Information Theory

Mahdi Roozbahani
Georgia Tech

These slides are based on slides from Le Song, Roni Rosenfeld, Chao Zhang, and Maneesh Sahani.

Outline

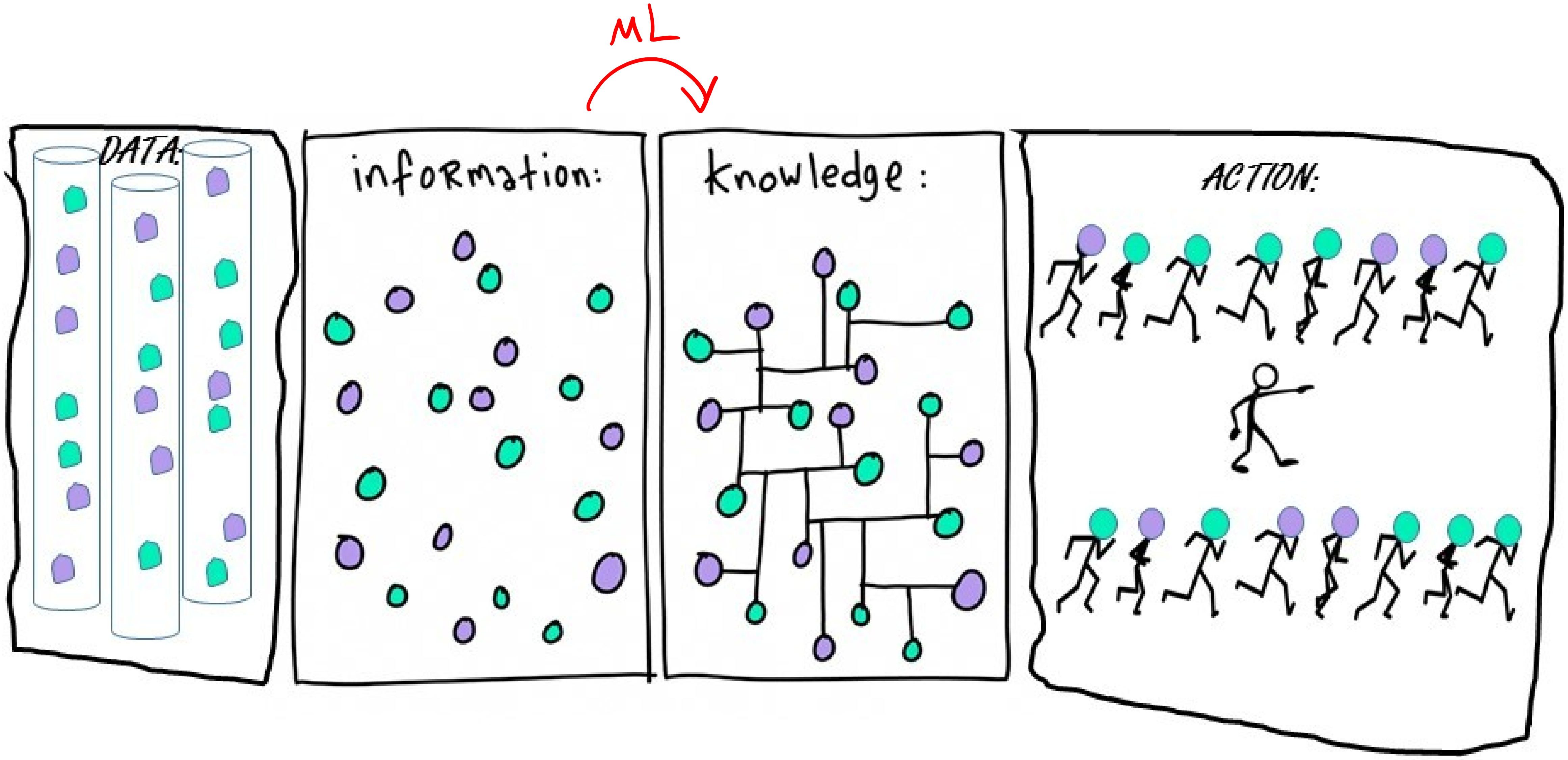
- Motivation
- Entropy
- Conditional Entropy and Mutual Information
- Cross-Entropy and KL-Divergence

Uncertainty and Information

Information is processed data whereas **knowledge** is **information** that is modeled to be useful.

