香港城市大學
City University of Hong Kong
專業 創新 胸懷全球
Professional·Creative
For The World

Information Theory

CS5483 Data Warehousing and Data Mining

The origin of …

$$\mathrm{Info}(D) \coloneqq \sum_k p_k \log \frac{1}{p_k}$$

$$\mathrm{Info}_{\mathrm{X}}(D) \coloneqq \sum_j \frac{|D_j|}{|D|} \mathrm{Info}(D_j)$$

$$\mathrm{Gain}_{\mathrm{X}}(D) \coloneqq \mathrm{Info}(D) - \mathrm{Info}_{\mathrm{X}}(D)$$

$$\mathrm{SplitInfo}_{\mathrm{X}}(D) \coloneqq \sum_j \frac{|D_j|}{|D|} \log \frac{1}{|D_j|/|D|}$$

- p_k: fraction of tuples in D with class k.
- X: splitting attribute.
- D_j: Data satisfying $\mathrm{X} = j$.

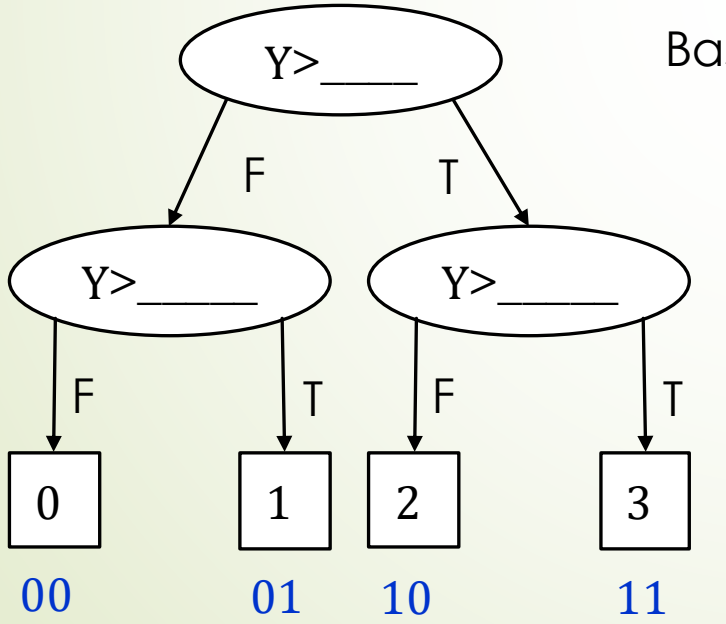
What is information?

"Information is the resolution of uncertainty"

Claude Elwood Shannon
(1916-2001)

A game of resolution of uncertainty

- Determine the value of Y by asking as few T/F questions as possible.
- Given Y takes value from $\{0,1\}$, we can ask whether _____. So, _____ question.
- What if Y takes value from $\{0,1,2,3\}$? _____ questions.



codewords

Basically a decision tree with splitting attribute __.

