Information Theory

Mahdi Roozbahani
Georgia Tech

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

Outline

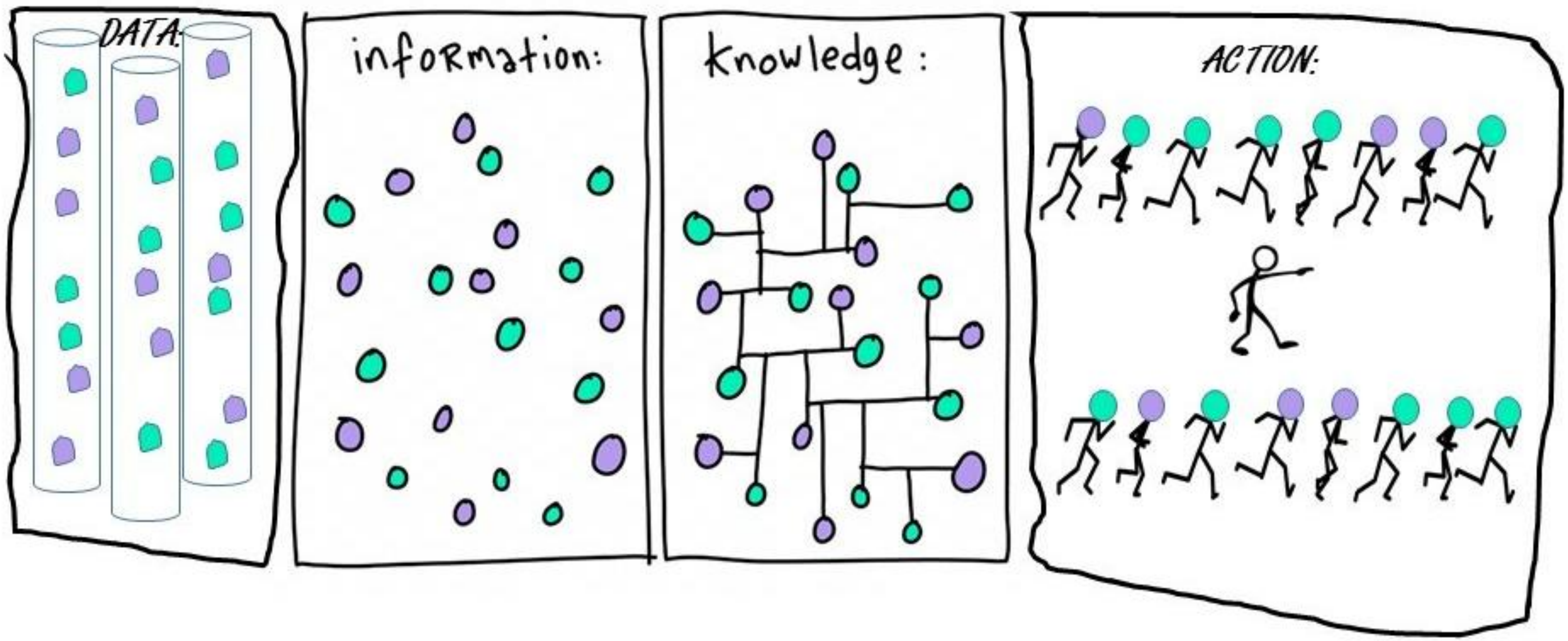
- Motivation
- Entropy
- Conditional Entropy and Mutual Information
- Cross-Entropy and KL-Divergence

