Introduction to Information Theory

(EDAN96)

Luigi Nardi

https://cs.lth.se/luigi-nardi

Lund University

Material

- T. M. Cover. Elements of Information Theory. John Wiley & Sons, 1999 – Sections 2.1 - 2.4
- C. M. Bishop. Pattern Recognition and Machine Learning. Information Science and Statistics. New York: Springer, 2006[1] – Section 1.6
- C. Olah. "Visual Information Theory". 2015. URL: http://colah.github.io/posts/2015-09-Visual-Information/
- L. Papenmeier. Reading group notes

[1] The Bishop PRML book is publicly available: https://www.microsoft.com/en-us/research/wp-content/uploads/2016/05/prml-web-sol-2009-09-08.pdf

Lecture Aims

Understand basic concepts in Information theory

- Entropy
- Kullback-Leibler Divergence (aka KL divergence)
- Mutual information (aka information gain)

Introduction

Information theory gives a precise language to describe:

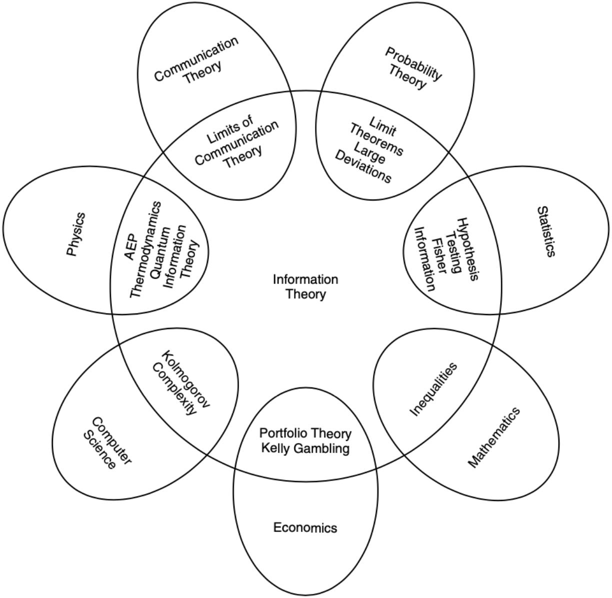
- How uncertain am I?
- How much does knowing the answer to question A tell me about the answer to question B?
- How similar is one set of beliefs to another?
- What is the ultimate data compression? (answer: the entropy)

It is used in a variety of applications:

- Machine Learning
- Data compression
- Etc.

Entropy: Entropy is a measure of uncertainty or disorder in a set of data. It's often used to describe the randomness or unpredictability of a dataset. Higher entropy indicates higher disorder or unpredictability.

Relationship to Other Fields[2]



[2] T. M. Cover. Elements of Information Theory. John Wiley & Sons, 1999.

Scenario

Joe

Images by freepik

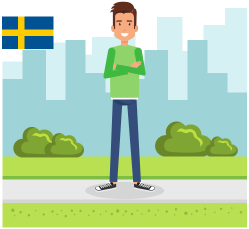
Scenario

Joe

Images by freepik

Scenario

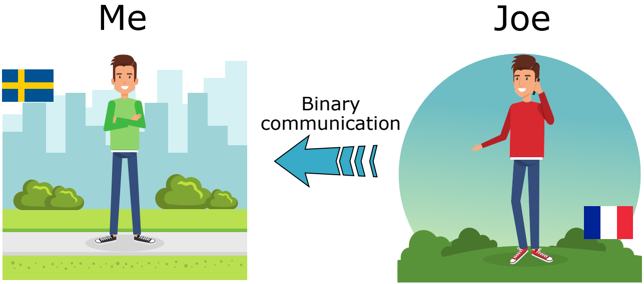
Me



Joe

Images by freepik

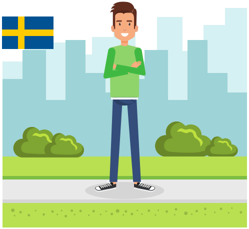
Scenario



Images by freepik

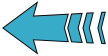
Scenario

Me

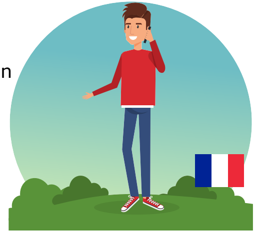


I like to listen about his animals

Binary communication



Joe



Joe likes animals: dog, cat, fish, bird

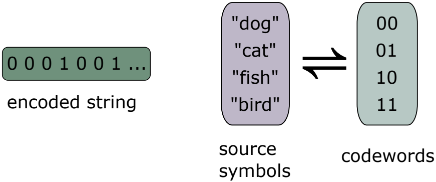
Images by freepik

Codes

0 0 0 1 0 0 1 ...

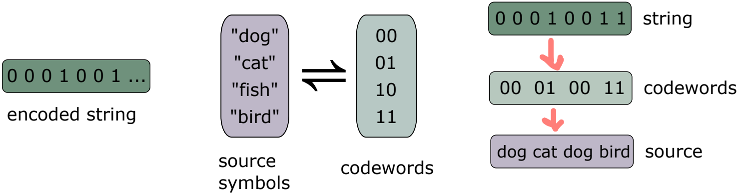
encoded string

Codes



0 0 0 1 0 0 1 ...
encoded string
"dog"
"cat"
"fish"
"bird"
00
01
10
11
source symbols
codewords

Codes

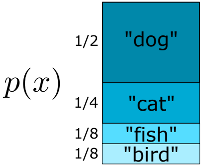

0 0 0 1 0 0 1 ...
encoded string
"dog"
"cat"
"fish"
"bird"
00
01
10
11
source symbols
codewords
0 0 0 1 0 0 1 1
string
00 01 00 11
codewords
dog cat dog bird
source

Word Frequency and Fixed-length Code

- Our code uses codewords that are 2 bits long: $L(x) = 2, \forall x \in \{\text{“}dog\text{”}, \text{“}cat\text{”}, \text{“}fish\text{”}, \text{“}bird\text{”}\}$

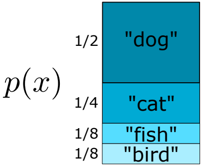
Word Frequency and Fixed-length Code

- Our code uses codewords that are 2 bits long:
  $L(x) = 2, \forall x \in \{\text{“}dog\text{”}, \text{“}cat\text{”}, \text{“}fish\text{”}, \text{“}bird\text{”}\}$
- However, some words are more common:
  Dog lover's example: $p(\text{“}dog\text{”}) > p(\text{“}bird\text{”})$