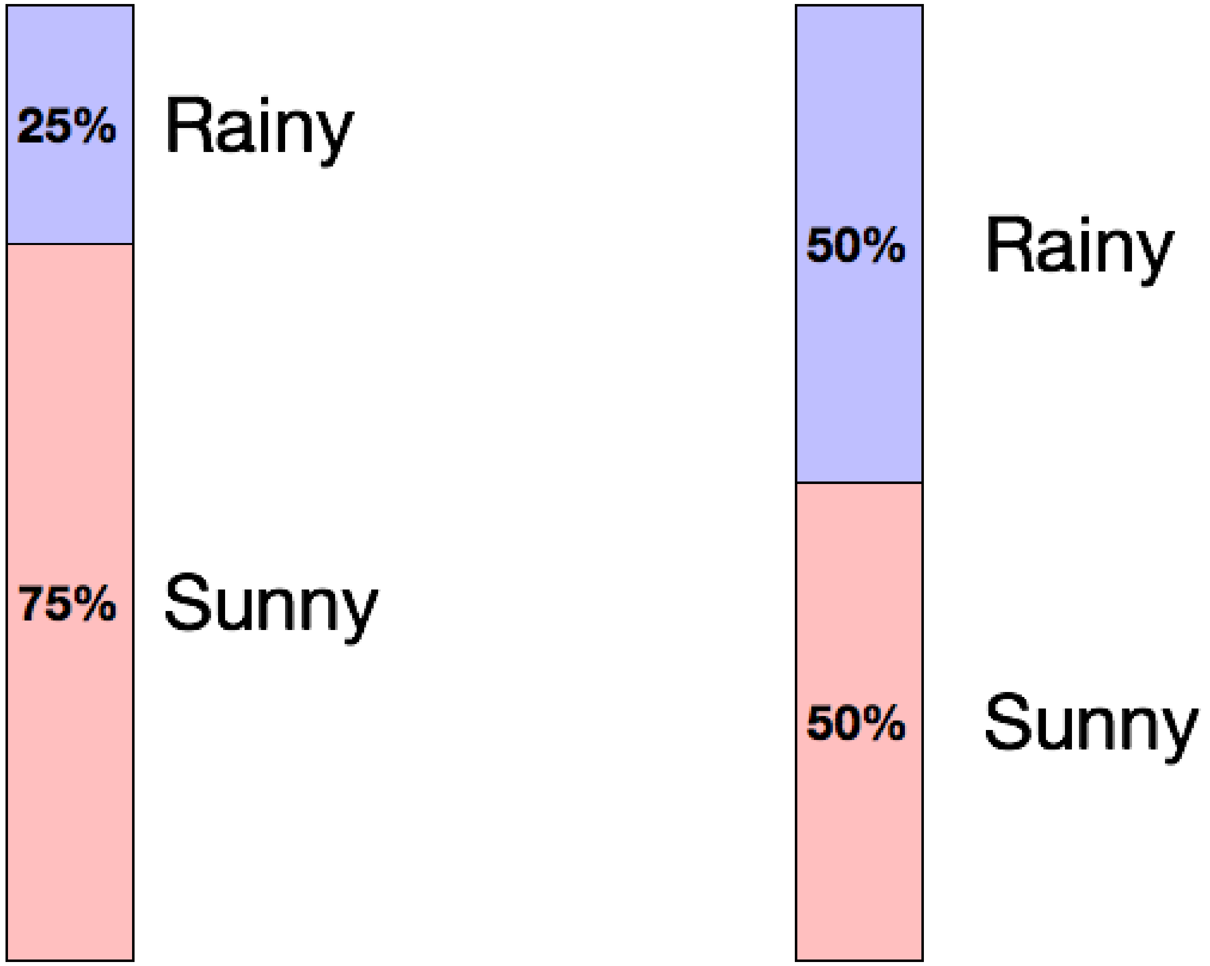
You need **information** to be able to get **knowledge**

- information $\neq$ knowledge
  Concerned with abstract possibilities, not their meaning



Created by Bruce Campbell: “DIKA – ancient Chinese saying for get up and DO! Data-Information-Knowledge-Action.”

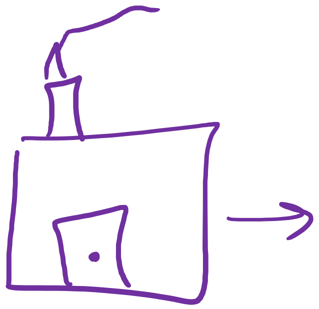
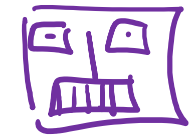
Uncertainty and Information



Which day is more uncertain?

How do we quantify uncertainty?

