Recap

- 3 rules in probability
- Expectation and Arithmetic Mean
- Covariance and Correlation
- Probability vs Likelihood
- MLE

Georgia Tech

Machine Learning CS 4641

Information Theory

Nimisha Roy

Lecturer, SCI, College of Computing, Georgia Tech

Director of Online Undergradute Initiatives

Slides adapted from Max Roozbahani, Le Song , Roni Rosenfeld, Maneesh Sahani, and Chao Zhang.

Georgia Tech

Outline

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

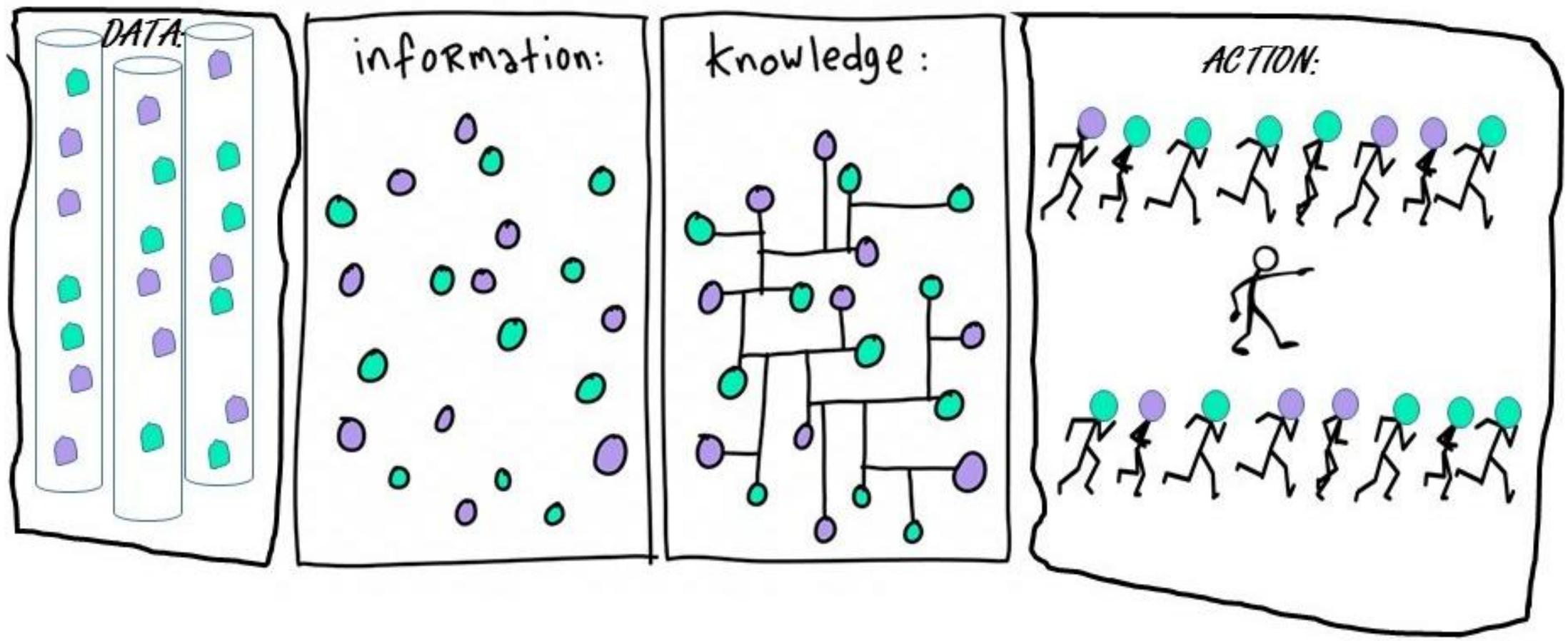
Uncertainty and Information

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

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

- **More information when an unlikely event occurs than when something certain occurs** (in fact, it should be zero when the event is certain)

- **Example:** You are in sunny Los Angeles, California and you are told it did not rain yesterday → **not a lot of information since it rarely rains in SoCal**