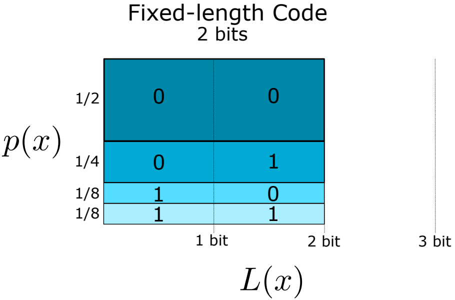


Word Frequency and Fixed-length Code

- Our code uses codewords that are 2 bits long: $L(x) = 2, \forall x \in \{\text{“}dog\text{”}, \text{“}cat\text{”}, \text{“}fish\text{”}, \text{“}bird\text{”}\}$
- However, some words are more common: Dog lover's example: $p(\text{“}dog\text{”}) > p(\text{“}bird\text{”})$

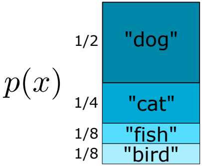


Visualization: prob. of each word $p(x)$ vs length of codeword $L(x)$

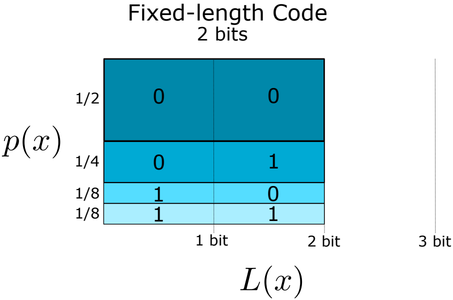


Word Frequency and Fixed-length Code

- Our code uses codewords that are 2 bits long: $L(x) = 2, \forall x \in \{\text{“}dog\text{”}, \text{“}cat\text{”}, \text{“}fish\text{”}, \text{“}bird\text{”}\}$
- However, some words are more common: Dog lover’s example: $p(\text{“}dog\text{”}) > p(\text{“}bird\text{”})$

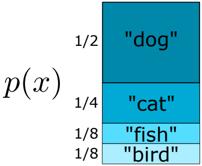


Visualization: prob. of each word $p(x)$ vs length of codeword $L(x)$



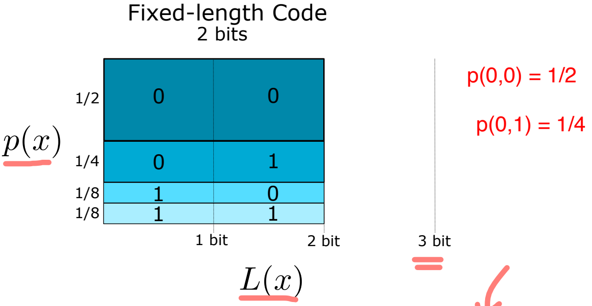
Word Frequency and Fixed-length Code

- Our code uses codewords that are 2 bits long: $L(x) = 2, \forall x \in \{\text{“}dog\text{”}, \text{“}cat\text{”}, \text{“}fish\text{”}, \text{“}bird\text{”}\}$
  Length
- However, some words are more common:
  Dog lover’s example: $p(\text{“}dog\text{”}) > p(\text{“}bird\text{”})$



Some words are more common than others

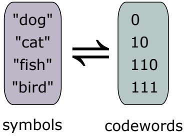
<u>Visualization</u>: prob. of each word $p(x)$ vs length of codeword $L(x)$



The area is the average length of a word we send: $\bar{L}(x) = 2$ bits

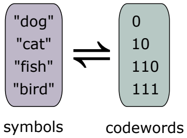
Variable-length Code

- Idea: common codewords are made very short



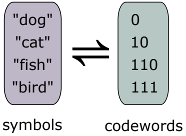
Variable-length Code

- Idea: common codewords are made very short
- Challenge: competition between codewords

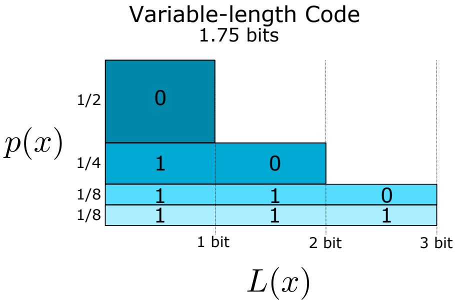


Variable-length Code

- Idea: common codewords are made very short
- Challenge: competition between codewords

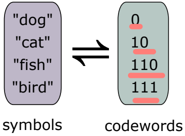


Visualization: prob. of each word $p(x)$ vs length of codeword $L(x)$

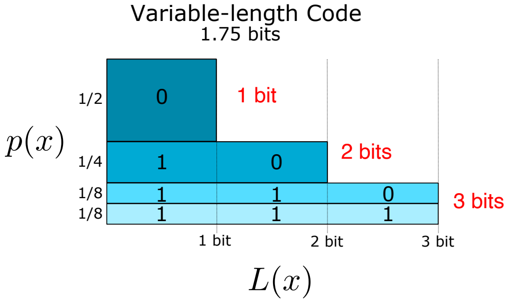


Variable-length Code

- Idea: common codewords are made very short
- Challenge: competition between codewords



Visualization: prob. of each word $p(x)$ vs length of codeword $L(x)$



The area is the average length of a word we send:

$$\Rightarrow \bar{L}(x) = \frac{1}{2} * 1 + \frac{1}{4} * 2 + \frac{1}{8} * 3 + \frac{1}{8} * 3 = 1.75 \text{ bits} = \text{smaller area}$$

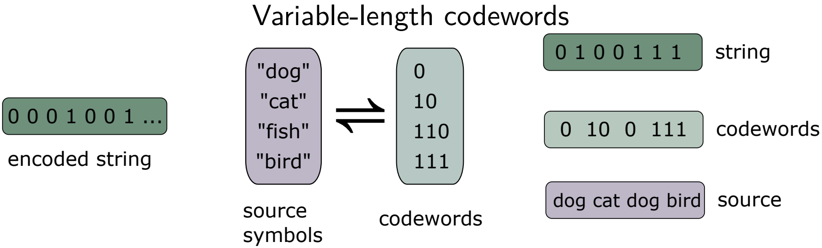
Entropy

- The code of the previous example is the best possible code
- For this distribution, $\bar{L}(x) = 1.75$ bits is the best you can do
- There is simply a fundamental limit
- We call this fundamental limit the entropy of the distribution

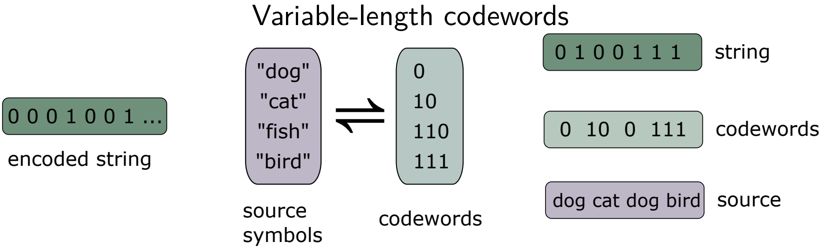
entropy: the minimum average length of a message without losing any information

