Probability → sum rule
Probability → Product rule

↓

Bayes's rule $P(x|y) = \dfrac{P(x,y) = P(y|x)\,P(x)}{P(y)}$

↓

Gaussian distribution or normal ⇝ $N(x \mid \mu, \sigma) = \dfrac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

↓

pdf ? Yes

optimize parameters $\theta \in \{\mu, \sigma\}$

$L(\theta|x)$

$L(\mu,\sigma|x)$

MLE → all data points are independent $L(\theta|x) = P(x_1, x_2, \ldots, x_n | \theta)$

$\underset{\sim}{\overset{MLE}{=}} P(x_1|\theta)\,P(x_2|\theta) \cdots P(x_n|\theta)$

$$\boxed{\log L(\Theta|x) = l(\Theta|x)}$$

$$\underset{\Theta}{\operatorname{Argmax}} \; \log L(\Theta|x)$$

x^2  $-x^2$ 

$$\underset{\Theta}{\operatorname{Argmin}} \; -\log L(\Theta|x)$$

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \; \ldots$$

is it for multivarciate or univarciate gaussian?

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi|\Sigma|^2}}$$

covariance $d \times d$

Information Theory

Mahdi Roozbahani
Georgia Tech

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

Outline

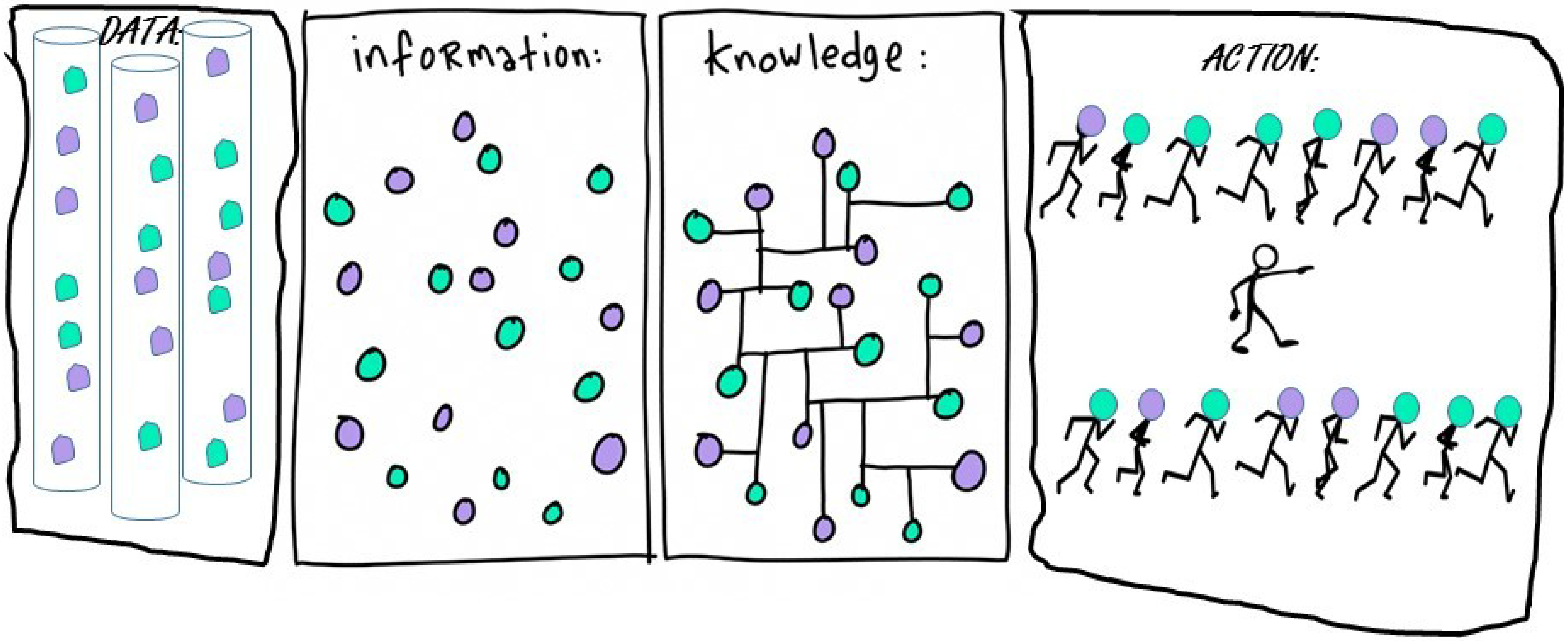
- Motivation
- Entropy
- Conditional Entropy and Mutual Information
- Cross-Entropy and KL-Divergence