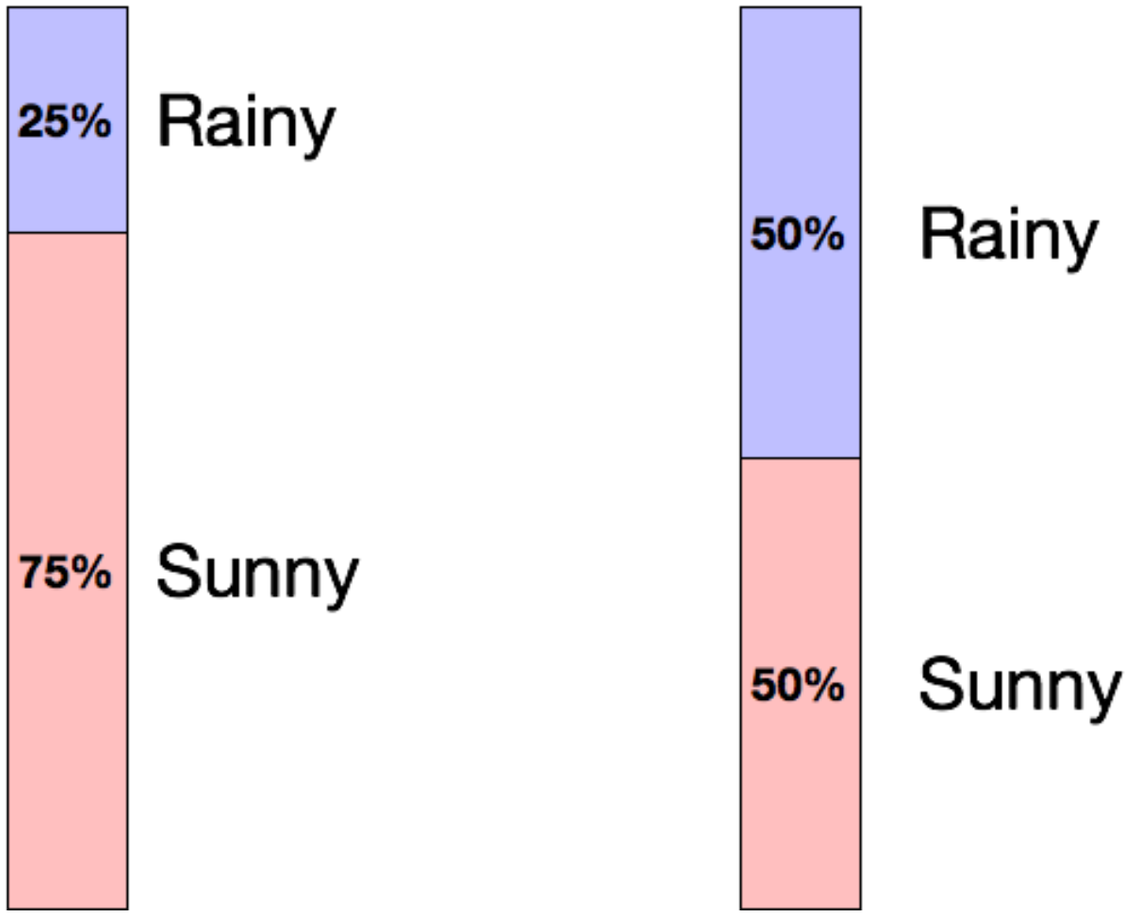


Created by Bruce Campbell: “DIKA – ancient Chinese saying for get up and DO! Data-Information-Knowledge-Action.”

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 can we encode more efficiently?
  - Is the **information carried by different random variables complementary or redundant**?

Uncertainty and Information



Which day is more uncertain?

How do we quantify uncertainty?

Less likely → More uncertainty → More information → High entropy

Information of an Outcome

Rain in Atlanta in July

P(x) = 0.8

Snow in Atlanta in July

P(x) = 0.05

Which event is more likely?

Which will give more information if it happens?

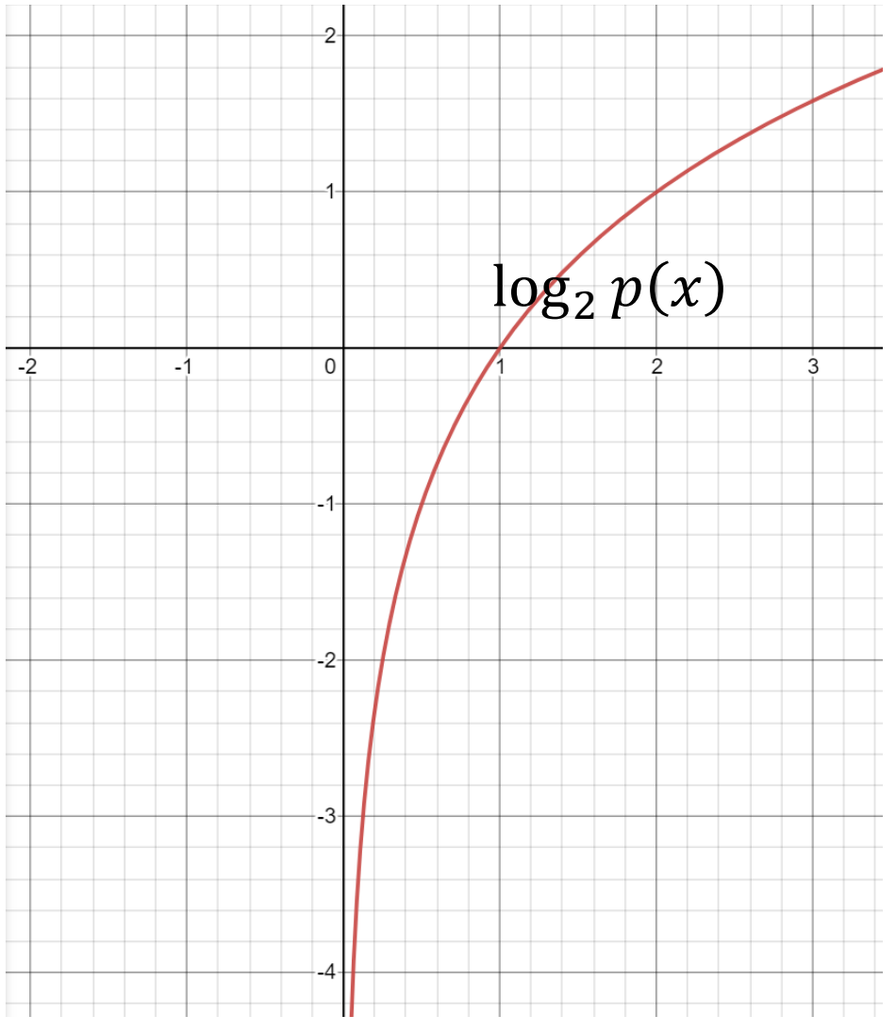
More likely → Less uncertainty → Less information?