Remember: $\bar{L}(x)$ is the average codeword length

Prefix Property

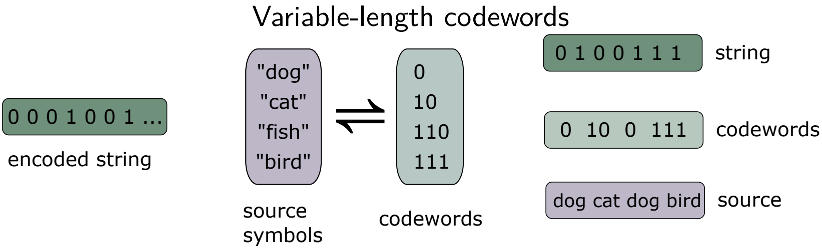


Prefix Property



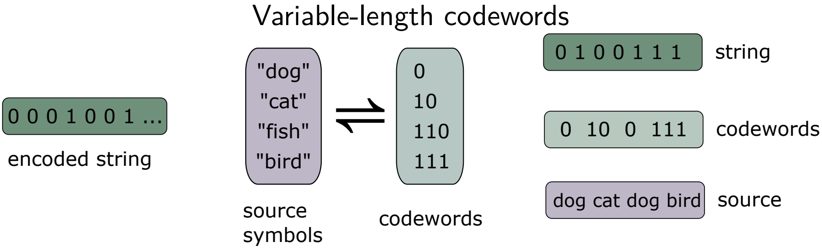
- How do we split the encoded string back into the codewords?

Prefix Property



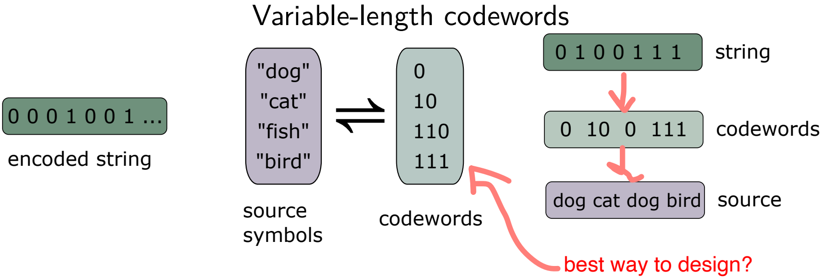
- How do we split the encoded string back into the codewords?
- Easy for fixed-length codes, e.g., split every 2 steps

Prefix Property



- How do we split the encoded string back into the codewords?
- Easy for fixed-length codes, e.g., split every 2 steps
- Code must be uniquely decodable. Example:
  - if 0, 1 and 01 are codewords, can you decode 001?

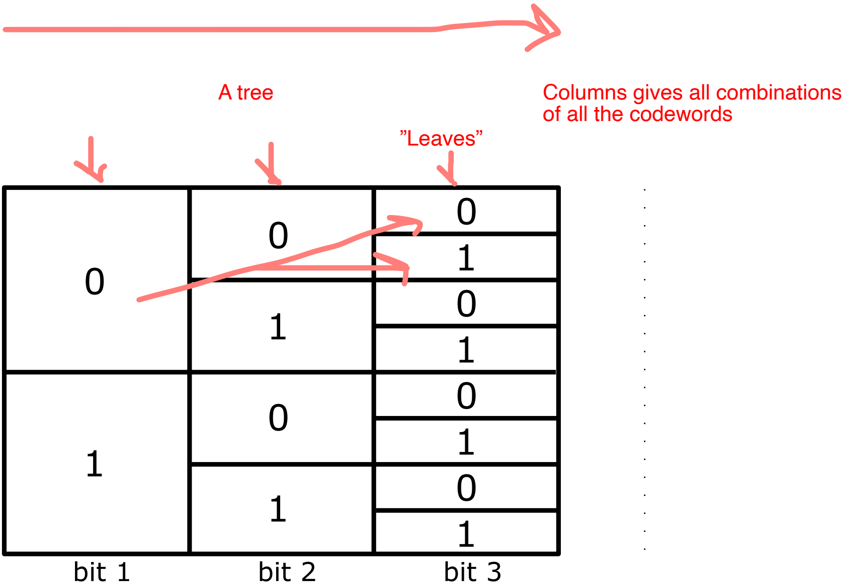
Prefix Property



- How do we split the encoded string back into the codewords?
- Easy for fixed-length codes, e.g., split every 2 steps
- Code must be uniquely decodable. Example:
  - if 0, 1 and 01 are codewords, can you decode 001? Not decodable
- Prefix codes: no codeword is the prefix of another codeword

Eg. 0 and 1 in 01

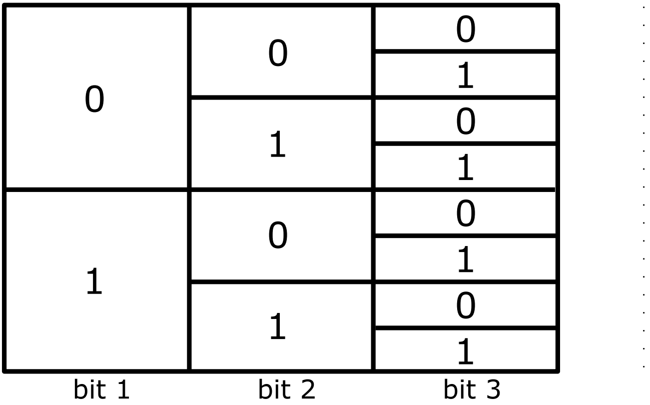
The Space of Codewords



The Space of Codewords

One useful way to think about Prefix codes is:

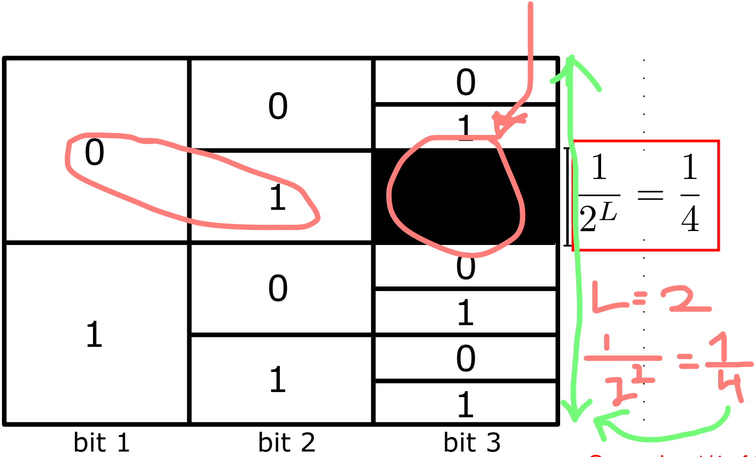
- Every codeword requires a sacrifice from the codewords space
- If we take the codeword 01, we lose codewords with that prefix
  $\Rightarrow$ We can't use 010 or 011010110 anymore (ambiguity)