Uncertainty and Information

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

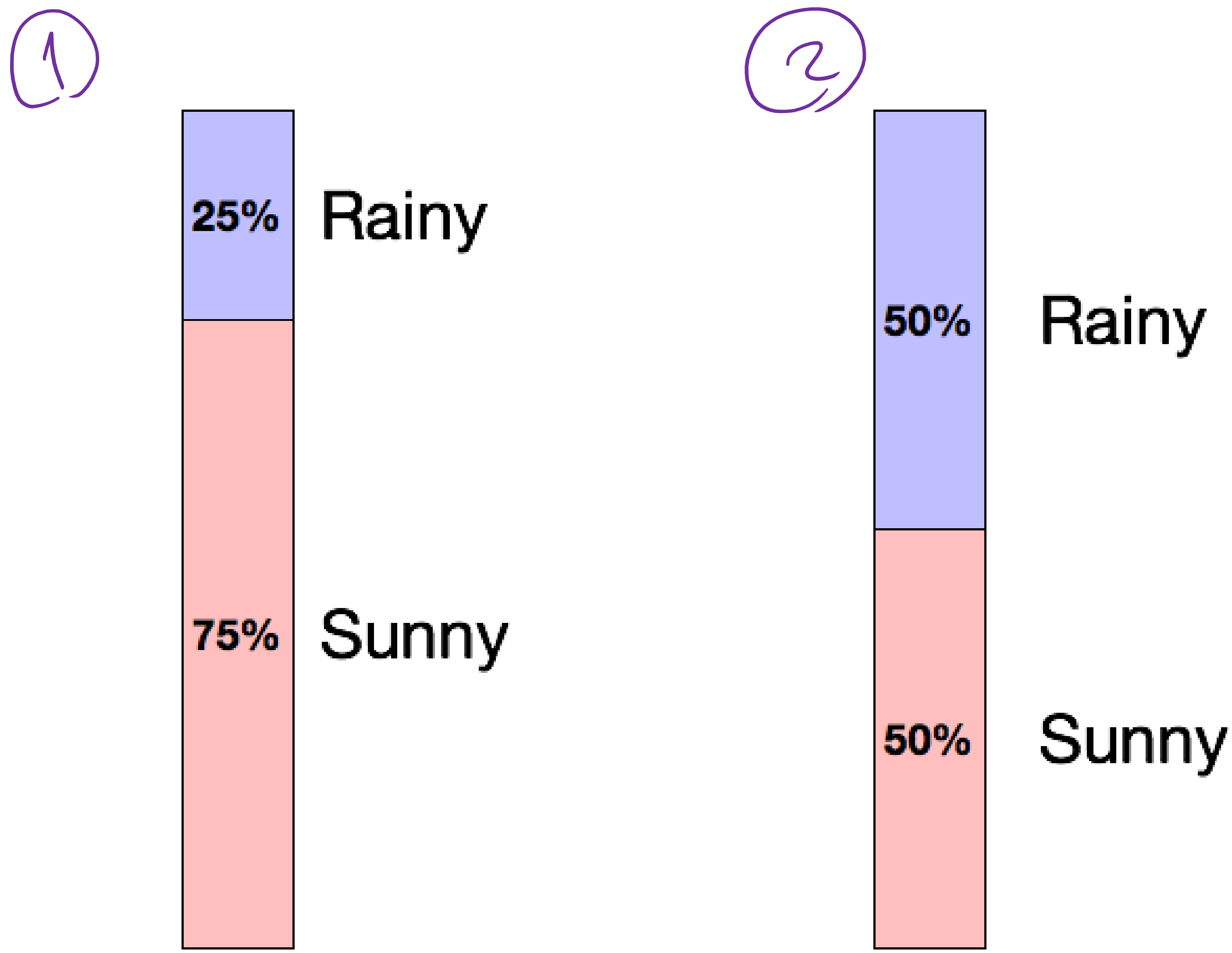
You need **information** to be able to get **knowledge**

- 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."