Georgia Tech

Information

We can associate our measure of information with probability of an event occurring. Let X be a random variable with distribution $p(x)$:

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



High probability means less information

Information: Example

X = [cat,]

Example: is a picture worth 1,000 words?

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

$$I(word) = \log_2\left(\frac{1}{p(x)}\right) = \log_2\left(\frac{1}{1/100{,}000}\right) = 16.61\ bits$$

- A 1,000-word document from same source:

$$I(document) = 1000 \times I(word) = 16{,}610\ bits$$

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

$$I(picture) = \log_2\left(\frac{1}{1/16^{640\times480}}\right) = 1{,}228{,}800\ bits$$

A picture is worth (a lot more than) 1,000 words!

Motivation: Data 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 the information
- Another way to think about this**: how much information does the sequence of events actually contain?**

Example: Data compression

- Consider the problem of recording coin tosses in unary

T, T, T, T, H

- **Approach 1:**

Heads	Tails
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?

Example: Data compression

- Consider the problem of recording coin tosses in unary

T, T, T, T, H

- **Approach 2:**

Heads	Tails
00	0

0, 0, 0, 0, 00

We used **6** characters

- Which one has a higher probability: T or H?
- Which one should carry more information: T or H?

Motivation: Data compression

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

Information of an Outcome

Rain in Atlanta in July

P(x) = 0.8

Snow in Atlanta in July

P(x) = 0.05

Which event is more likely?

Which will give more information if it happens?