The Space of Codewords

One useful way to think about Prefix codes is:

- Every codeword requires a sacrifice from the codewords space
- If we take the codeword 01, we lose codewords with that prefix
  $\Rightarrow$ We can't use 010 or 011010110 anymore (ambiguity) => cost



$$\boxed{\frac{1}{2^L} = \frac{1}{4}}$$

Occupying 1/4 of the total codewords

Optimal Encoding

- Limited budget to spend on getting short codewords
  - Short codewords $\Rightarrow$ short average message lengths

Optimal Encoding

- Limited budget to spend on getting short codewords
  - Short codewords $\Rightarrow$ short average message lengths
- One codeword costs a fraction of possible codewords
  - Examples: the cost of a codeword of length
    - 0 is 1, all possible codewords
    - 1, like "0", is $1/2$
    - 2, like "01", is $1/4$

Optimal Encoding

- Limited budget to spend on getting short codewords
  - Short codewords $\Rightarrow$ short average message lengths
- One codeword costs a fraction of possible codewords
  - Examples: the cost of a codeword of length
    - 0 is 1, all possible codewords
    - 1, like "0", is $1/2$
    - 2, like "01", is $1/4$
- Law: the cost of codewords decreases exponentially: $\frac{1}{2^{L(x)}}$

Optimal Encoding

Find the best codeword for a given problem

- Limited budget to spend on getting short codewords
  - Short codewords $\Rightarrow$ short average message lengths
- One codeword costs a fraction of possible codewords
  - Examples: the cost of a codeword of length
    - 0 is 1, all possible codewords
    - 1, like "0", is 1/2
    - 2, like "01", is 1/4
- Law: the cost of codewords decreases exponentially: $\boxed{\frac{1}{2^{L(x)}}}$

What's the best way to use our limited budget?

Distribute our budget in proportion to how common an event is

- This is optimal: Huffman coding $\Leftarrow$ optimal

- Examples: if one event happens this number of times
  - 50%: we spend 50% of our budget buying a short codeword
  - 1%: we only spend 1% of our budget

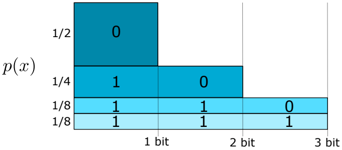
Calculating Entropy

$$cost(x) = \frac{1}{2^{L(x)}} \Rightarrow \boxed{L(x) = \log_2\left(\frac{1}{cost(x)}\right)}$$

Remember: the entropy is the average message length using the best possible code. It is a fundamental limit.

Calculating Entropy

$$cost(x) = \frac{1}{2^{L(x)}} \Rightarrow L(x) = \log_2\left(\frac{1}{cost(x)}\right)$$



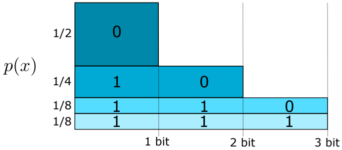
$$L(x) = \log_2\left(\frac{1}{p(x)}\right)$$

- Optimal encoding (Huffman): $cost(x) = p(x)$
- $L(x)$ are the optimal lengths given $p(x)$

Remember: the entropy is the average message length using the best possible code. It is a fundamental limit.

Calculating Entropy

$$cost(x) = \frac{1}{2^{L(x)}} \Rightarrow L(x) = \log_2\left(\frac{1}{cost(x)}\right)$$



$$L(x) = \log_2\left(\frac{1}{p(x)}\right)$$

- Optimal encoding (Huffman): $cost(x) = p(x)$
- $L(x)$ are the optimal lengths given $p(x)$

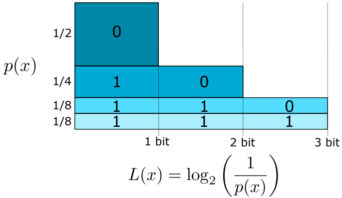
$$\boxed{H(p) = \underbrace{\sum_x p(x)\log_2\left(\frac{1}{p(x)}\right)}_{\text{Entropy of } p} = 1.75 \text{ bits}}$$

Entropy is measured in bits, shortest average message length which is the fundamental limit without losing any info

Remember: the entropy is the average message length using the best possible code. It is a fundamental limit.

Calculating Entropy

$$cost(x) = \frac{1}{2^{L(x)}} \Rightarrow L(x) = \log_2\left(\frac{1}{cost(x)}\right)$$



- Optimal encoding (Huffman): $cost(x) = p(x)$
- $L(x)$ are the optimal lengths given $p(x)$

$$\underbrace{H(p) = \sum_x p(x)\log_2\left(\frac{1}{p(x)}\right)}_{\text{Entropy of } p} = 1.75 \text{ bits}$$

Remark: entropy is often written as $H(p) = -\sum_x p(x)\log_2 p(x)$

Remember: the entropy is the average message length using the best possible code. It is a fundamental limit.

Entropy High-level View

- The entropy has clear implications for compression
- But are there other reasons we should care about it? Yes!
  - It describes how uncertain I am
  - Gives a way to quantify information

The entropy of a random variable

is the average level of "information", "surprise", or "uncertainty" inherent in the variable's possible outcomes

Thus, by knowing the entropy of a random variable we can in ML know how much information it is giving

History: The concept of information entropy was introduced by Claude Shannon in "A Mathematical Theory of Communication (1948)". It is sometimes called Shannon entropy.

Entropy High-level View

- The entropy has clear implications for compression
- But are there other reasons we should care about it? Yes!
  - It describes how uncertain I am
  - Gives a way to quantify information

The entropy of a random variable

is the average level of "information", "surprise", or "uncertainty" inherent in the variable's possible outcomes

- Can the entropy of a discrete RV be negative?

Can never be negative, a message can't have a length less than 0

History: The concept of information entropy was introduced by Claude Shannon in "A Mathematical Theory of Communication (1948)". It is sometimes called Shannon entropy.