Encode Y into a sequence of bits with:
F↔ 0 and T↔ 1

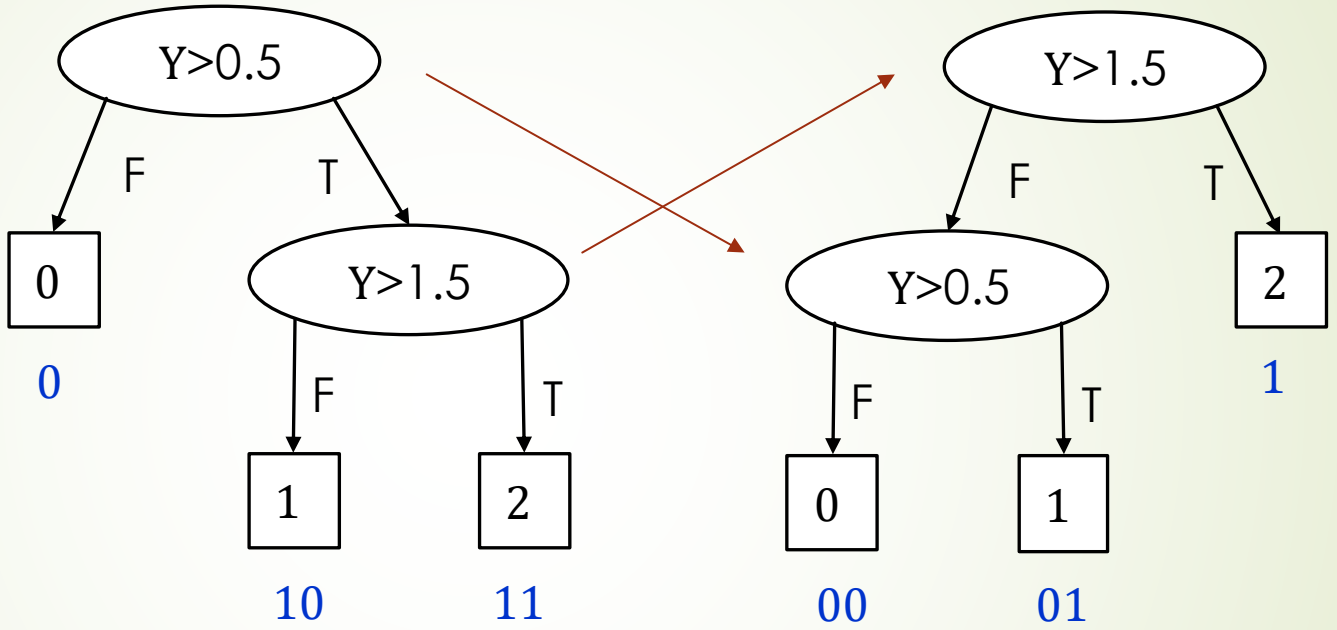
#Questions=#_____ required to encode Y,
i.e., to resolve the *uncertainty* of Y.

A game of resolution of uncertainty

- What if Y takes value from $\{0,1,\dots,2^m-1\}$? _____ questions.
- What if Y takes value from $\{0,1,2\}$? _____questions because $\log_2 3 = 1.585$.
- Can we ask only $\log_2 3 = 1.585$ questions to resolve Y?
- **C**_____ n variables
  - Resolve $\mathrm{Y}_1, \mathrm{Y}_2, \dots, \mathrm{Y}_n$ together using _____ questions.
  - Information rate $\frac{1}{n}$_____ $\rightarrow \log_2 3$ questions per variable
- Can we do even better?

Expected codeword length

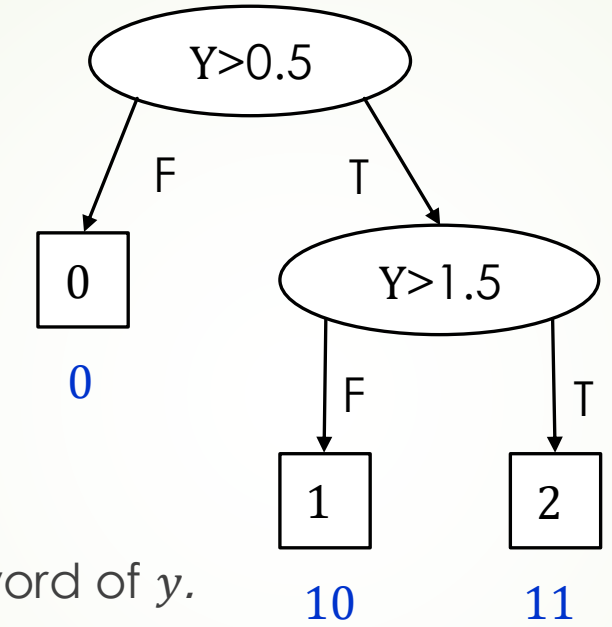
y	$p_Y(y)$
0	½
1	¼
2	¼



- Which encoding is better?
  - Left/right because the expected codeword is larger/smaller.
  - In general, assign shorter/longer codeword for more probable y.

Entropy coding

y	$p_Y(y)$	$\log\frac{1}{p_Y(y)}$
0	½	_
1	¼	_
2	¼	_



- $l(y) :=$ length of the codeword of y.
- *[Variable-length coding]* Y can be encoded with

$$l(y) = \left\lceil \log\frac{1}{p_Y(y)} \right\rceil$$

- Expected codeword length is _____ .