More likely → Less uncertainty → Less information if it happens

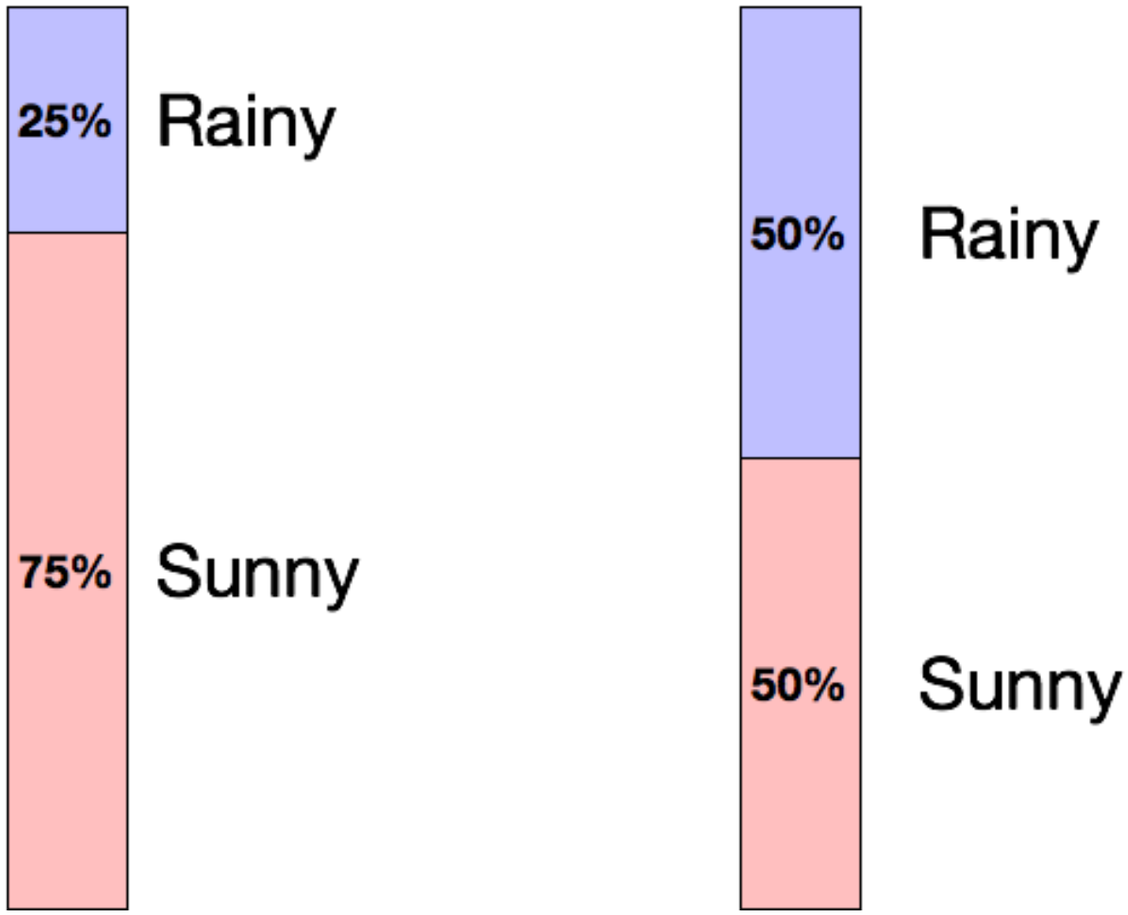
Georgia Tech

Outline

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

- Information measured *after an outcome occurs*.
- Sometimes, we care about how uncertain a system is *before* we observe anything

Uncertainty and Entropy



Which day is more uncertain?

How do we quantify uncertainty?

Now we are not asking which outcome is more informative. We are asking: which day is harder to predict?

Georgia Tech

Entropy

- Average amount of information to encode a random variable X with respect to its distribution $p(x)$ is the entropy $H(x)$:

$$E[g(x)] = \sum_x g(x)p(x)$$

$$H(x) = E[I(x)] = \sum_x I(x)p(x) = \sum_x p(x)\log_2\left(\frac{1}{p(x)}\right) = -\sum_x p(x)\log_2 p(x)$$

Considering a random variable X with k possible states:

$$H(x) = -\sum_{k=1}^{K} p(x=k)\log_2 p(x=k) = \sum_{k=1}^{K} p(x=k)\log_2 \frac{1}{p(x=k)}$$

- Entropy is the **average amount of surprise (information)** you gain when observing outcomes of X.

Entropy

- Max possible entropy a random variable X with k possible states is $log_2 k$

- Entropy is non-negative

- Most efficient code assigns $-\log_2 P(x = k)$ bits to encode the message $x = k$.
  - because it matches the information content of that event.

Example: entropy computation for coin toss

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

Head	0
Tail	6

$$p(H) = \frac{0}{6} = 0, \qquad p(T) = \frac{6}{6} = 1$$
$$H = -0 \log_2 0 - 1 \log_2 1 = 0$$

Head	1
Tail	5

$$p(H) = \frac{1}{6}, \qquad p(T) = \frac{5}{6}$$
$$H = -\frac{1}{6}\log_2 \frac{1}{6} - 5/6 \log_2 \frac{5}{6} = 0.65$$

Head	2
Tail	4

$$p(H) = \frac{2}{6}, \qquad p(T) = \frac{4}{6}$$
$$H = -\frac{2}{6}\log_2 \frac{2}{6} - 4/6 \log_2 \frac{4}{6} = 0.92$$