Entropy (Formally)

$\sum p(x) \log \frac{1}{p(x)}$

Average

How probable it is out of the total budget

Definition (Entropy)

The *entropy* $H(X)$ of a discrete random variable X is defined by

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x)$$

- X is a random variable with alphabet $\mathcal{X}$ and probability mass function $p(x) = Pr\{X = x\}, \quad x \in \mathcal{X}$

Entropy as Expectation

The entropy depends only on the probabilities of the RV X

- Not on the outcomes

If $X \sim p(x)$, then the expectation of the RV $g(X)$ is written as

$$\mathbb{E}_p[g(X)] = \sum_{x \in \mathcal{X}} p(x) g(x)$$

Now, if $g(X) = \log \frac{1}{p(X)}$

$$H(X) = \mathbb{E}_p[g(X)] = \mathbb{E}_p\left[\log \frac{1}{p(X)}\right] = \mathbb{E}_p[-\log p(X)]$$

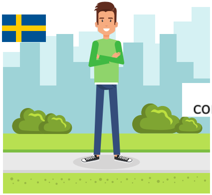
This definition is related to the definition of entropy in thermodynamics

Let's Continue with Our Story

Joe

Joe likes animals: dog, cat, fish, bird
He likes dogs more than anything else

Me



I like to listen about their animals

Binary communication

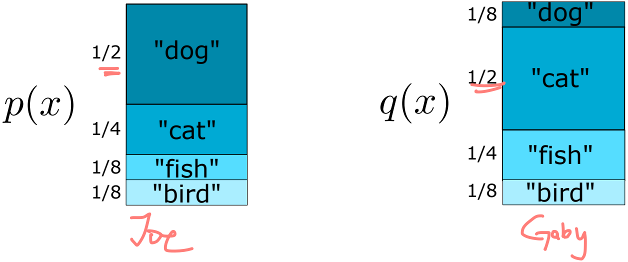
Gaby

Gaby likes animals too: dog, cat, fish, bird
She likes cats more than anything else

Images by freepik

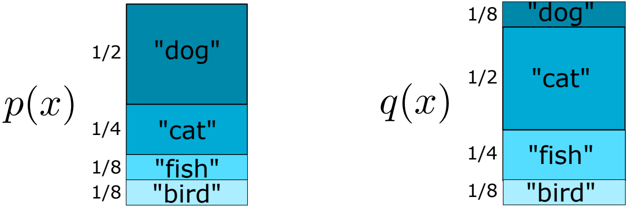
Cross-entropy

- Joe and Gaby say the same words, just at different frequencies



Cross-entropy

- Joe and Gaby say the same words, just at different frequencies

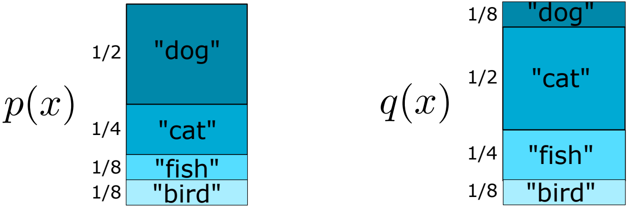


- Joe sends a message: $\bar{L}_{Joe}(x) = 1.75$ bits
- Gaby uses Joe's code: $\bar{L}_{Gaby}(x) = 2.25$ bits ≥ 1.75

Need to adjust for
every probability
distribution

Cross-entropy

- Joe and Gaby say the same words, just at different frequencies



- Joe sends a message: $\bar{L}_{Joe}(x) = 1.75$ bits
- Gaby uses Joe's code: $\underline{\bar{L}_{Gaby}(x) = 2.25 \text{ bits}} \geq 1.75$

$$H_p(q) = \underbrace{\underbrace{\sum_x q(x)}_{\text{Gaby}} \underbrace{\log_2\left(\frac{1}{p(x)}\right)}_{\text{Joe}}}_{\text{Cross-entropy}} = 2.25 \text{ bits}$$

Reads: the cross-entropy of q wrt p

Cross-entropy is Asymmetric

$$H_p(q) \neq H_q(p)$$

Code Used	(Cross-)Entropy	Value
Joe using his own code	$H(p)$	1.75 bits
Gaby using Joe's code	$H_p(q)$	2.25 bits
Gaby using her own code	$H(q)$	1.75 bits
Joe using Gaby's code	$H_q(p)$	2.375 bits

Cross-entropy is Asymmetric

$$H_p(q) \neq H_q(p)$$

unsymmetrical

Code Used	(Cross-)Entropy	Value
Joe using his own code	$H(p)$	1.75 bits
Gaby using Joe's code	$H_p(q)$	2.25 bits
Gaby using her own code	$H(q)$	1.75 bits
Joe using Gaby's code	$H_q(p)$	2.375 bits

Why should we care about cross-entropy?

It expresses how different two probability distributions are

The more difference $\Rightarrow$ the more $H_q(p)$ will be bigger than $H(p)$

Kullback-Leibler Divergence

The interesting thing is:

- The difference between the entropy and the cross-entropy
- Price to pay for using optimized code wrt another distribution
- The distributions are the same $\Leftrightarrow$ difference is zero

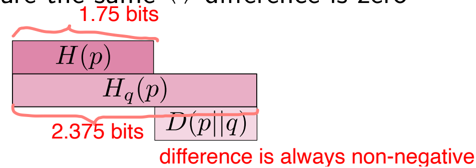
---

KL divergence as a distance: the KL divergence acts like a distance metric between two distributions. However, formally it is not a distance because a distance is symmetric while $D(p||q) \neq D(q||p)$.

Kullback-Leibler Divergence

The interesting thing is:

- The difference between the entropy and the cross-entropy
- Price to pay for using optimized code wrt another distribution
- The distributions are the same $\Leftrightarrow$ difference is zero



$$\underbrace{D(p||q) = H_q(p) - H(p)}_{\text{Kullback-Leibler (KL) divergence}}$$

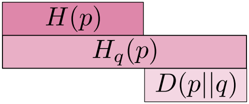
Reads: the KL divergence of p with respect to q

KL divergence as a distance: the KL divergence acts like a distance metric between two distributions. However, formally it is not a distance because a distance is symmetric while $D(p||q) \neq D(q||p)$.

Kullback-Leibler Divergence

The interesting thing is:

- The difference between the entropy and the cross-entropy
- Price to pay for using optimized code wrt another distribution
- The distributions are the same $\Leftrightarrow$ difference is zero



$$\underbrace{D(p||q) = H_q(p) - H(p)}_{\text{Kullback-Leibler (KL) divergence}}$$

Cross-entropy: measure of how well the predicted probabilities match the true distribution

<u>Reads</u>: the KL divergence of p with respect to q

Cross-Entropy and KL divergence are incredibly useful in ML

- Example: we want predicted and GT distributions to be close

---

<u>KL divergence as a distance</u>: the KL divergence acts like a distance metric between two distributions. However, formally it is not a distance because a distance is symmetric while $D(p||q) \neq D(q||p)$.