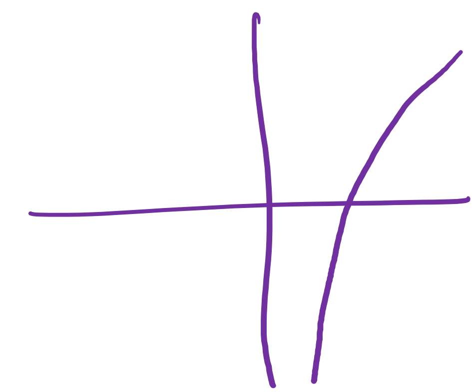
High entropy correlates to high information or the more uncertain

 $\rightarrow$ $X = [cat, cat, cat, dog] \rightsquigarrow$  $\rightsquigarrow$

$P(X = cat) = 1$ $\quad I(x) = \log_2 \frac{1}{P(x)} \Rightarrow I(X = cat) = \log_2 \frac{1}{1} = 0$

$P(X = dog) = \frac{1}{4}$ $\quad I(X = dog) = \log_2 \frac{1}{\frac{1}{4}} = \log_2 2^2 = 2$ bits

$P(X = cat) = \frac{3}{4}$ $\quad I(X = cat) = \log_2 \frac{4}{3}$ bit



$$E[g(x)] = \sum p(x)\, g(x)$$

$$E[I(x)] = \sum p(x)\, I(x) = H(X)$$

$$H(X) = \sum_{\substack{X \\ \text{cat} \quad \text{dog}}} p(x)\, I(x) = p(X=\text{cat})\, I(X=\text{cat}) + p(X=\text{dog})\, I(X=\text{dog})$$

$$= \frac{3}{4} \log_2 \frac{4}{3} + \frac{1}{4}\; 2$$

$$H(X) = -\sum p(x) \log_2 p(x) = \sum p(x) \log_2 \frac{1}{p(x)}$$

[0 0 1 0 0 - - - 1]$_{16 \text{ bits}}$ $= 2^{0} \times 1 + \cdots$

Information

Let X be a random variable with distribution p(x)

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

Have you heard a picture is worth 1000 words?

Information obtained by random word from a 100,000 word vocabulary:

$$I(word) = \log_2\left(\frac{1}{p(x)}\right) = \log_2\left(\frac{1}{1/100000}\right) = 16.61\ bits$$

A 1000 word document from same source:

$$I(document) = 1000 \times I(word) = 16610 \quad \text{bits}$$

A 640*480 pixel, 16-greyscale video picture (each pixel has 16 bits information):

$$I(Picture) = \log_2\left(\frac{1}{1/16^{640*480}}\right) = 1228800$$

$\log_2 \frac{1}{\frac{1}{16}} = 4$

$640 \times 480 \times \log_2 \frac{1}{\frac{1}{16}} =$

A picture is worth (a lot more than) 1000 words!

$I(X = one\ bit) = ?$

Motivation: Compression

- Suppose we observe a sequence of events:
  - Coin tosses
  - Words in a language
  - notes in a song
  - etc.
- We want to record the sequence of events in the smallest possible space.
- In other words we want the shortest representation which preserves all information.
- Another way to think about this: How much information does the sequence of events actually contain?

MOTIVATION: COMPRESSION

1 3

0 00

To be concrete, consider the problem of recording coin tosses in unary.

$$T, T, T, T, H$$

Approach 1:

H	T
0	00

$$00, 00, 00, 00, 0$$

We used 9 characters

Which one has a higher probability: T or H?

Which one should carry more information: T or H?

Motivation: Compression

To be concrete, consider the problem of recording coin tosses in unary.

$$T, T, T, T, H$$

Approach 2:

H	T
00	0

$$0, 0, 0, 0, 00$$

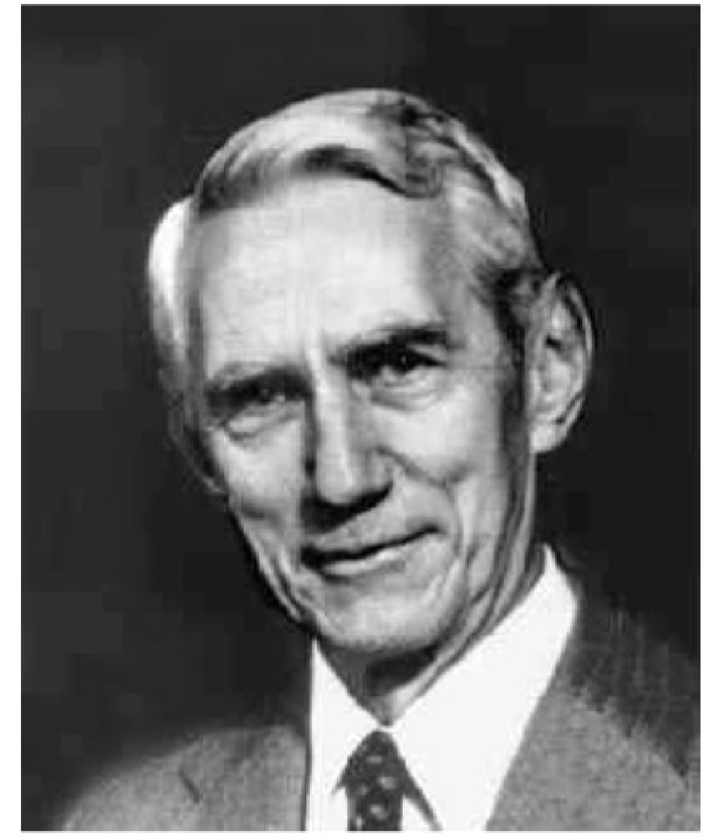
We used 6 characters

Motivation: Compression

- Frequently occuring events should have short encodings
- We see this in english with words such as “a”, “the”, “and”, etc.
- We want to maximise the information-per-character
- seeing common events provides little information
- seeing uncommon events provides a lot of information

Information Theory

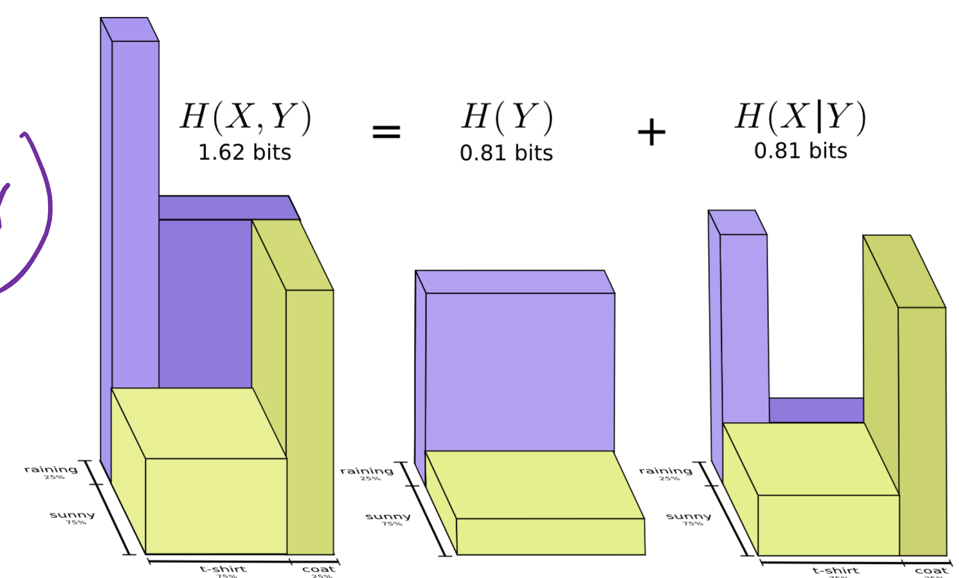
- Information theory is a mathematical framework which addresses questions like:
  - How much information does a random variable carry about?
  - How efficient is a hypothetical code, given the statistics of the random variable?
  - How much better or worse would another code do?
  - Is the information carried by different random variables complementary or redundant?



Claude Shannon

$$P(x,y) = P(x|y)\,P(y)$$

$$H(x,y) = H(x|y) + H(y)$$



Outline

- Motivation
- Entropy
- Conditional Entropy and Mutual Information
- Cross-Entropy and KL-Divergence