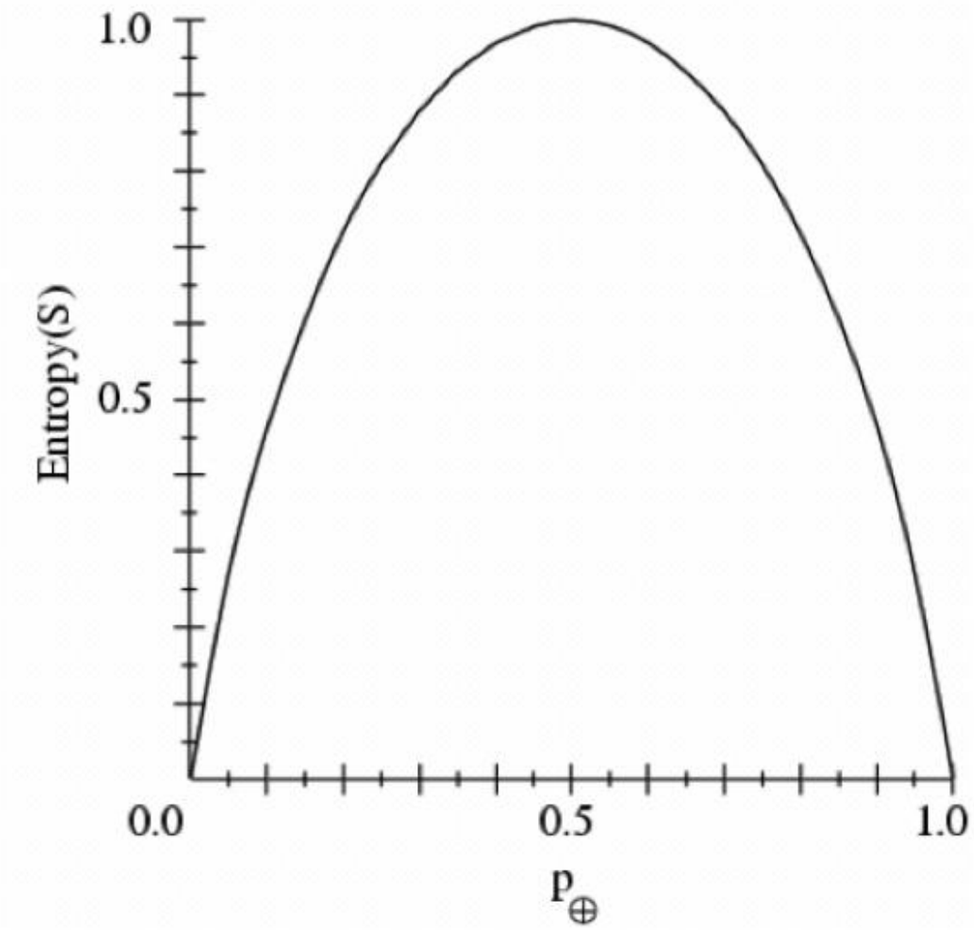
Head	3
Tail	3

$$p(H) = \frac{1}{2}, \qquad p(T) = \frac{1}{2}$$
$$H = -\frac{1}{2}\log_2 \frac{1}{2} - 1/2 \log_2 \frac{1}{2} = 1$$

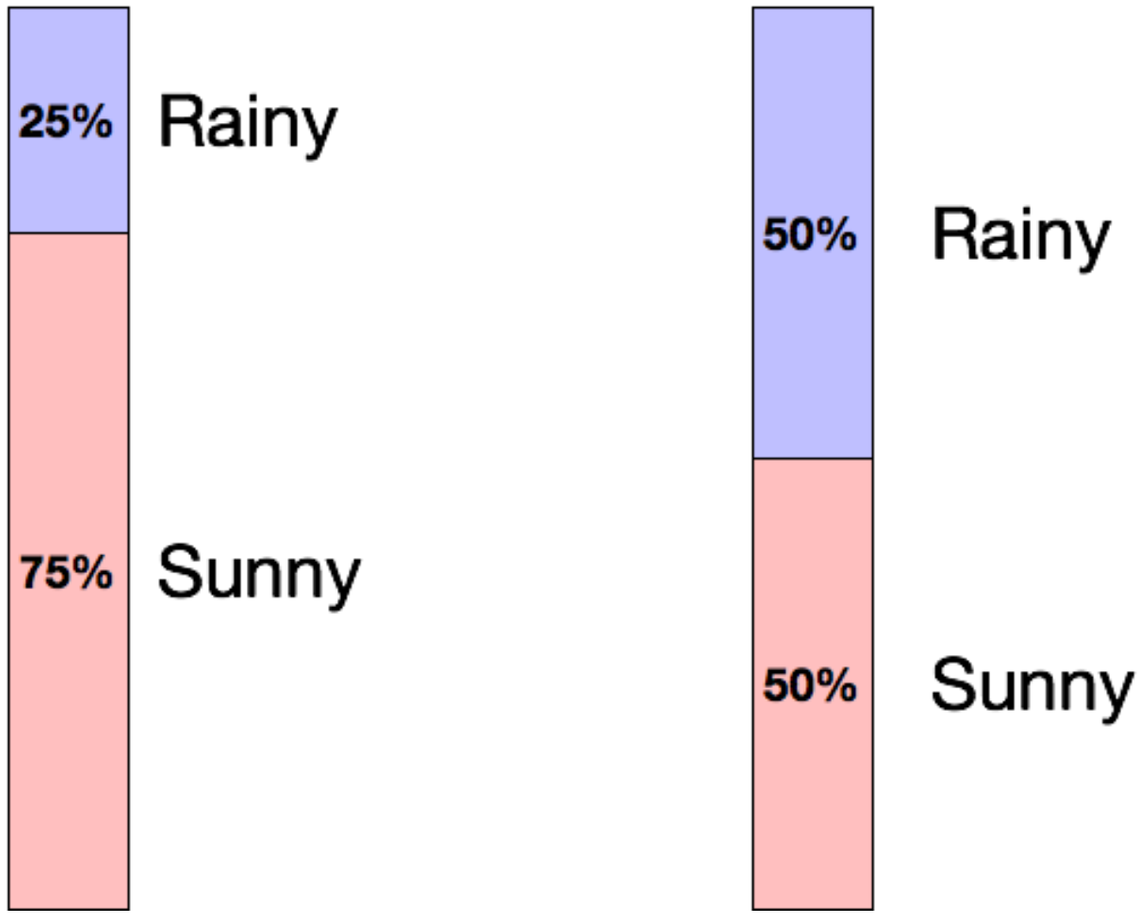
Example: entropy

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

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



Uncertainty and Entropy



Which day is more uncertain?