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**

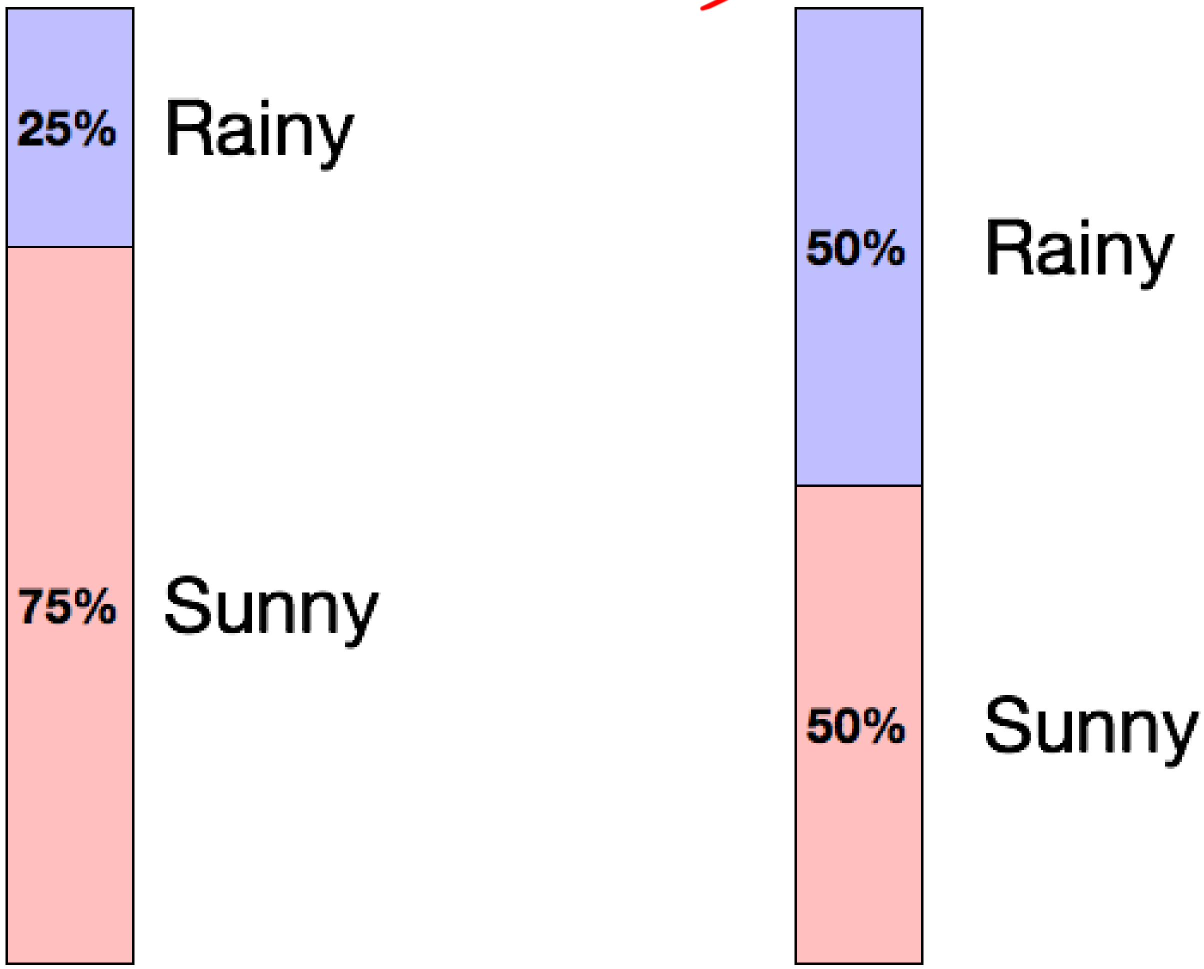
- information $\neq$ knowledge
  Concerned with abstract possibilities, not their meaning



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

Uncertainty and Information

1 2



Which day is more uncertain?

How do we quantify uncertainty?

High entropy correlates to high information or the more uncertain

 $\rightsquigarrow$ [cat, cat, cat, dog] $\rightsquigarrow$  $\longrightarrow$ output

data — information — knowledge — take action

$$P(x = \text{cat}) = \frac{3}{4} \qquad P(x = \text{dog}) = \frac{1}{4}$$

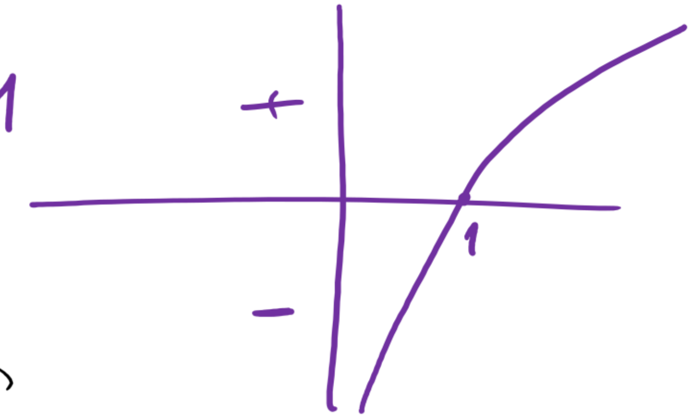
$$\underline{I(x)} = \log_2 \left(\frac{1}{P(x)} \right) = -\log_2 P(x)$$