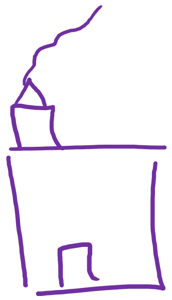
Uncertainty and Information



Which day is more uncertain?

How do we quantify uncertainty?

High entropy correlates to high information or the more uncertain



→ [Cat, cat, cat, dog] ⟶

⟶ cat.

$P(cat) = 1$ $\quad$ $P(cat) = 1$ $\quad$ $P(cat) = 1$ $\quad$ $P(cat) = \frac{3}{4}$ $\quad$ $P(dog) = \frac{1}{4}$

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

at first

$$I(cat) = \log_2 1 = 0$$

$P(cat) = \frac{3}{4}$ $\quad$ $P(dog) = \frac{1}{4}$

$$I(cat) = \log_2 \frac{4}{3}$$

bits of information

$$I(dog) = \log_2 2^2 = 2$$

bits of information

$$E[g(x)] = \sum p(x)\, g(x) \Rightarrow E[I(x)] = \underbrace{H(x)}_{\text{Entropy}} = \sum p(x)\, I(x)$$

① [cat, cat, cat, dog)

$$H(x) = \frac{3}{4}\log_2 \frac{4}{3} + \frac{1}{4}\log_2 4$$

② [cat, cat, dog, dog]

$$H_2 > H_1$$

Pixe [1 0 0 ..... 1] = $2^0 \times 1 + ...$

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:

$$I(document) = 1000 \times I(word) = 16610$$

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

P(one bit in a pixel) = $\frac{1}{16}$

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

$640 \times 480 \times \log_2 16 = 1228800$

$I(X = one\ bit) = ?$ $\log_2 16$

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

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.