Entropy coding with concatenation

- Encode i.i.d. sequence $\mathrm{Y}^{\mathrm{n}} := (\mathrm{Y}_1, \mathrm{Y}_2, \dots, \mathrm{Y}_{\mathrm{n}})$:

$$l(y_1, y_2, \dots, y_n) = \left\lceil \log_2 \frac{1}{p_{\mathrm{Y}_1,\dots,\mathrm{Y}_{\mathrm{n}}}(y_1,\dots,y_n)} \right\rceil$$

$$= \left\lceil \log_2 \frac{1}{p_{\mathrm{Y}}(y_1)p_{\mathrm{Y}}(y_2)\dots p_{\mathrm{Y}}(y_n)} \right\rceil$$

$$= \left\lceil \sum_{i=1}^{n} \log_2 \frac{1}{p_{\mathrm{Y}}(y_i)} \right\rceil$$

$$\frac{1}{n} l(y_1, y_2, \dots, y_n) = \frac{1}{n} \left\lceil \sum_{i=1}^{n} \log_2 \frac{1}{p_{\mathrm{Y}}(y_i)} \right\rceil$$

$$= \frac{1}{n} \sum_{i=1}^{n} \log_2 \frac{1}{p_{\mathrm{Y}}(y_i)} + o\left(\frac{1}{n}\right)$$

$$\to E\left[\log_2 \frac{1}{p_{\mathrm{Y}}(\mathrm{Y})}\right] = H(\mathrm{Y})$$

- Length is minimal by *Kraft's inequality*.

y_1	y_2	$p_{\mathrm{Y}}(y_1)$	$p_{\mathrm{Y}}(y_2)$	$p_{\mathrm{Y}_1\mathrm{Y}_2}(y_1, y_2)$	$\log_2 \frac{1}{p_{\mathrm{Y}_1\mathrm{Y}_2}(y_1,y_2)}$
0	0	2^{-1}	2^{-1}	2^{-2}	2
	1		2^{-2}	2^{-3}	3
	2		2^{-2}	2^{-3}	3
1	0	2^{-2}	2^{-1}	2^{-3}	3
	1		2^{-2}	2^{-4}	4
	2		2^{-2}	2^{-4}	4
2	0	2^{-2}	2^{-1}	2^{-3}	3
	1		2^{-2}	2^{-4}	4
	2		2^{-2}	2^{-4}	4

Conditional entropy and mutual information

- Join entropy:

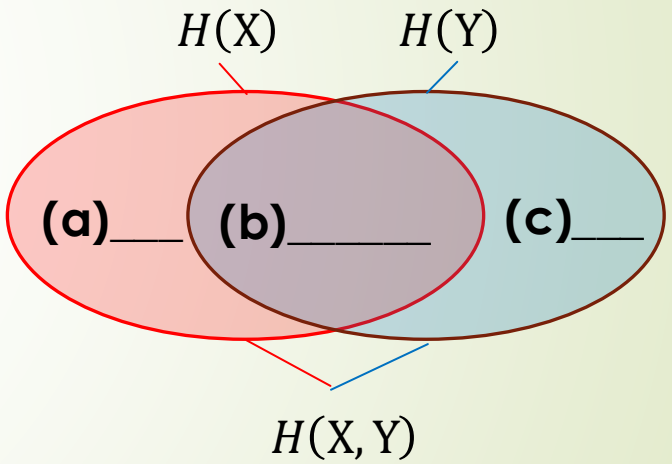
$$H(\mathrm{X},\mathrm{Y}) \coloneqq E\left[\log \frac{1}{p_{\mathrm{XY}}(\mathrm{X},\mathrm{Y})}\right]$$

- Conditional entropy:

$$\begin{aligned} H(\mathrm{Y}|\mathrm{X}) &\coloneqq H(\mathrm{X},\mathrm{Y}) - H(\mathrm{X}) \\ &= E\left[\log \frac{1}{p_{\mathrm{XY}}(\mathrm{X},\mathrm{Y})} - \log \frac{1}{p_{\mathrm{X}}(\mathrm{X})}\right] \\ &= E\left[\log \frac{p_{\mathrm{X}}(\mathrm{X})}{p_{\mathrm{XY}}(\mathrm{X},\mathrm{Y})}\right] \\ &= E\left[\log \frac{1}{p_{\mathrm{Y}|\mathrm{X}}(\mathrm{Y}|\mathrm{X})}\right] \end{aligned}$$