Entropy

- Entropy $H(Y)$ of a random variable Y

$$H(Y) = -\sum_{k=1}^{K} P(y = k) \log_2 P(y = k)$$

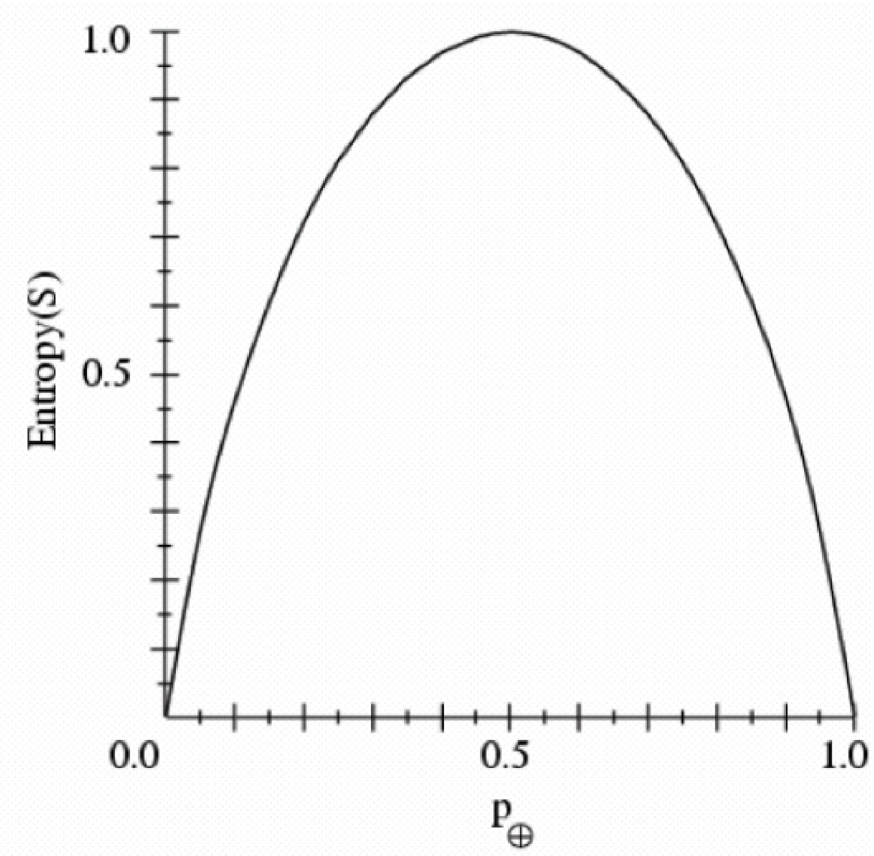
- $H(Y)$ is the expected number of bits needed to encode a randomly drawn value of Y (under most efficient code)

- Information theory:

  Most efficient code assigns $-\log_2 P(Y = k)$ bits to encode the message $Y = k$, So, expected number of bits to code one random Y is:

$$-\sum_{k=1}^{K} P(y = k) \log_2 P(y = k)$$

Entropy

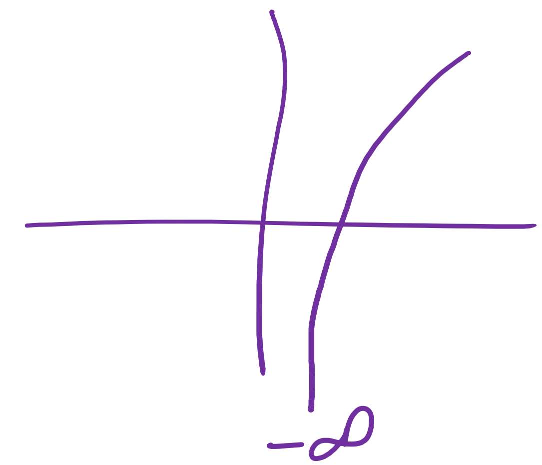


- S is a sample of coin flips
- p_+ is the proportion of heads in S
- p_- is the proportion of tails in S
- Entropy measure the uncertainty of S

$$H(S) \equiv -p_+ \log_2 p_+ - p_- \log_2 p_-$$

Entropy Computation: An Example

$$H(S) \equiv -p_+ \log_2 p_+ - p_- \log_2 p_-$$



head	**0**
tail	**6**

P(h) = 0/6 = 0   P(t) = 6/6 = 1

Entropy = – 0 log 0 – 1 log 1 = – 0 – 0 = 0

$0 \times (-\infty)$

head	**1**
tail	**5**

P(h) = 1/6   P(t) = 5/6

Entropy = – (1/6) $\log_2$ (1/6) – (5/6) $\log_2$ (5/6) = 0.65

head	**2**
tail	**4**

P(h) = 2/6   P(t) = 4/6

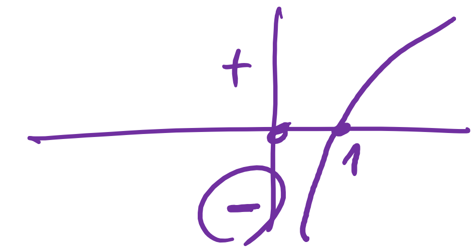
Entropy = – (2/6) $\log_2$ (2/6) – (4/6) $\log_2$ (4/6) = 0.92

Properties of Entropy

$y = \begin{bmatrix} \text{cat} \\ \text{dog} \\ \text{fish} \end{bmatrix}$

$y = \begin{bmatrix} [1 & 0 & 0] \\ [0 & 1 & 0] \\ [0 & 0 & 1] \end{bmatrix}$ p q $\hat{y} = \begin{bmatrix} [0.8 & 0.1 & 0.1] \end{bmatrix}$

$\log(a-b) = \log \frac{a}{b}$



$$H(P) = \sum_i p_i \cdot \log \frac{1}{p_i}$$

$1 \times \log \frac{0.8}{1} + 0 \times \log \frac{0.1}{0} + 0 \times \log \frac{0.1}{0}$