$$P(x)^{-1}$$

$$E[g(x)] = \sum P(x)\, g(x) \rightsquigarrow \underline{E[I(x)]} = \sum P(x)\, I(x) = H(x)$$

$$H(x) = P(x = \text{cat})\, I(x = \text{cat}) + P(x = \text{dog})\, I(x = \text{dog})$$

$$H(x) = \frac{3}{4} \log_2 \frac{4}{3} + \frac{1}{4} \underbrace{\log_2 4}_{\log_2 2^2 = 2}$$



$$k = 2 \Rightarrow H(x) = \frac{1}{2}\log_2 2 + \frac{1}{2}\log_2 2 = \log_2 2 = k = \log_2 k = 1$$

$$k = 6 \Rightarrow H(x) = \frac{1}{6}\log_2 6 + \dots + \frac{1}{6}\log_2 6 = \log_2 6 = \log_2 k$$

$$k = \infty \Rightarrow H(x) = \log_2 k \rightsquigarrow \infty$$

$$0 \leq H(x) \leq +\infty$$

$P(x = \text{cat}) = 0 \qquad P(x = \text{dog}) = 1$

$$H(x) = -\underbrace{1\log 1}_{0} - \underbrace{0\log 0}_{0} = 0$$

Information

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

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

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

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

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

① $P(X=T) = \frac{4}{5}$

② $P(X=H) = \frac{1}{5}$ ⇝ ② higher surprised information

"It needs higher bits of information associated to it"

Approach 1:

① ②

0 00

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

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

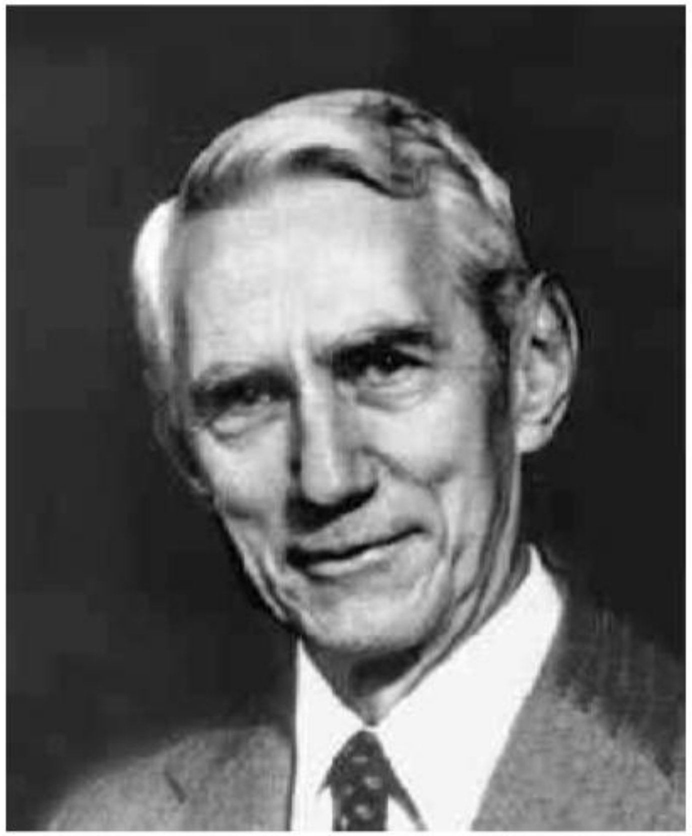
We used 6 characters

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

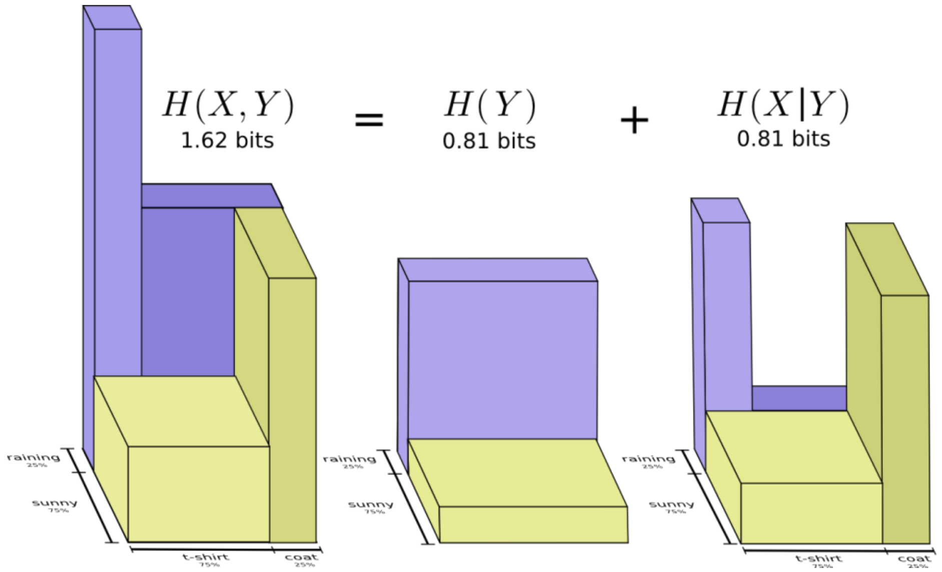
- Information theory is a mathematical framework which addresses questions like:
  - How much information does a random variable carry about?
  - How efficient is a hypothetical code, given the statistics of the random variable?
  - How much better or worse would another code do?
  - Is the information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

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

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

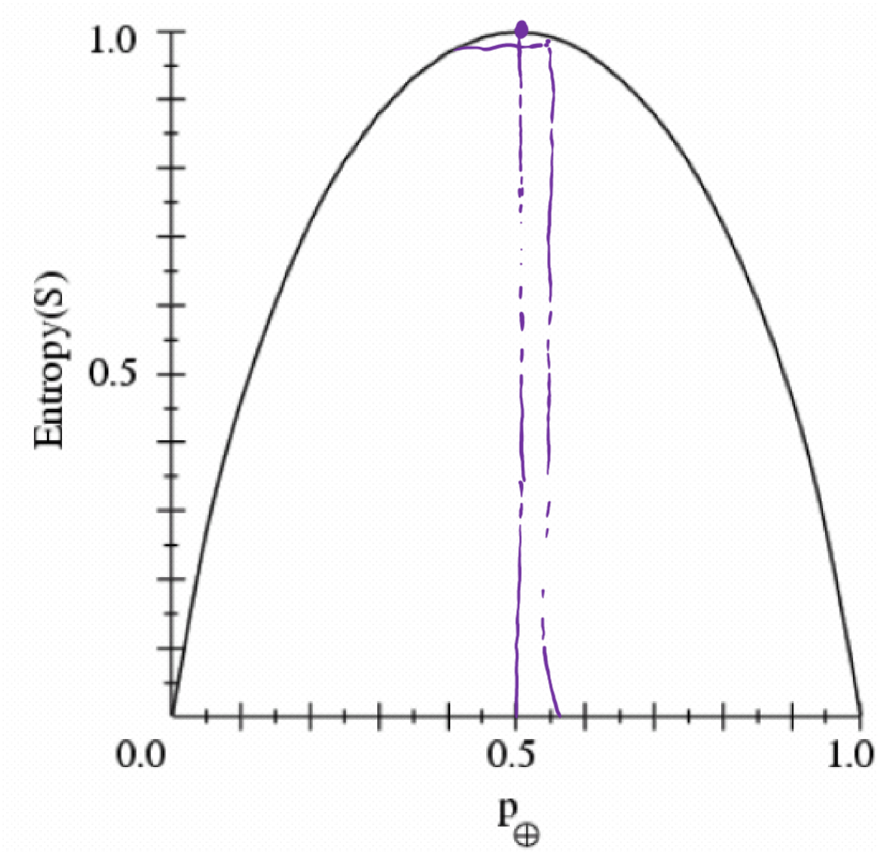
- $H(Y)$ is the expected number of bits needed to encode a randomly drawn value of Y (under most efficient code)