How do we quantify uncertainty?

Less likely → More uncertainty → More information → High entropy

Properties of Entropy

- Non-negative: $H(P) \geq 0$
- Invariant with respect to permutation of its inputs:

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

- $H(P) \leq \log_2 \mathrm{k}$, with equality iff $p_i = \frac{1}{\mathrm{k}}, \forall i$

- The further P is from uniform, **the lower the entropy**

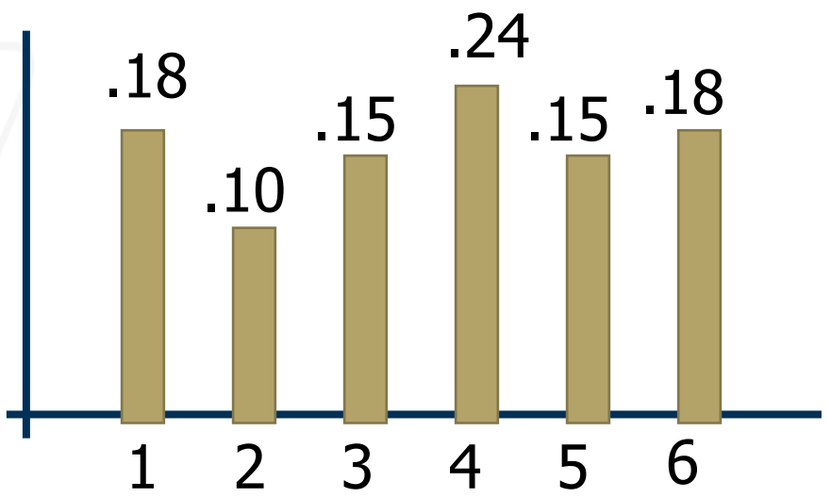
Properties of Entropy

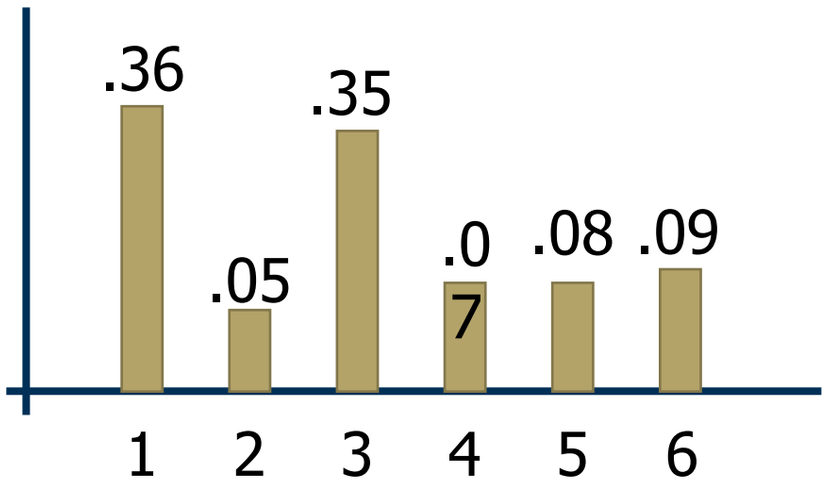
- For any other probability distribution $\{q_1, q_2, \ldots, q_k\}$

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

Example: entropy for rolling a die

$$H(x) = -\sum_{k=1}^{K} p(x = k) \log_2 p(x = k)$$





In context: Why does Entropy matter?

Rare outcomes have high information, and systems with balanced probabilities have high entropy.

Entropy measures uncertainty.

- It tells us how much surprise to expect on average.
- High entropy = we should expect to learn more once the outcome is revealed.

Entropy helps us compare systems

- Coin that always lands heads → entropy = 0 (no uncertainty, no info).
- Fair coin → entropy = 1 bit (max for binary outcomes).
- Six-sided die → entropy ≈ 2.58 bits (more uncertainty than a coin).

👉 **Why helpful?**

- It lets us compare “information richness” between random processes.
- Used in feature selection. Features with higher entropy have more capacity to carry information.