Kullback-Leibler Divergence (Formally)

$H_g(p(x)) = \sum p(x) \log g(x) -$

$H(p(x)) = \sum p(x) \log p(x)$

Definition (KL Divergence)

The *Kullback-Leibler divergence*, aka *relative entropy*, between two probability mass functions $p(x)$ and $q(x)$ is defined as

$$\begin{aligned} D(p||q) &= H_q(p(X)) - H(p(X)) \\ &= \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \\ &= \mathbb{E}_p \left[\log \frac{p(X)}{q(X)} \right] \end{aligned}$$

KL or Relative Entropy: Properties

Non-negative

Intuition: If $D(p||q) < 0$, we could transmit p more efficiently using the code of q. But we coded p according to $H(p)$, which already gives us the optimal average code length for p.

Non-symmetric

Intuition: $p(x)$ encodes all symbols with same probability ($p(x)$ has a uniform distribution), $q(x)$ puts almost all probability mass on one symbol. Then using the code of q for p is more wasteful than the other way round ("multi-purpose" vs. "specialized" code).

Zero iff equal

$D(p||q) = 0 \Leftrightarrow p = q$.

Entropy and Multiple Variables

- X is a RV for the weather: $X \in \{\text{sun}, \text{rain}\}$
- Y is a RV for clothing: $Y \in \{\text{tee}, \text{coat}\}$
- Message about clothing and weather today:

  $(X, Y) \in \{\underbrace{(\text{sun}, \text{tee})}_{55\%}, \underbrace{(\text{sun}, \text{coat})}_{20\%}, \underbrace{(\text{rain}, \text{tee})}_{5\%}, \underbrace{(\text{rain}, \text{coat})}_{20\%}\}$

Entropy and Multiple Variables

- X is a RV for the weather: $X \in \{\text{sun}, \text{rain}\}$
- Y is a RV for clothing: $Y \in \{\text{tee}, \text{coat}\}$
- Message about clothing and weather today:

  $(X, Y) \in \{\underbrace{(\text{sun}, \text{tee})}_{55\%}, \underbrace{(\text{sun}, \text{coat})}_{20\%}, \underbrace{(\text{rain}, \text{tee})}_{5\%}, \underbrace{(\text{rain}, \text{coat})}_{20\%}\}$
- Now we can figure out the optimal average message length:

$$\Longrightarrow \textbf{Joint Entropy}$$

$$H[p] = -\sum_{x,y} p(x,y) \log p(x,y)$$

Entropy and Multiple Variables

- X is a RV for the weather: $X \in \{\text{sun}, \text{rain}\}$
- Y is a RV for clothing: $Y \in \{\text{tee}, \text{coat}\}$
- Message about clothing and weather today:
  $(X, Y) \in \{(\underbrace{\text{sun}, \text{tee}}_{55\%}), (\underbrace{\text{sun}, \text{coat}}_{20\%}), (\underbrace{\text{rain}, \text{tee}}_{5\%}), (\underbrace{\text{rain}, \text{coat}}_{20\%})\}$
- Now we can figure out the optimal average message length:

$\Longrightarrow$ **Joint Entropy**

- Suppose you know the weather (check it on the news)
  - How much information do I need to provide for the clothing?
- I need to send less, the weather strongly implies clothing
  - When it's sunny: sunny-optimized code
  - When it's raining: raining-optimized code

$\Longrightarrow$ **Conditional Entropy**

Joint Entropy (Formally)

Definition (Joint Entropy)

The *joint entropy* $H(X, Y)$ of a pair of discrete RVs X, Y with a joint distribution $p(x, y)$ is defined as

$$\boxed{H(X,Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{1}{p(x,y)}}$$

It can also be expressed as

$$H(X,Y) = -\mathbb{E}_p[\log p(X,Y)]$$

Conditional Entropy (Formally)

Definition (Conditional Entropy)

The *conditional entropy* $H(Y|X)$ of a pair of discrete RVs X ,Y with a joint distribution $p(x,y)$ is defined as

$$\boxed{H(Y|X) = \sum_{x \in \mathcal{X}} p(x) H(Y|X=x)}$$

$$= -\sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log p(y|x)$$

$$= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log p(y|x)$$

$$= -\mathbb{E}_{p(x,y)}[\log p(Y|X)]$$

Example: Joint and Conditional Entropy

$Y \backslash X$	1	2	3	4	$p(y)$
1	$1/8$	$1/16$	$1/32$	$1/32$	$1/4$
2	$1/16$	$1/8$	$1/32$	$1/32$	$1/4$
3	$1/16$	$1/16$	$1/16$	$1/16$	$1/4$
4	$1/4$	0	0	0	$1/4$
$p(x)$	$1/2$	$1/4$	$1/8$	$1/8$	1

Marginal distribution
Sum all whole row

Sum the whole columns

$P(x=1, y=1) = \frac{1}{8}$

$P(x=4, y=4) = 0$

Example: Joint and Conditional Entropy

$Y \backslash X$	1	2	3	4	$p(y)$
1	$1/8$	$1/16$	$1/32$	$1/32$	$1/4$
2	$1/16$	$1/8$	$1/32$	$1/32$	$1/4$
3	$1/16$	$1/16$	$1/16$	$1/16$	$1/4$
4	$1/4$	0	0	0	$1/4$
$p(x)$	$1/2$	$1/4$	$1/8$	$1/8$	1

H = entropy

$$H(p) = \sum_x p(x) \log_2\left(\frac{1}{p(x)}\right)$$

$$H(X) = \frac{1}{2}\log_2 2 + \frac{1}{4}\log_2 4 + \frac{1}{8}\log_2 8 + \frac{1}{8}\log_2 8 = \frac{7}{4} \text{ bits}$$

$$H(Y) = \frac{1}{4}\log_2 4 + \frac{1}{4}\log_2 4 + \frac{1}{4}\log_2 4 + \frac{1}{4}\log_2 4 = 2 \text{ bits}$$