$\sum p_i \log \frac{1}{p_i} - \sum p_i \log \frac{1}{q_i} < 0$

1. Non-negative: $H(P) \geq 0$

$\sum p_i \log \frac{1}{p_i} - \log \frac{1}{q_i} < 0$

2. Invariant wrt permutation of its inputs:
$$H(p_1, p_2, \ldots, p_k) = H(p_{\tau(1)}, p_{\tau(2)}, \ldots, p_{\tau(k)})$$

$\sum p_i \log \frac{q_i}{p_i} < 0$ $p_i = 1$

3. For any *other* probability distribution $\{q_1, q_2, \ldots, q_k\}$:

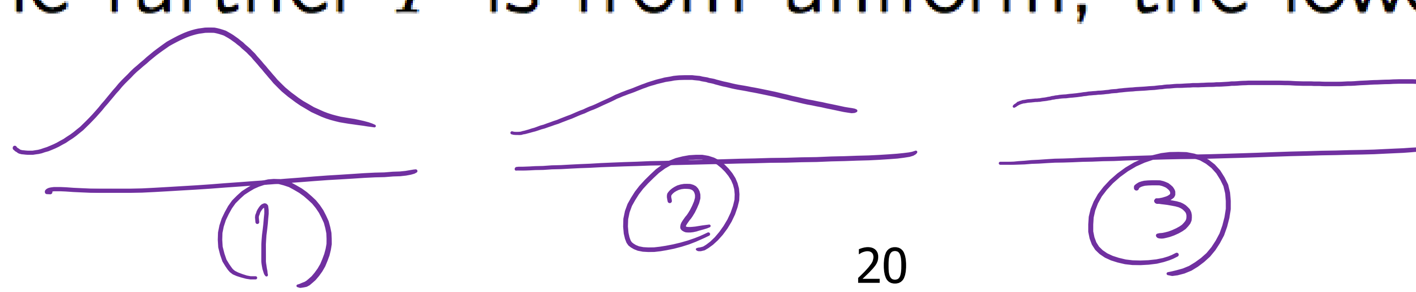
actual output

$$H(P) = \sum_i p_i \cdot \log \frac{1}{p_i} < \sum_i p_i \cdot \log \frac{1}{q_i}$$

predicted

4. $H(P) \leq \log k$, with equality iff $p_i = 1/k \;\; \forall i$

5. The further P is from uniform, the lower the entropy.



Outline

- Motivation
- Entropy
- Conditional Entropy and Mutual Information
- Cross-Entropy and KL-Divergence

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

$E[g(x)] = \sum p(x)\, g(x)$

$H(x) = \sum p(x)\, I(x) = -\sum p(x) \log_2 p(x)$

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

$H(x,y) = \underbrace{H(x|y)}_{\text{Average conditional entropy}} + H(y)$

Joint Entropy

$P(T=c, M=low) = 0.1$

$H(x) = -\sum P(x) \log_2 P(x)$

$H(x) = \sum P(x) \log_2 \frac{1}{P(x)}$

$H(x) = \sum P(x) I(x)$

$P(M=low) = 0.6$

Temperature

huMidity

	cold	mild	hot	
low	0.1	0.4	0.1	0.6
high	0.2	0.1	0.1	0.4
	0.3	0.5	0.2	1.0

- $H(T) = H(0.3, 0.5, 0.2) = 1.48548$ $-0.3 \times \log 0.3 - 0.5 \log 0.5 - 0.2 \log 0.2$
- $H(M) = H(0.6, 0.4) = 0.970951$
- $H(T) + H(M) = 2.456431$
- **Joint Entropy**: consider the space of (t, m) events $H(T, M) = \sum_{t,m} P(T = t, M = m) \cdot \log \frac{1}{P(T=t, M=m)}$
  $H(0.1, 0.4, 0.1, 0.2, 0.1, 0.1) = 2.32193$

$\underbrace{I(x)}= \log_2 \frac{1}{P(M=low, T=cold)}$

$-\underbrace{0.1}_{P(M=low,\ T=cold)} \log 0.1 + \cdots - 0.1 \log 0.1$

Notice that $H(T, M) \leqslant H(T) + H(M)$!!!

$$H(T, M) = H(T|M) + H(M) = H(M|T) + H(T)$$

Conditional Entropy

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

average conditional entropy

$$P(T = t | M = m)$$

$p(T = cold \mid M = low) = \frac{1}{6}$

T

M

	cold	mild	hot	
low	1/6	4/6	1/6	1.0
high	2/4	1/4	1/4	1.0

$p(M = low)\, p(T \mid M = low)$

$+ p(M = high)\, p(T \mid M = h)$

Conditional Entropy:

- $H(T|M = low) = H(1/6, 4/6, 1/6) = 1.25163$
- $H(T|M = high) = H(2/4, 1/4, 1/4) = 1.5$
- **Average Conditional Entropy** (aka equivocation):
  $H(T/M) = \sum_m P(M = m) \cdot H(T|M = m) =$
  $0.6 \cdot H(T|M = low) + 0.4 \cdot H(T|M = high) = 1.350978$

Conditional Entropy

$$P(M = m|T = t)$$

	cold	mild	hot
low	1/3	4/5	1/2
high	2/3	1/5	1/2
	1.0	1.0	1.0

Conditional Entropy:

- $H(M|T = cold) = H(1/3, 2/3) = 0.918296$
- $H(M|T = mild) = H(4/5, 1/5) = 0.721928$
- $H(M|T = hot) = H(1/2, 1/2) = 1.0$
- Average Conditional Entropy (aka Equivocation):
  $H(M/T) = \sum_t P(T = t) \cdot H(M|T = t) =$
  $0.3 \cdot H(M|T = cold) + 0.5 \cdot H(M|T = mild) + 0.2 \cdot H(M|T = hot) = 0.8364528$

Conditional Entropy

- Conditional entropy $H(Y|X)$ of a random variable Y given X_i

Discrete random variables:

$$H(Y|X) = \sum_{x \in X} p(x_i) H(Y|X = x_i) = \sum_{x \in X, y \in Y} p(x_i, y_i) log \frac{p(x_i)}{p(x_i, y_i)}$$

Continuous: $$H(Y|X) = -\int \left(\sum_{k=1}^{K} P(y = k|x_i) \log_2 P(y = k) \right) p(x_i) dx_i$$

Mutual Information

- Mutual information: quantify the reduction in uncertainty in Y after seeing feature X_i

$$I(X_i, Y) = \underbrace{H(Y)}_{\text{Entropy of Parent}} - \boxed{H(Y|X_i)} \rightarrow \text{average conditional entropy children}$$

- The more the reduction in entropy, the more informative a feature.

- Mutual information is symmetric
  - $I(X_i, Y) = I(Y, X_i) = H(X_i) - H(X_i|Y)$
  - $I(Y|X) \;= \int \sum_k^K p(x_i, y = k) \log_2 \frac{p(x_i, y=k)}{p(x_i)p(y=k)} \; dx_i$
  - $= \int \sum_k^K p(x_i|y = k)p(y = k) \log_2 \frac{p(x_i|y = k)}{p(x_i)} \; dx_i$

Properties of Mutual Information

$$
\begin{aligned}
I(X,Y) &= H(X) - H(X|Y) \\
&= \sum_x P(x) \cdot \log \frac{1}{P(x)} - \sum_{x,y} P(x,y) \cdot \log \frac{1}{P(x|y)} \\
&= \sum_{x,y} P(x,y) \cdot \log \frac{P(x|y)}{P(x)} \\
&= \sum_{x,y} P(x,y) \cdot \log \frac{P(x,y)}{P(x)P(y)}
\end{aligned}
$$

Properties of Average Mutual Information:

