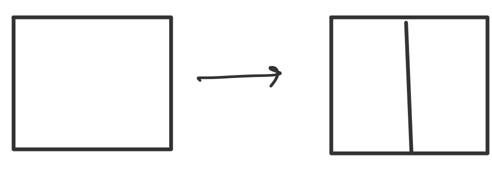
  our interest : Material $\rightarrow$ experiment $\rightarrow$ Property

  (if our expt is perfect, we donot have to study information theory at all)

- This is very relevant to the disussions we have had so far wherein we represented each experiment as a sampling from a probability distribution

- Suppose we now want to measure how much of a choice there is for any particular sample. Alternatively, we can pose this as measuring the uncertainity of the particular outcome we obtained so far

- it is also common to throw around the term called "information" that simply means how much do we cut our space by having access to certain Information



$I = -\log(1/p)$

$\Rightarrow$ by asking a particular question, we got down our uncertainity of the system by half then it has $-\log_2(1/2) = 1$

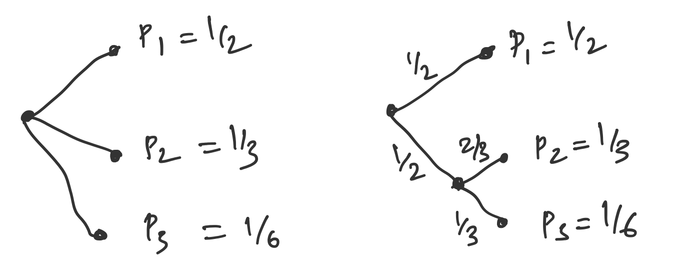
$$\left(\frac{1}{2}\right)^I = P \Rightarrow I = -\log_2 {}^{1/p}$$

- If we are given probabilities of different outcomes $p_1, p_2, \dots, p_n$ we can define a measure $H(p_1, p_2, \dots, p_n)$ such that :

  a) H is continuous in p_i (no sudden jumps)

  b) if all p_i's are equal H is a monotonically increasing function of n. (if all the events are equally likely then we have more choice and uncertainity on which one we pick)

  c) if H can be broken down into two successive choices, the original H should be a weighted sum



$$\Rightarrow H(P_1, P_2, P_3) = H(1/2, 1/2) + \frac{1}{2} H(2/3, 1/3)$$

- Shanon showed that only function that satisfies this assumption is of the form

$$H = -\sum_{i=1}^{n} p_i \log(p_i)$$

$\hookrightarrow$ grows linearly with size

(This is equivalent to $E[I] = \sum p(x) \log(1/p(x))$)

- For uniform distribution $p_i = 1/8$; $H = -8 \times \frac{1}{8} \log_2 1/8 = 3$

  For non-uniform $1/2, 1/4, 1/8, 1/16, (1/64)4$

  $H = 2$

  i.e. more disorder is less entropy

- Original work of Shanon is towards data transmisson (he was working with AT&T Bell labs) where a disoreder means, we require smaller number of bits to transmit.

- For continuous cases : $H = -\int P(x) \ln P(x)\, dx$

- For conditional distribution of $y|x$

$$H(y|x) = -\iint P(y|x) \ln P(y|x)\, dy\, dx$$

Relative entropy :

- consider the case of two distributions $p(x)$ and $q(x)$ where $p(x)$ is unknown but it can be approximated using $q(x)$ which is a known distribution

  Q: How much information is needed to specify x from P if we know the information of specifying x from q

KL divergence :

$$D_{KL}(P \| q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)}\, dx$$

- example: experimental design

  H_0 — null hypothesis

  H_1 — Alternate hypothesis

  $D_{KL}(x|H_0) - D_{KL}(x|H_1) \gtrless 0$ $\quad H_1$ if < 0, H_0 if ≥ 0

- example: approximating complex probabilities to simple

  $P(x)$ — set of microscope images of nano particles

  $q(x)$ — A Gaussian

  we can try to figure out a mapping between $p(x)$ and $q(x)$ that minimizes $D_{KL}(P \| q)$ such that sampling from q is equivalent to sampling from P

Conditinoal entropy :

$$I(x, y) = H(x) - H(x|y)$$

This measures the reduction in uncertainity after observing 'y' also called mutual information

example: collect data to reduce uncertanity of your model by maximizing mutual information

- Fun example: Application of information entropy to wordle example by 3blueBrown (youtube)

- The idea factory : Ch 7,8 (The informationist, Man and the machine)