- Mutual Information: $I(\mathrm{X};\mathrm{Y}) \coloneqq H(\mathrm{Y}) - H(\mathrm{Y}|\mathrm{X})$

Venn diagram interpretation



Conditional entropy

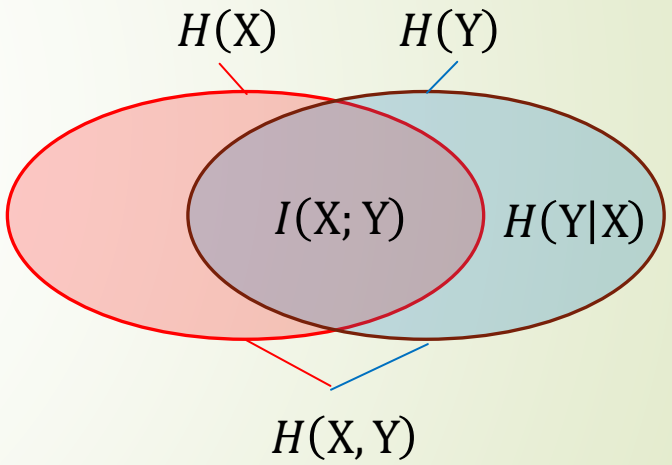
- $\mathrm{Info}_{\mathrm{X}}(D) \coloneqq \sum_j \frac{|D_j|}{|D|} \mathrm{Info}(D_j)$

  $= \sum_j p_{\mathrm{X}}(j) \sum_y p_{\mathrm{Y|X}}(y|j) \log \frac{1}{p_{\mathrm{Y|X}}(y|j)}$

  $= E\left[\log \frac{1}{p_{\mathrm{Y|X}}(\mathrm{Y}|\mathrm{X})}\right] =$_____
- $\mathrm{Gain}_{\mathrm{X}}(D) \coloneqq \mathrm{Info}(D) - \mathrm{Info}_{\mathrm{X}}(D)$

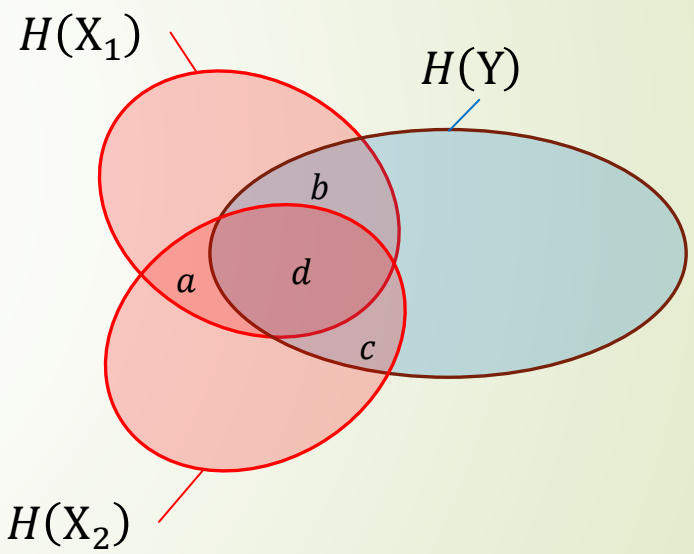
  $= H(\mathrm{Y}) - H(\mathrm{Y}|\mathrm{X}) =$_____
- $\mathrm{Gain}_{\mathrm{X}}(D)/\mathrm{SplitInfo}_{\mathrm{X}}(D) =$_____