- Symmetric
- Non-negative
- Zero iff X, Y independent

CE and MI: Visual Illustration

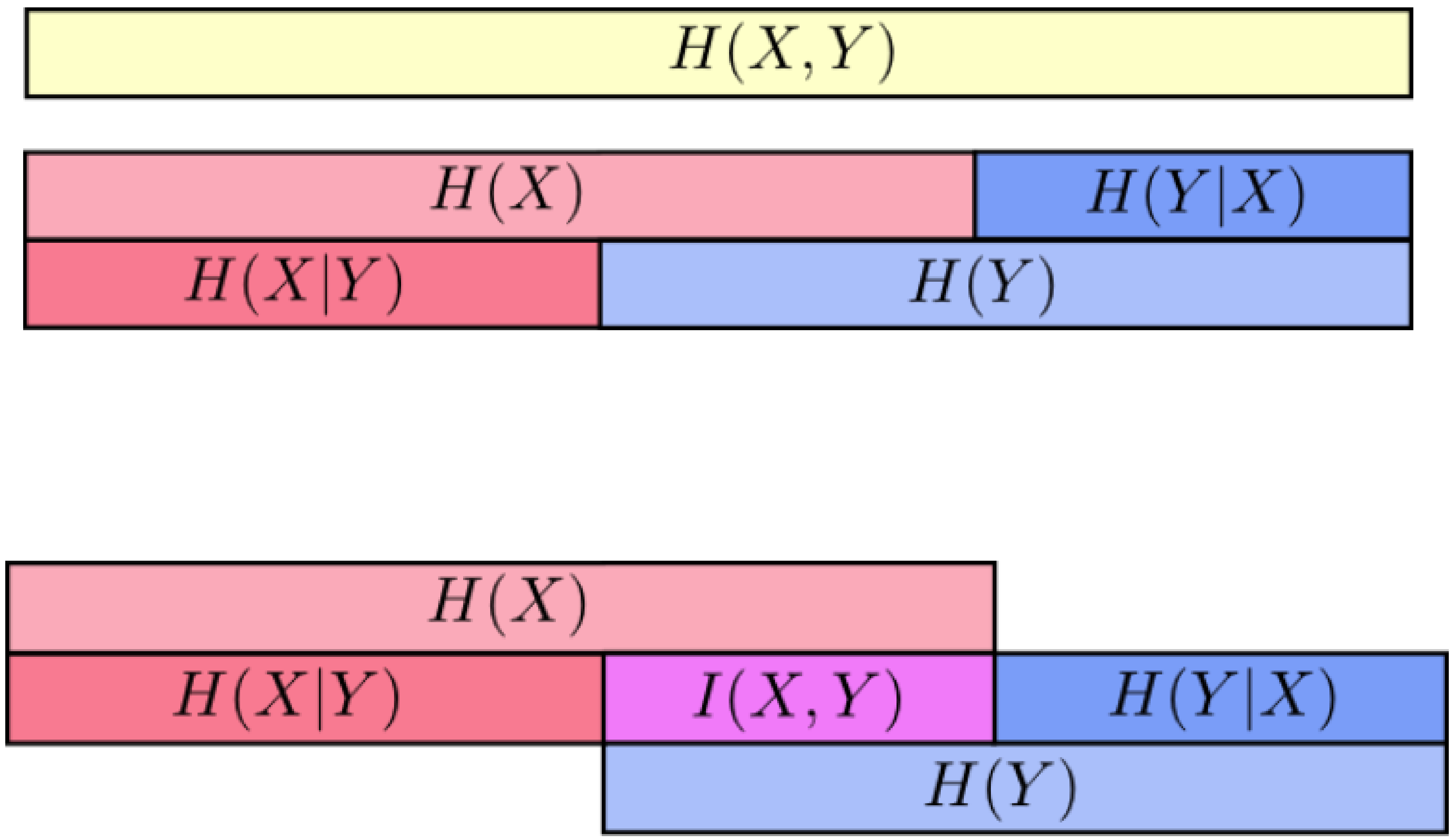


Image Credit: Christopher Olah.

Outline

- Motivation
- Entropy
- Conditional Entropy and Mutual Information
- Cross-Entropy and KL-Divergence

Let's work on this subject in our Optimization lecture

Cross Entropy

Cross Entropy: The expected number of bits when a wrong distribution Q is assumed while the data actually follows a distribution P

$$H(p,q) = -\sum_{x \in \mathcal{X}} p(x)\, \log q(x) = H(P) + KL[P][Q]$$

This is because:

$$H(p,q) = \mathrm{E}_p[l_i] = \mathrm{E}_p\left[\log \frac{1}{q(x_i)}\right]$$

$$H(p,q) = \sum_{x_i} p(x_i)\, \log \frac{1}{q(x_i)}$$

$$H(p,q) = -\sum_{x} p(x)\, \log q(x).$$

Kullback-Leibler Divergence

Another useful information theoretic quantity measures the difference between two distributions.

$$\mathbf{KL}[P(S)\|Q(S)] = \sum_s P(s) \log \frac{P(s)}{Q(s)}$$

$$= \underbrace{\sum_s P(s) \log \frac{1}{Q(s)}}_{\text{cross entropy}} - \mathbf{H}[P] = H(P,Q) - H(P)$$

Excess cost in bits paid by encoding according to Q instead of P.

KL Divergence is a **KIND OF** distance measurement

$$-\mathbf{KL}[P\|Q] = \sum_s P(s) \log \frac{Q(s)}{P(s)}$$

log function is concave or convex?

$$\sum_s P(s) \log \frac{Q(s)}{P(s)} \leq \log \sum_s P(s) \frac{Q(s)}{P(s)}$$

By Jensen Inequality

$$= \log \sum_s Q(s) = \log 1 = 0$$

So $\mathbf{KL}[P\|Q] \geq 0$. Equality iff $P = Q$

When $P = Q,\ KL[P||Q] = 0$

Take-Home Messages

- Entropy
  - A measure for uncertainty
  - Why it is defined in this way (optimal coding)
  - Its properties
- Joint Entropy, Conditional Entropy, Mutual Information
  - The physical intuitions behind their definitions
  - The relationships between them
- Cross Entropy, KL Divergence
  - The physical intuitions behind them
  - The relationships between entropy, cross-entropy, and KL divergence