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

$I(T) < I(H)$

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

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

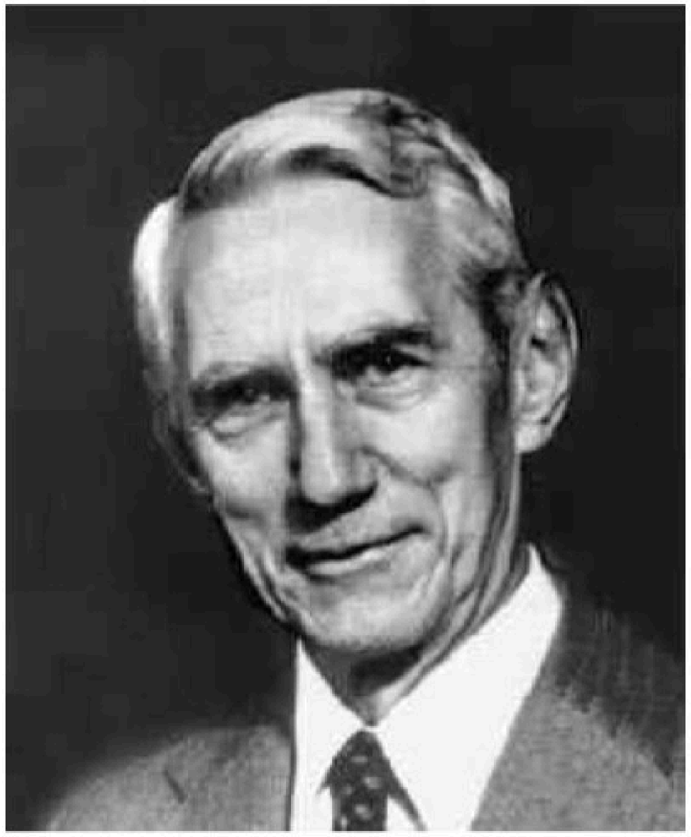
We used 6 characters

MOTIVATION: COMPRESSION

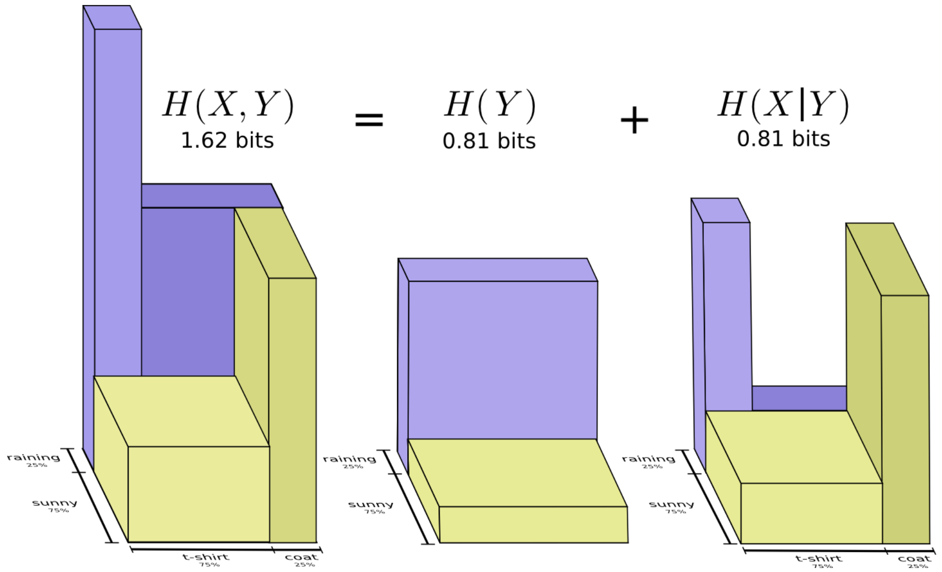
- Frequently occuring events should have short encodings
- We see this in english with words such as “a”, “the”, “and”, etc.
- We want to maximise the information-per-character
- seeing common events provides little information
- seeing uncommon events provides a lot of information

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



Outline

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

Entropy

$H(y) = \sum p(y) \log_2 p(y)^{-1}$

$= - \sum p(y) \log_2 p(y)$

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

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

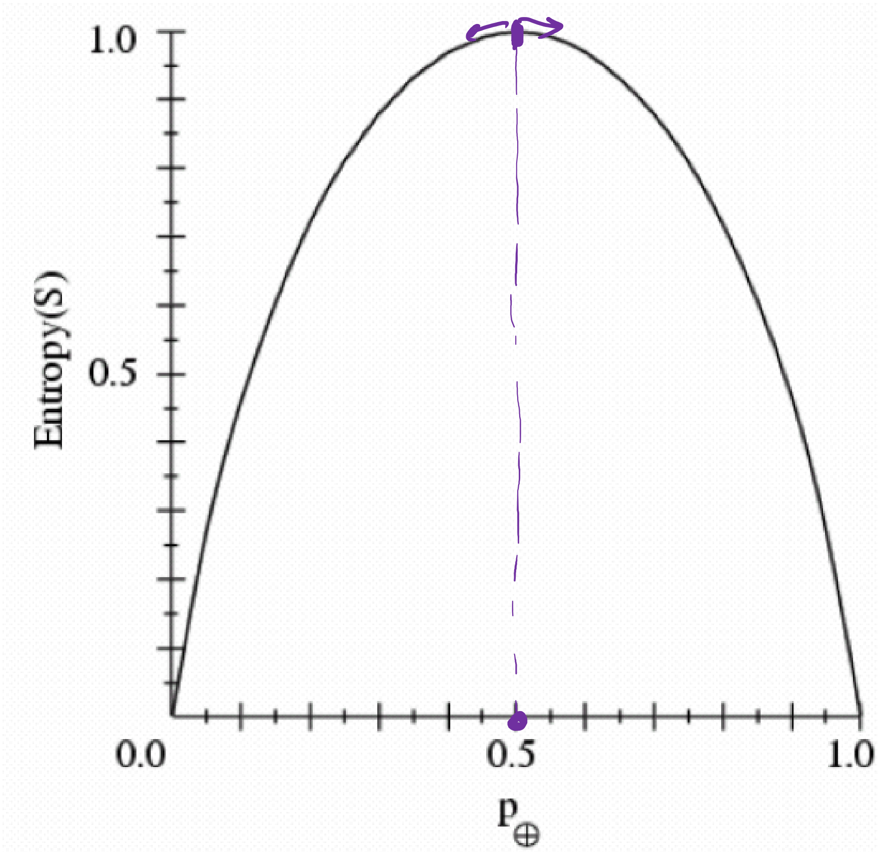
- $H(Y)$ is the expected number of bits needed to encode a randomly drawn value of Y (under most efficient code)