Venn diagram interpretation



Properties of information

- Chain rule:
  - $I(\mathrm{X}_1, \mathrm{X}_2; \mathrm{Y}) = I(\mathrm{X}_1; \mathrm{Y}) + I(\mathrm{X}_2; \mathrm{Y}|\mathrm{X}_1) = I(\mathrm{X}_2; \mathrm{Y}) + I(\mathrm{X}_1; \mathrm{Y}|\mathrm{X}_2)$
- Conditioning reduces entropy:
  - $H(\mathrm{Y}) \geq H(\mathrm{Y}|\mathrm{X}_1) \geq H(\mathrm{Y}|\mathrm{X}_1, \mathrm{X}_2)$
  - Equiv., $I(\_\_\_\_\_), I(\_\_\_\_\_) \geq 0$
- Data processing inequality:
  - $I(\mathrm{X}_2; \mathrm{Y}|\mathrm{X}_1) = 0 \Rightarrow I(\mathrm{X}_1; \mathrm{Y}) \leq / \geq I(\mathrm{X}_2; \mathrm{Y})$
- ITIP (Information Theory Inequality Prover)



Reference

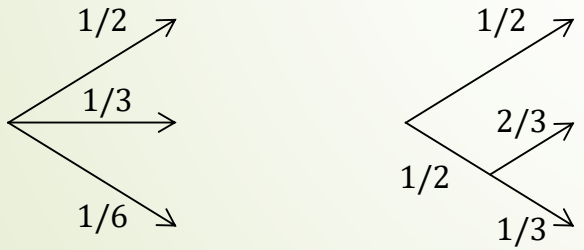
- C. E. Shannon: *A mathematical theory of communication*. Bell System Technical Journal, vol. 27, pp. 379–423 and 623–656, July and October 1948
- Cover, T., & Thomas, J. (2006). Elements of information theory (2nd ed.). Hoboken, N.J.: Wiley-Interscience.

- Source coding with known distribution
  - Entropy coding
  - Huffman coding
- Universal source coding
  - Lampel-Ziv algorithms (LZ77 and LZ78)
  - Deflate algorithm (Huffman coding + LZ77) commonly used in ZIP, gzip, and PNG.

Entropy is the only function H that satisfies…

1. H is a continuous function of a probability distribution $[p_1, p_2, \ldots, p_n]$.
2. If $p_i = \frac{1}{n}$ for all i, then H is an increasing function of n.
3. If a choice is broken down into successive choices, the original H should be a weighted sum of the individual values of H. E.g.,

$$H\left(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}\right) = H\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{1}{2} H\left(\frac{2}{3}, \frac{1}{3}\right)$$



$H = K \sum_i p_i \log \frac{1}{p_i}$ unique up to a scaling factor $K > 0$.

Why measure information in “bits”?

"The choice of a logarithmic base corresponds to the choice of a unit for measuring information. If the base 2 is used the resulting units may be called ***binary digits, or more briefly bits****, a word suggested by J. W. Tukey."*

Shannon July 1948

John Wilder Tukey
(1915-2000)

Why the name "entropy"?

"My greatest concern was what to call it. I thought of calling it 'information', but the word was overly used, so I decided to call it 'uncertainty'. When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, 'You should call it entropy, for two reasons. In the first place your uncertainty function ***has been used in statistical mechanics under that name****. In the second place, and more importantly,* ***no one knows what entropy really is****, so in a debate you will always have the advantage."*

Shannon April 1961

John von Neumann
(1903-1957)

Information theory and machine learning

"Turing and I had an awful lot in common, and we would talk about that kind of question. He had already written his famous paper about Turing Machines, so called, as they call them now, Turing Machines. They didn't call them that then. And we spent much time discussing the concepts of what's in the human brain. How the brain is built, how it works and what can be done with machines and ***whether you can do anything with machines that you can do with the human brain*** *and so on. And that kind of thing. And I had talked to him several times about my notions on Information Theory, I know, and he was interested in those."*

Shannon July 1982

Alan Mathison Turing
(1912 –1954)

Concerning learning machines

"If, as is usual with computing machines, the operators of the machine itself could alter its instructions, there is the possibility that a ***learning process could by this means completely alter the programme*** *in the machine…*

In many types of investigation, e.g., in the theory of information, it is a useful ***assumption that 'computing costs nothing'****. It is important however* ***not to let this assumption become a belief****. In particular when one is considering brains and computers the assumption, and the theories based on it, are not applicable…"*

Turing contribution at conference on Information theory 1950

Alan Mathison Turing
(1912 –1954)