Example: Joint and Conditional Entropy

Y \ X	1	2	3	4	$p(y)$
1	$1/8$	$1/16$	$1/32$	$1/32$	$1/4$
2	$1/16$	$1/8$	$1/32$	$1/32$	$1/4$
3	$1/16$	$1/16$	$1/16$	$1/16$	$1/4$
4	$1/4$	0	0	0	$1/4$
$p(x)$	$1/2$	$1/4$	$1/8$	$1/8$	1

$$H(X|Y) = \sum_{y \in \mathcal{Y}} p(y) H(X|Y=y)$$

$p(x,y) = p(x|y) \cdot p(y)$

$$= \sum_{y \in \mathcal{Y}} p(y) \sum_{x \in \mathcal{X}} p(x|y) \log_2 \left(\frac{1}{p(x|y)} \right)$$

$$= \sum_{y \in \mathcal{Y}} p(y) \sum_{x \in \mathcal{X}} \frac{p(x,y)}{p(y)} \log_2 \left(\frac{1}{\frac{p(x,y)}{p(y)}} \right)$$

$$= \frac{1}{4} H\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right) + \frac{1}{4} H\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{8}, \frac{1}{8}\right) + \frac{1}{4} H\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) + \frac{1}{4} H(1,0,0,0) = \frac{11}{8} \text{ bits}$$

Example: Joint and Conditional Entropy

$Y \backslash X$	1	2	3	4	$p(y)$
1	$1/8$	$1/16$	$1/32$	$1/32$	$1/4$
2	$1/16$	$1/8$	$1/32$	$1/32$	$1/4$
3	$1/16$	$1/16$	$1/16$	$1/16$	$1/4$
4	$1/4$	0	0	0	$1/4$
$p(x)$	$1/2$	$1/4$	$1/8$	$1/8$	1

Similarly:

$$H(Y|X) = \sum_{x \in \mathcal{X}} p(x) H(Y|X = x) = \frac{13}{8} \text{ bits}$$

$$H(X,Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \frac{1}{p(x,y)} \log p(x,y) = \frac{27}{8} \text{ bits}$$

Remark: $\underbrace{H(Y|X)}_{\frac{13}{8}\text{ bits}} \neq \underbrace{H(X|Y)}_{\frac{11}{8}\text{ bits}}$ conditional entropy is asymmetric

Mutual Information and Entropy

Knowing X can mean that communicating Y requires less info:

joint entropy $\geq$ marginal entropy $\geq$ conditional entropy

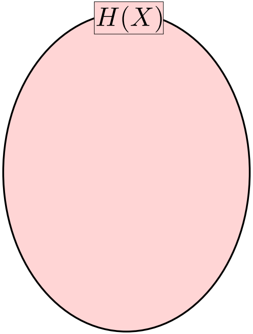
$$\underbrace{H(X,Y)} \geq \underbrace{H(X)} \geq \underbrace{H(X|Y)}$$

Mutual Information and Entropy

Knowing X can mean that communicating Y requires less info:

joint entropy $\geq$ marginal entropy $\geq$ conditional entropy

$$H(X,Y) \geq H(X) \geq H(X|Y)$$



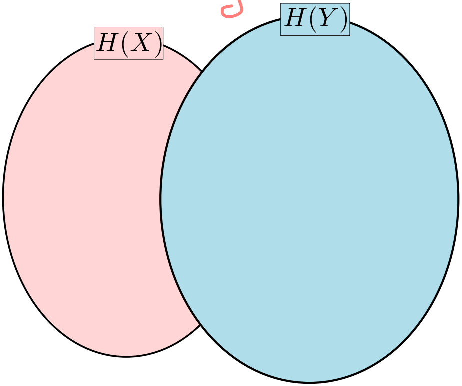
Mutual Information and Entropy

Knowing X can mean that communicating Y requires less info:

joint entropy $\geq$ marginal entropy $\geq$ conditional entropy

$$H(X,Y) \geq H(X) \geq H(X|Y)$$

E.g.

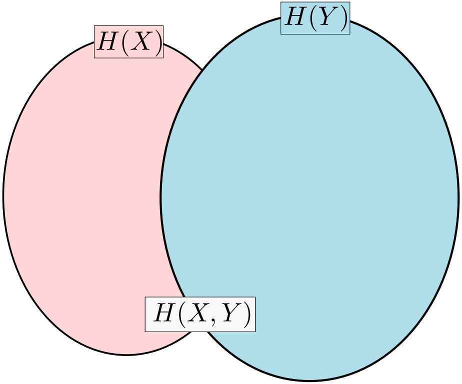


Mutual Information and Entropy

Knowing X can mean that communicating Y requires less info:

joint entropy $\geq$ marginal entropy $\geq$ conditional entropy

$$H(X,Y) \geq H(X) \geq H(X|Y)$$



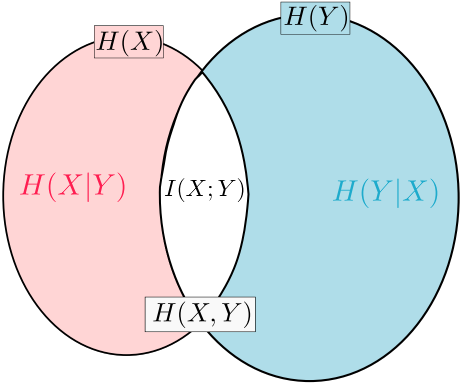
Mutual Information and Entropy

Knowing X can mean that communicating Y requires less info:

joint entropy $\geq$ marginal entropy $\geq$ conditional entropy

$$H(X,Y) \geq H(X) \geq H(X|Y)$$

Mutual information: $I(X;Y)$



Mutual Information and Entropy

Knowing X can mean that communicating Y requires less info:

joint entropy $\geq$ marginal entropy $\geq$ conditional entropy

$$H(X,Y) \geq H(X) \geq H(X|Y)$$

Mutual information: $I(X;Y)$

$I(X;Y) = H(X) - H(X|Y)$

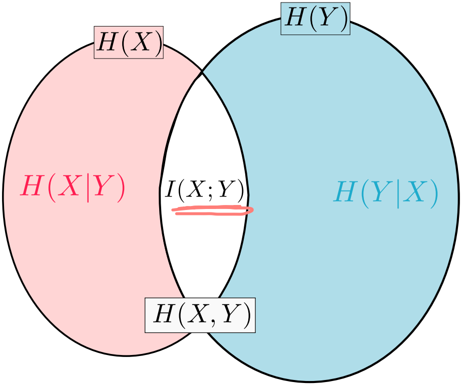
$I(X;Y) = H(Y) - H(Y|X)$

$I(X;Y) = H(X) + H(Y) - H(X,Y)$

$I(X;Y) = I(Y;X)$

$I(X;X) = H(X)$

$I(X;Y) \geq 0$



Mutual Information (Formally)

Definition (Mutual Information)