- 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

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

1

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

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

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

2

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

3

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

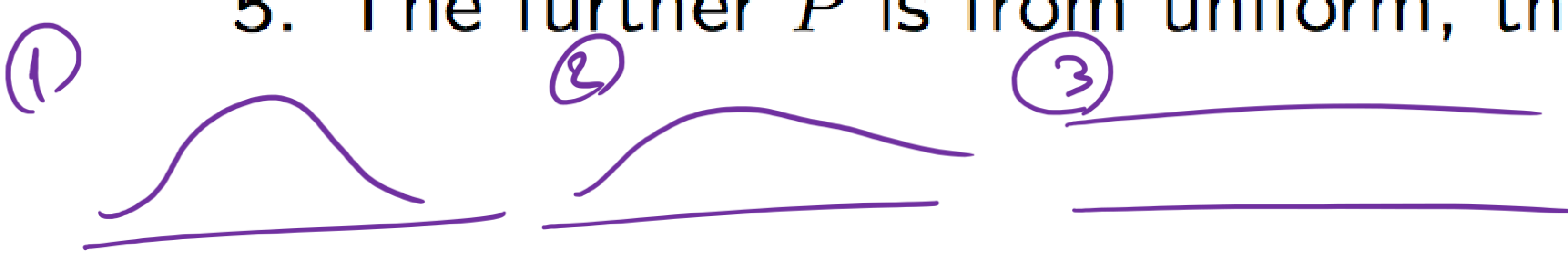
P(h) = 2/6 **P(t) = 4/6**

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

Properties of Entropy

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

1. Non-negative: $H(P) \geq 0$
2. Invariant wrt permutation of its inputs:
   $H(p_1, p_2, \ldots, p_k) = H(p_{\tau(1)}, p_{\tau(2)}, \ldots, p_{\tau(k)})$
3. For any *other* probability distribution $\{q_1, q_2, \ldots, q_k\}$:
   $$\boxed{H(P) \;=\; \sum_i p_i \cdot \log \frac{1}{p_i} < \sum_i p_i \cdot \log \frac{1}{q_i}}$$
   actual pdf
   Predicted pdf
4. $H(P) \leq \log_2 k$, 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

for dice $\log_2 6 = \log_2 (k)$

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

$E[I(x)] = \sum P(x) I(x) = H(x)$

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

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

$0 \leq H(x) < \infty$

$P(M=\text{low}, T=\text{cold}) = 0.1$

$P(M=\text{low}) = 0.6$

$P(M=\text{low} \mid T=\text{cold}) = \frac{0.1}{0.3} = \frac{1}{3}$

$P(T=\text{cold} \mid M=\text{low}) = \frac{0.1}{0.6} = \frac{1}{6}$

Joint Entropy

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

$P(T=\text{hot}) \log_2 \frac{1}{P(T=\text{hot})}$

- $H(T) = H(\underbrace{0.3}_{\text{cold}}, \underbrace{0.5}_{\text{mild}}, \underbrace{0.2}_{\text{hot}}) = 1.48548$ $\quad \underbrace{P(T=\text{cold}) \log_2 \frac{1}{P(T=\text{cold})}}_{0.3 \log_2 \frac{1}{0.3}} + P(T=\text{mild}) \log_2 \frac{1}{P(T=\text{mild})} +$
- $H(M) = H(0.6, 0.4) = 0.970951$
- $\boxed{H(T) + H(M) = 2.456431}$

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

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

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

$H(T,M) = P(T=\text{cold}, M=\text{low}) \log_2 \frac{1}{} + P(T=\text{cold}, M=\text{high}) \log_2 \frac{1}{} + \dots + P(T=\text{hot}, M=\text{high}) \log_2 \frac{1}{}$

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

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

$\boxed{\text{Average}}$ Conditional Entropy

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

$H(T|M) = P(M=low)\, H(T|M=low) + P(M=high)\, H(T|M=high)$

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

$0.6\ H(T|M=low) + 0.4\ H(T|M=high)$

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

cold mild hot

Conditional Entropy:

- $H(T|M=low) = H(1/6, 4/6, 1/6) = 1.25163$ $\quad \frac{1}{6}\log_2 6 + \frac{4}{6}\log_2 \frac{6}{4} + \frac{1}{6}\log_2 6$
- $H(T|M=high) = H(2/4, 1/4, 1/4) = 1.5$
- **Average Conditional Entropy** (aka equivocation):
  $H(T/M) = \sum_m P(M=m) \cdot H(T|M=m) =$
  $0.6 \cdot H(T|M=low) + 0.4 \cdot H(T|M=high) = 1.350978$

Average

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:

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

Mixed setting: Continuous (over x) and discrete (over y):