Big picture intuition

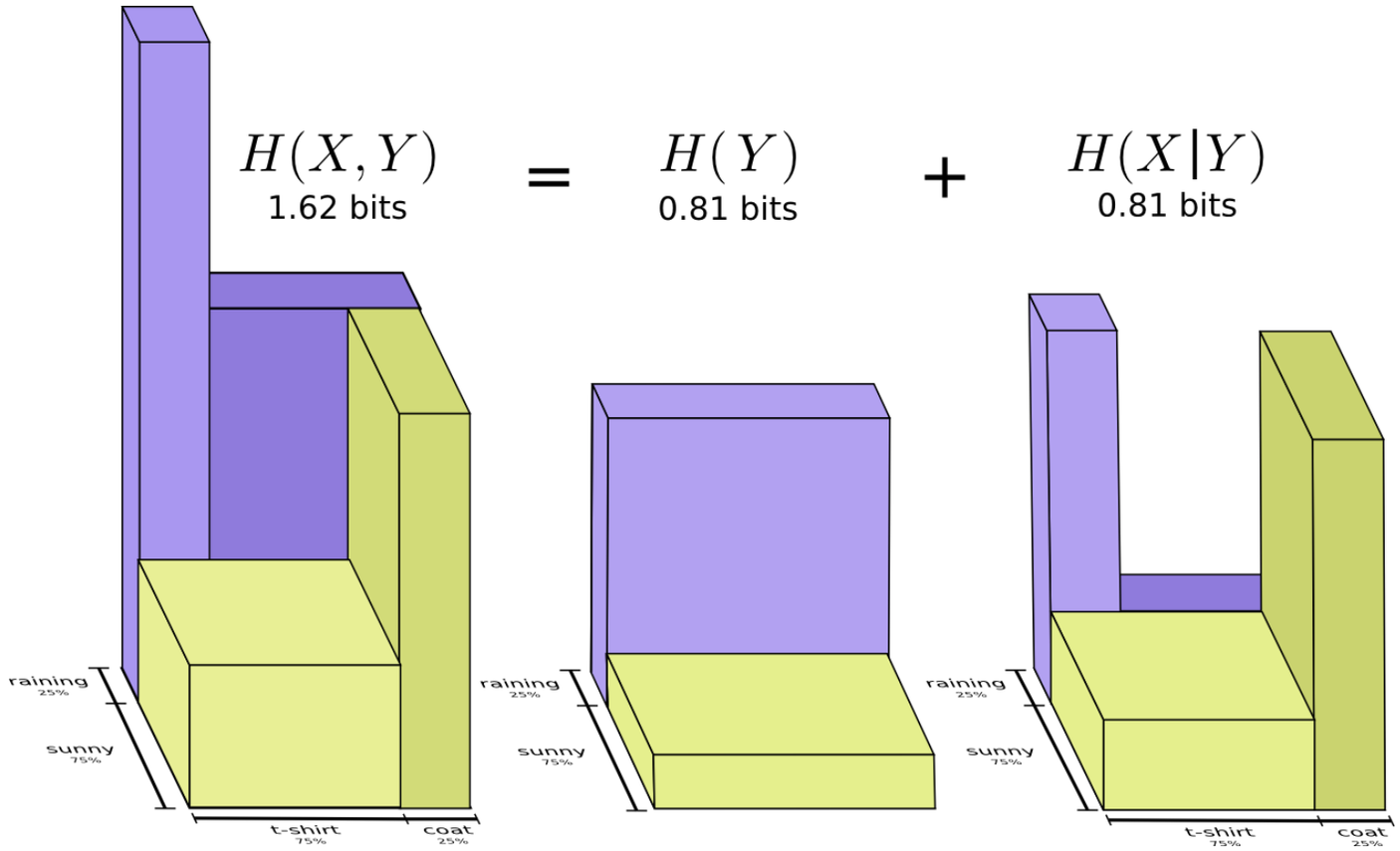
- Entropy matters because it tells us how much “space” we need to represent uncertainty.
- The more uncertain the world is, the more information we gain by observing it — and entropy quantifies that.

Georgia Tech

Outline

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

Entropy


$H(X,Y)$
1.62 bits
=
$H(Y)$
0.81 bits
+
$H(X|Y)$
0.81 bits
raining
25%
sunny
75%
t-shirt
75%
coat
25%
raining
25%
sunny
75%
raining
25%
sunny
75%
t-shirt
75%
coat
25%

Joint Entropy

- The joint entropy measures the entropy of a joint probability distribution of two random variables X and Y

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

- It helps us quantify the **dependence** between different variables and how much information they share

- From the product rule we can also see that

$$H(X,Y) = H(X) + H(Y|X)$$

Joint Entropy

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

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

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

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

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

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

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:

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

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

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

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

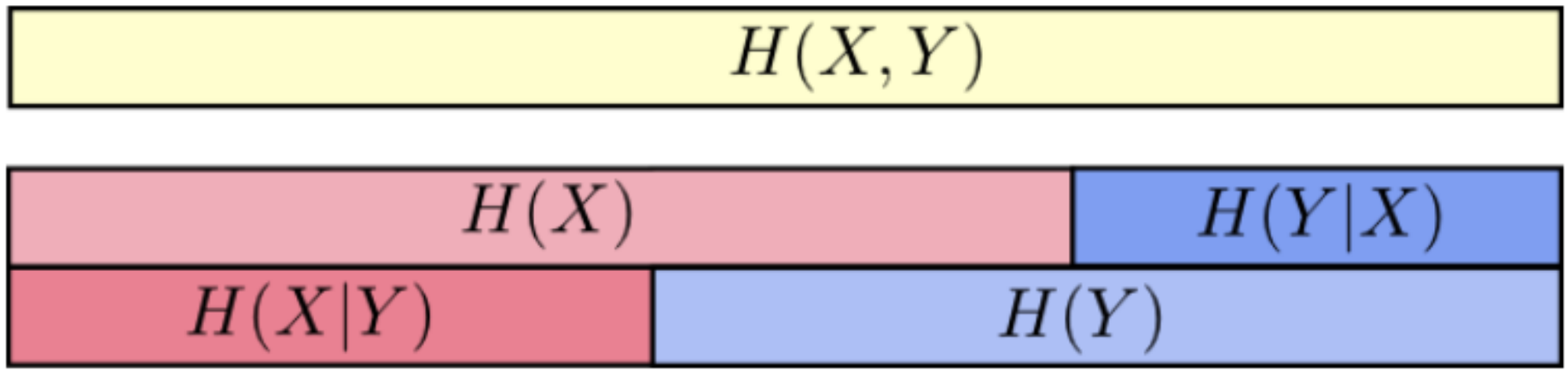
Properties of Mutual Information

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

Properties of Average Mutual Information:

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

CE and MI: Visual Illustration



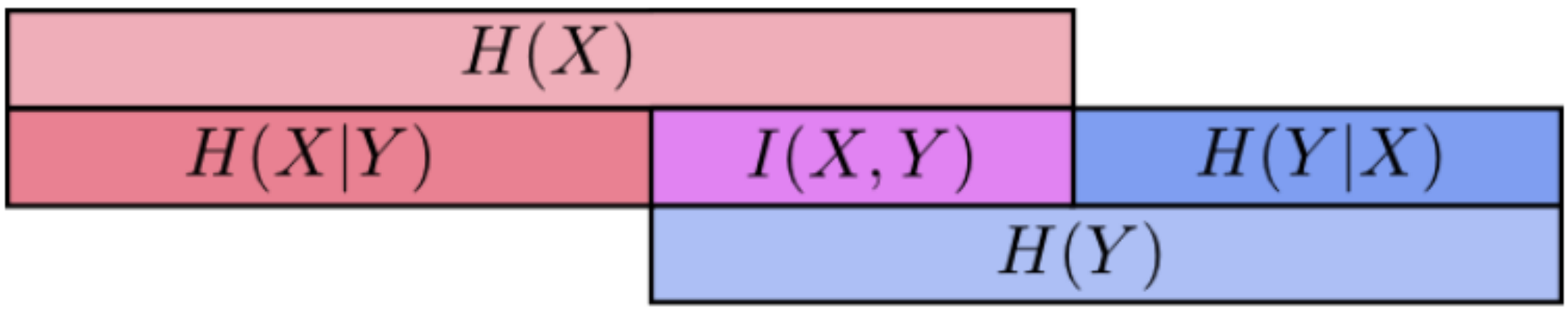


Image Credit: Christopher Olah.

Outline

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

We will cover this next class in our Optimization lecture

Cross Entropy

Suppose the **true distribution** of data is $P(x)$.

But you assume or model the data using another distribution $Q(x)$.

Cross entropy measures the *average number of bits required* to encode samples from P, **if you use a code optimized for** Q **instead of** P.

👉 In short: *How inefficient is your code when you use the wrong distribution?*

Entropy Summary

- Entropy (H(X)) is the average number of bits needed to encode the random variable X when using an optimal code matched to its distribution $p(x)$.

- Joint Entropy (H(X,Y)) is the average number of bits needed to encode both X and Y together, if you use an optimal code matched to their joint distribution.

- Conditional Entropy (H(Y|X)) is the average number of extra bits needed to encode Y, once you already know X.

- Cross Entropy is the average number of bits needed to encode data that really follows distribution P, if you instead use a code optimized for another distribution Q.

Cross Entropy

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

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

H(P): the true entropy (best-case cost, if you used the right distribution).

KL[P][Q]: the extra penalty (inefficiency) you pay for using the wrong distribution Q.

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.

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

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

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?

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

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
  - High uncertainty maps to High information and High entropy
  - 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