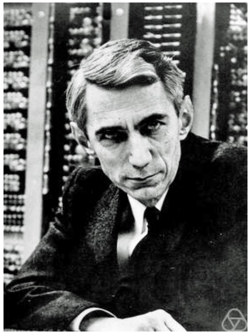
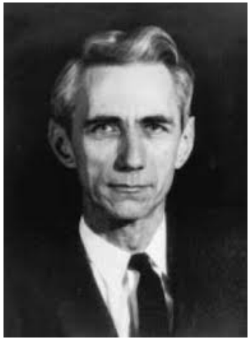
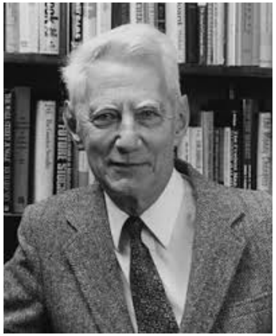
Information measurement and entropy

Hung-Hsuan Chen

Claude E. Shannon

- 1916 – 2001
- Master's thesis: a symbolic analysis of relay and switching circuits (1937)
- PhD thesis: an algebra for theoretical genetics

The foundation of practical digital circuit design

Howard Gardner called Shannon's thesis "possibly the most important, and also the most noted, **master's thesis** of the (20th) century."

"A Mathematical Theory of Communication" (1948) -- Shannon developed **information entropy** as a measure for the **uncertainty** in a message

- This essentially invents the field of information theory

How to measure “information”?

- If we observe the occurrence of an event E with probability p, how much information we get?
  - $I(\boldsymbol{p})$ = ?
  - Note that measure we use p (not E) as the input parameter
  - Essentially, given two events E_1 and E_2, if their occurrence chance are both p, observing E_1 and observing E_2 reveal the same amount of information

The desired properties of the information measure

- We want the information measure $I(p)$ to have the following properties
  1. $I(p) \geq 0$
  2. If p=1, we get no information from the occurrence of the event ➔ $I(p) = 0$
  3. If two **independent** events E (with probability p) and F (with probability q) occur, the information we get from observing both events is the sum of the two information ➔ $I(p * q) = I(p) + I(q)$
  4. The information measure should be a continuous and monotonic function of the probability
     - Observing a more likely events gives us fewer information
- Shannon discovered a proper function to meet the above properties:
  - ➢ $I(p) = \log(1/p) = -\log(p)$

Logarithm with different bases

- $\log_2$: **b**inary **i**nformation uni**t** ➔ bit
- $\log_3$: **tr**inary **i**nformation uni**t** ➔ trit
- $\log_e$: **na**tural information uni**t** ➔ nat

- Unless otherwise mentioned, we often use base 2
  - If you see $\log(p)$, typically we mean $\log_2(p)$

Examples

- If you draw a card at random from a standard N=52-card deck and get a spade-A, how much information you get?
    - ➢ p = 1/52, I(p) = $\log_2(52/1)$ = 5.7
- If the card is a heart, how much information you get?
    - ➢ p = 1/4, $I(p)$ = $\log_2(4/1)$ = 2

Entropy as the expected amount of information

- Suppose the probability of the events (a_1, a_2, …, a_n) are (p_1, p_2, …, p_n) respectively
  - $p_1 + p_2 + \ldots + p_n = 1$
- If we observe a_i, we get information $\log_2(1/p_i)$
  - The probability of observing a_i is p_i
- What is the **expected** amount of information we will get?
  - $\sum_{i=1}^{n} p_i \log_2(\frac{1}{p_i}) = \sum_{i=1}^{n}(-p_i \log_2(p_i))$

$$\sum_{i=1}^{n} p_i \log_2\left(\frac{1}{p_i}\right) = \sum_{i=1}^{n}\left(-p_i \log_2(p_i)\right)$$

Entropy as a measurement of uncertainty

- Example
  - Predicting the tossing result of a fair coin is harder (uncertainty is high)
  - Predicting the tossing result of an unfair coin is easier (uncertainty is low)
- Uniform distribution → every outcome is equally likely → hard to predict → high uncertainty → high entropy
- Gaussian distribution with small variance → certain outcomes are more likely → easier to predict → low uncertainty → low entropy

The range of entropy

ex: $-\sum_{i=1}^{2} P_i \log P_i = -1 \cdot \log 1 - 0 \cdot \log 0 = 0$

- Max: $\log_2(n)$
  - n: the number of possible outcomes
  - If n=2, the max entropy is 1
  - Max occurs when all the probabilities are the same
    - $p_1 = p_2 = p_3 = \ldots = p_n = 1/n$

ex: $-0.5 \log 0.5 - 0.5 \log 0.5 = 1$

- Min: 0
  - Min occurs when one of the probabilities is 1 and the rests are 0's
    - p_i=1; for all $j \neq i$, p_j=0

* $\log 0 = 0$ 看屏幕

Shannon game

- Guess a short paragraph “character by character”
  - The expected value of the log of the number of guesses is the entropy of the paragraph
- The following examples are listed in the book “The most human human” by Brian Christian (Chinese translation: 人性較量)

# of times to correctly guess the character

U N D E R N E A T H _ T H E _ B L U E _
22 1 1 1 1 1 1 1 1 1 1 2 1 1 1 5 6 5 1 2

C U S I O N _ I N _ T H E _ L I V I N G _
2 7 11 5 1 1 1 2 6 5 2 1 1 1 1 1 1 1 1 1 1

R O O M _ I S _ A _ H A N D F U L _ O F _
1 1 1 1 1 1 1 1 1 1 19 3 1 2 13 5 1 1 1 1 1

C H A N G E _ A N D _ T H E _ R E M O T E _
1 21 1 1 2 1 1 1 1 1 1 6 1 1 1 4 2 9 5 1 1 1

C O N T R O L
1 1 1 1 1 1 1

- Information entropy is highly imbalanced
  - Some are easy to guess (low entropy)
  - Some requires much effort (high entropy)

E V E N _ T H O U G H _ Y O U _ D O N T _

K N O W _ H O W _ T O _ F L Y _ Y O U _

M I G H T _ B E _ A B L E _ T O _ L I F T _

Y O U R _ S H O E _ L O N G _ E N O U G H _

F O R _ T H E _ C A T _ T O _ M O V E _ O U T _

F R O M _ U N D E R _ Y O U R _ F O O T

- Brian reported “Y”, “C”, and “M” are the ones with highest entropy (most guesses)
- It seems that “you”, “cat”, and “move” are the essence of the paragraph

Search function and Shannon game

- When using search engines, we tend to pick the less common words (high entropy)
  - Because we know that common words lead you to less relevant pages
- When search for a certain paragraph in a large document, we tend to search for the “special words”
  - Because we know the common words may appear in many paragraphs

Summary

- Information entropy provides a possible way to measure the “information” based on uncertainty
  - A highly certain event provides little information
- We may use information entropy to help build a decision tree classifier
  - We want after a split, each child node is “pure” (less uncertain)
    - i.e., the information entropy is low

Quiz

- Calculate the entropy of the following cases
  1. (O,O,X,X)
     - ➢ ½ $\log_2(2)$ + ½ $\log_2(2) = 1$
  2. (O,O,O,O)
     - ➢ 0
  3. (O,O,X,X,A,A,B,B)
     - ➢ Max entropy ➔ $\log_2(4) = 2$
     - ➢ Or, based on the definition: ¼ $\log_2(4)$ + ¼ $\log_2(4)$ + ¼ $\log_2(4)$ + ¼ $\log_2(4) = \log_2(4) = 2$