- 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

$H(S) = -P(H)\log_2 P(H) - P(T)\log_2 P(T)$

- 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

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

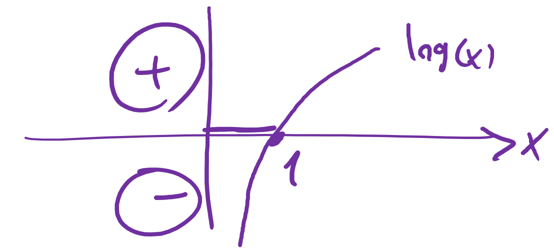
p_i

Actual dog $\rightsquigarrow [1 \quad 0 \quad 0]$ $[0.8 \quad 0.1 \quad 0.1]$ q_i predicted $\Rightarrow [1 \quad 0 \quad 0]$

cat $\rightsquigarrow [0 \quad 1 \quad 0]$

fish $\rightsquigarrow [0 \quad 0 \quad 1]$

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



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

$\sum p_i \left(\log \frac{1}{p_i} - \log \frac{1}{q_i}\right) < 0$

2. Invariant wrt permutation of its inputs: $\sum p_i \log \frac{q_i}{p_i} < 0$

   $H(p_1, p_2, \ldots, p_k) = H(p_{\tau(1)}, p_{\tau(2)}, \ldots, p_{\tau(k)})$

   $0 < \quad < 1$

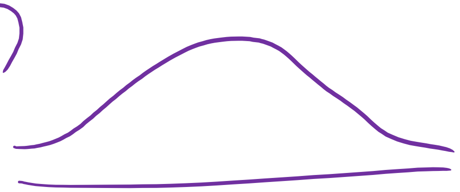
3. For any *other* probability distribution $\{q_1, q_2, \ldots, q_k\}$:

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

actual distribution predicted one

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.

1   2

Outline

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

Joint Entropy

$\log(T{=}\text{cold}) = \log_2 \frac{1}{0.3}$

$P(T = \text{cold}) = 0.3$

$P(T = \text{mild}) = 0.5$

$P(T = \text{hot}) = 0.2$

huMidity / Temperature

	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

- $H(T) = H(0.3, 0.5, 0.2) = 1.48548$ $\quad H = \underset{\text{cold}}{0.3 \log_2 \frac{1}{0.3}} + \underset{\text{mild}}{0.5 \times \log_2 \frac{1}{0.5}} + \underset{\text{hot}}{0.2 \log_2 \frac{1}{0.2}}$
- $H(M) = H(0.6, 0.4) = 0.970951$
- $H(T) + H(M) = 2.456431$
- **Joint Entropy**: consider the space of (t, m) events $H(T, M) = \sum_{t,m} P(T = t, M = m) \cdot \log \frac{1}{P(T=t, M=m)}$
  $H(0.1, 0.4, 0.1, 0.2, 0.1, 0.1) = 2.32193$ $\quad 0.1 \log_2 \frac{1}{0.1} + \cdots + 0.1 \log_2 \frac{1}{0.1}$

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

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

Conditional Entropy

$\sum p(x)\, g(x)$

average conditional Entropy

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

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

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

average conditional entropy $= 0.6 \times 1.25163 + 0.4 \times 1.5$

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 uncerntainty in Y after seeing feature X_i