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

Mutual Information

100 → weigh

dog 30 70 50 cat 20 dg

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

$X = \begin{bmatrix} \; \end{bmatrix}_{n \times d}$ $Y = \begin{bmatrix} \; \end{bmatrix}_{n \times 1}$ label

$$I(X_i, Y) = H(Y) - H(Y|X_i)$$

- 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

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

$$= \sum_x P(x) \cdot \log \frac{1}{P(x)} - \sum_{x,y} P(x,y) \cdot \log \frac{1}{P(x|y)}$$

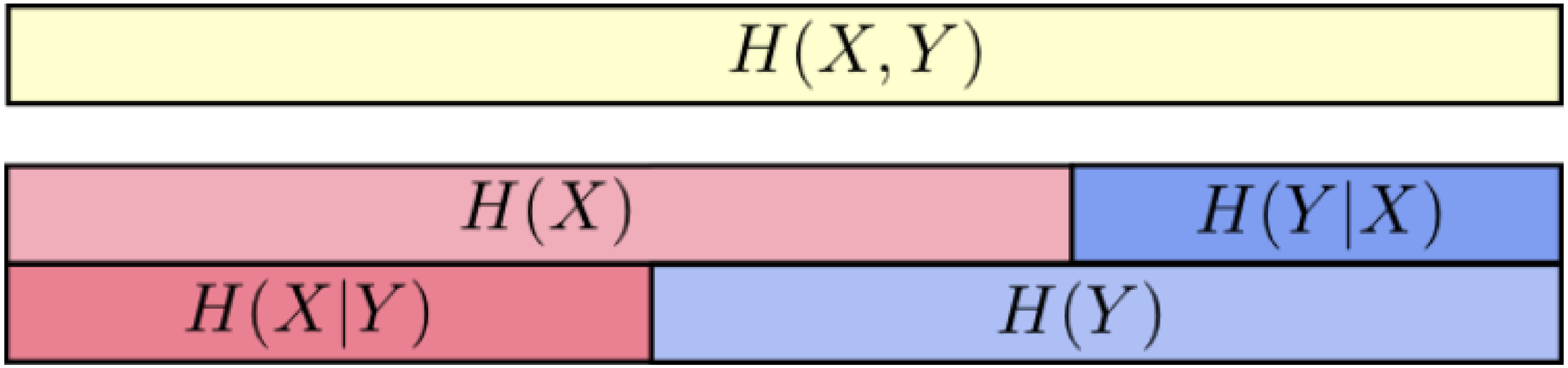
$$= \sum_{x,y} P(x,y) \cdot \log \frac{P(x|y)}{P(x)}$$

$$= \sum_{x,y} P(x,y) \cdot \log \frac{P(x,y)}{P(x)P(y)}$$

Properties of Average Mutual Information:

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

CE and MI: Visual Illustration



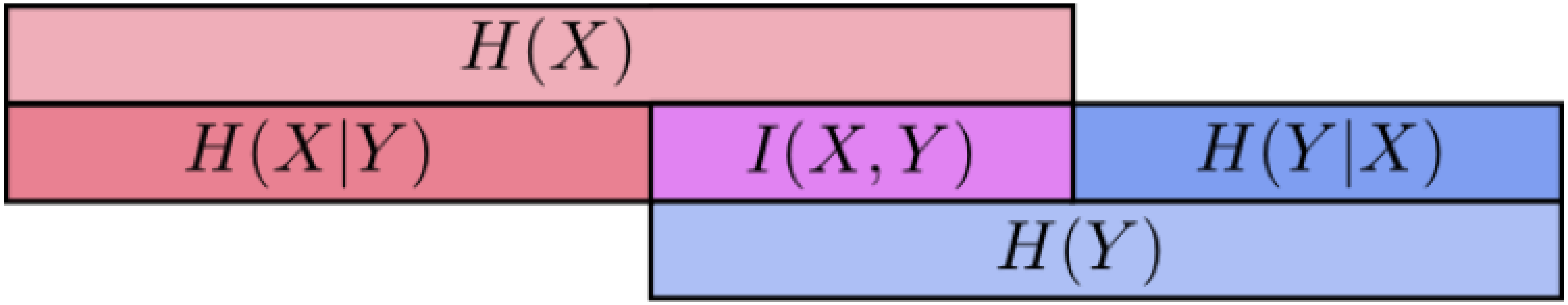


Image Credit: Christopher Olah.

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

True or actual pdf

predicted

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

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?

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

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

By Jensen Inequality

So $\mathbf{KL}[P\|Q] \ge 0$. 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