Consider two RVs X and Y with a joint probability mass function $p(x, y)$ and marginal probability mass function $p(x)$ and $p(y)$. The *mutual information* $I(X; Y)$ is the relative entropy between the joint distribution and the product distribution $p(x)p(y)$, i.e.,

$$\boxed{I(X;Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}}$$

$$= D(p(x,y) || p(x)p(y))$$

$$= \mathbb{E}_{p(x,y)} \left[\log \frac{p(X,Y)}{p(X)p(Y)} \right]$$

Mutual information is also known as ***Information gain***

Entropy and Mutual Information Relationship

The mutual information can be rewritten as

$$\boxed{I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}}$$

$$= \sum_{x,y} p(x,y) \log \frac{p(x|y)}{p(x)}$$

$$= -\sum_{x,y} p(x,y) \log p(x) + \sum_{x,y} p(x,y) \log p(x|y)$$

$$= -\sum_{x} p(x) \log p(x) - \left(-\sum_{x,y} p(x,y) \log p(x|y) \right)$$

$$\boxed{= \underbrace{H(X) - H(X|Y)}_{\substack{\text{Reduction of uncertainty of } X \\ \text{due to the knowledge of } Y}}}$$

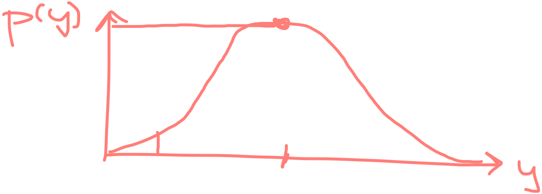
By symmetry, it also follows that $I(X;Y) = H(Y) - H(Y|X)$

Mutual Information: Example I

$Y \backslash X$	1	2	3	4	$p(y)$
1	$1/16$	$1/16$	$1/16$	$1/16$	$1/4$
2	$1/16$	$1/16$	$1/16$	$1/16$	$1/4$
3	$1/16$	$1/16$	$1/16$	$1/16$	$1/4$
4	$1/16$	$1/16$	$1/16$	$1/16$	$1/4$
$p(x)$	$1/4$	$1/4$	$1/4$	$1/4$	1

$$H(X) = 2$$
$$H(Y) = 2$$
$$I(X;Y) = 0$$
$$H(X|Y) = 2$$
$$H(Y|X) = 2$$
$$H(X,Y) = 4$$

Information gain: how much do we learn about Y if we know X?
I(X;Y) = 0 means that X and Y are independent, no exchange of information



Mutual Information: Example I

$Y \backslash X$	1	2	3	4	$p(y)$
1	$1/16$	$1/16$	$1/16$	$1/16$	$1/4$
2	$1/16$	$1/16$	$1/16$	$1/16$	$1/4$
3	$1/16$	$1/16$	$1/16$	$1/16$	$1/4$
4	$1/16$	$1/16$	$1/16$	$1/16$	$1/4$
$p(x)$	$1/4$	$1/4$	$1/4$	$1/4$	1

$$H(X) = 2$$
$$H(Y) = 2$$
$$\underline{I(X;Y) = 0}$$
$$H(X|Y) = 2$$
$$H(Y|X) = 2$$
$$H(X,Y) = 4$$

Corollary

$I(X;Y) = 0 \Leftrightarrow X$ *and* Y *are independent*

Proof.

"$\Leftarrow$": $I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = \sum_{x,y} p(x,y) \log \frac{p(x)p(y)}{p(x)p(y)} = 0$

"$\Rightarrow$": $\log \frac{p(x,y)}{p(x)p(y)} = 0 \Leftrightarrow \frac{p(x,y)}{p(x)p(y)} = 1 \Leftrightarrow p(x)p(y) = p(x,y)$ $\square$

Mutual Information: Example II

$Y \backslash X$	1	2	3	4	$p(y)$
1	$1/8$	$1/16$	$1/32$	$1/32$	$1/4$
2	$1/8$	$1/16$	$1/32$	$1/32$	$1/4$
3	$1/8$	$1/16$	$1/32$	$1/32$	$1/4$
4	$1/8$	$1/16$	$1/32$	$1/32$	$1/4$
$p(x)$	$1/2$	$1/4$	$1/8$	$1/8$	1

$$H(X) = 1.75$$
$$H(Y) = 2$$
$$I(X;Y) = 0$$
$$H(X|Y) = 1.75$$
$$H(Y|X) = 2$$
$$H(X,Y) = 3.75$$

p(y | x = 1)

p(y | x = 2)

p(x, y) -> p(y | x = 1)
p(x,y) = p(y|x)p(x)

$$\frac{1/8}{1/2} = \frac{1}{4}$$

$p(y=1 \mid x=1) = \frac{1}{4}$

$p(y=2 \mid x=1) = \frac{1}{4}$

$\vdots$

$y(y=4, x=1) = 4$

Mutual Information: Example III

$Y \backslash X$	1	2	3	4	$p(y)$
1	$1/8$	$1/16$	$1/32$	$1/32$	$1/4$
2	$1/32$	$1/8$	$1/16$	$1/32$	$1/4$
3	$1/32$	$1/32$	$1/8$	$1/16$	$1/4$
4	$1/16$	$1/32$	$1/32$	$1/8$	$1/4$
$p(x)$	$1/4$	$1/4$	$1/4$	$1/4$	1

$$H(X) = 2$$
$$H(Y) = 2$$
$$I(X;Y) = 0.25$$
$$H(X|Y) = 1.75$$
$$H(Y|X) = 1.75$$
$$H(X,Y) = 3.75$$

The uncertainty about X is reduced after observing Y:

- Depending on Y, certain outcomes of $X|Y$ are more likely

Mutual Information: Example IV

$Y \backslash X$	1	2	$p(y)$
1	$1/2$	0	$1/2$
2	0	$1/2$	$1/2$
$p(x)$	$1/2$	$1/2$	1

$$H(X) = 1$$
$$H(Y) = 1$$
$$I(X;Y) = 1$$
$$H(X|Y) = 0$$
$$H(Y|X) = 0$$
$$H(X,Y) = 1$$

No uncertainty about X if Y has been observed

Speaker's Bio and Contact Info:

- Current:
  - Assistant Professor at Lund University
  - Researcher at Stanford University
  - Founder & CEO at DBtune
- Previously: Imperial College London, Sorbonne Université (Paris), LAAS (Toulouse), Universidad Autónoma de Madrid, La Sapienza (Rome)
- Research: Bayesian Optimization, AutoML
- Applications: HW design, compilers, CV, robotics, DBs
- Offices: E-huset 4128 (LU), Gates building (Stanford campus)

Contact:

Luigi Nardi

luigi.nardi@cs.lth.se

cs.lth.se/luigi-nardi

GitHub: @luinardi

Twitter: @luiginardi

Research Team

Luigi Nardi
Assistant Professor Lund University
Researcher Stanford University

Erik Hellsten
Postdoc

Leonard Papenmeier
Ph.D. Student

Simon Kristoffersson Lind
Ph.D. Student

Carl Hvarfner
Ph.D. Student