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

- 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

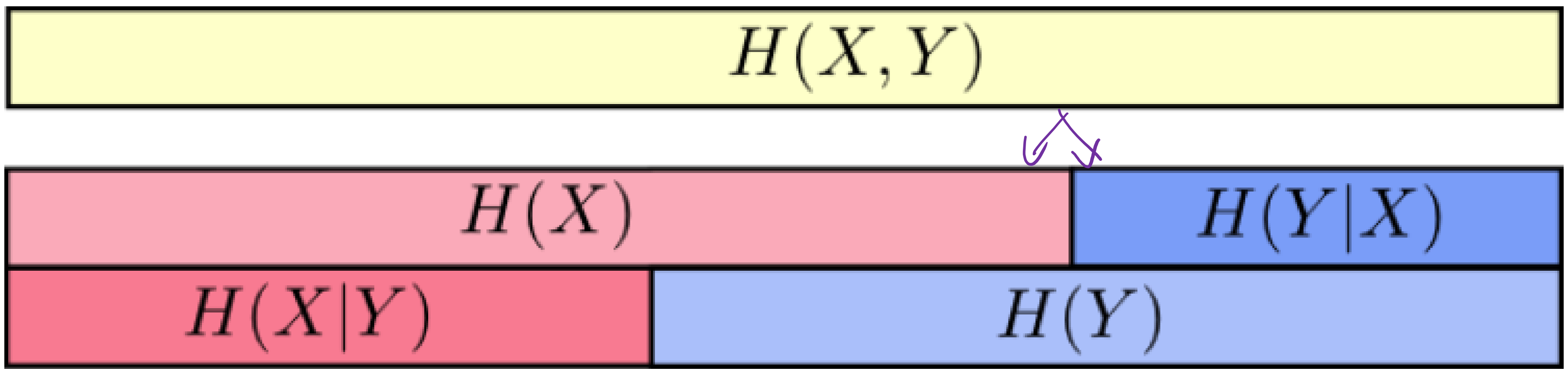
$$
\begin{aligned}
I(X,Y) = H(X) - H(X|Y)& \\
&= \sum_x P(x) \cdot \log \frac{1}{P(x)} - \sum_{x,y} P(x,y) \cdot \log \frac{1}{P(x|y)} \\
&= \sum_{x,y} P(x,y) \cdot \log \frac{P(x|y)}{P(x)} \\
&= \boxed{\sum_{x,y} P(x,y) \cdot \log \frac{P(x,y)}{P(x)P(y)}}
\end{aligned}
$$

Properties of Average Mutual Information:

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

CE and MI: Visual Illustration

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



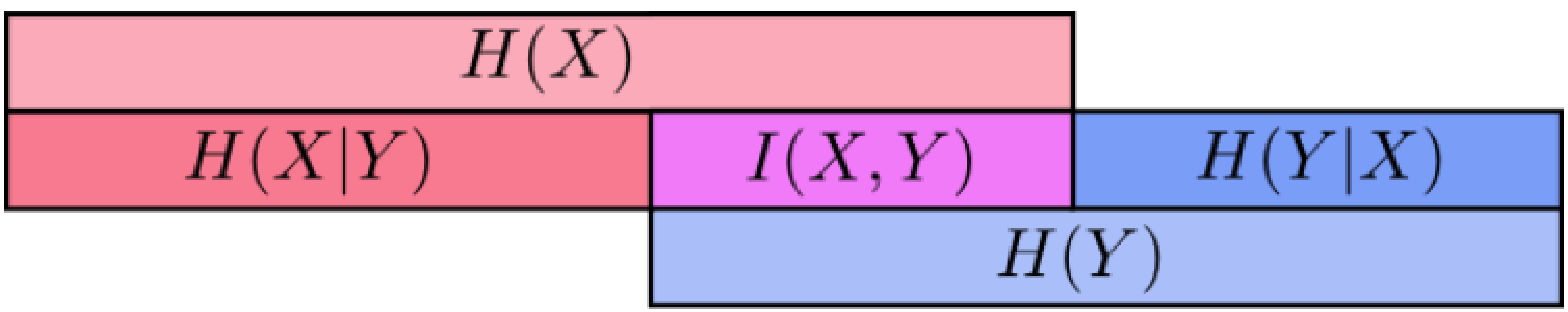


Image Credit: Christopher Olah.

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

actual Cat [1 0 0]

predicted cat [0.8 0.1 0.1]

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

actual pdf

predicted

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

↓

loss function → need to minimize this

$$E[g(x)] = \sum p(x) \log \frac{1}{q(x)}$$

$g(x)$

This is because:

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

$$= E\left[\log \frac{1}{q(x)}\right]$$

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

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

Kullback-Leibler Divergence

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

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

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

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

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

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

log function is concave or convex?

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

By Jensen Inequality

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

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