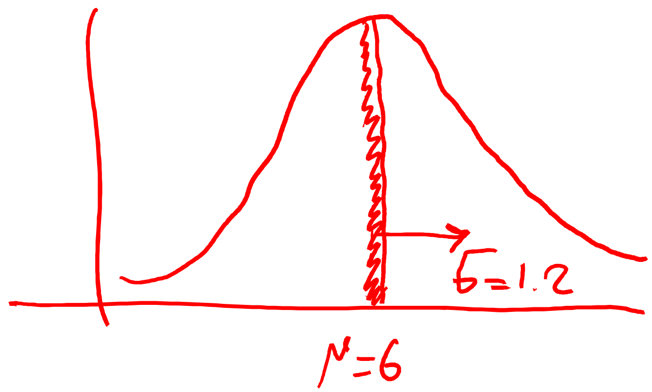
① Sum Rule: $P(x) = \sum_{y} P(x,y)$

② product Rule: $P(x,y) = P(x|y)\,P(y) = P(y|x)\,P(x)$

$\mu = 6 \quad \sigma = 1.2$

①

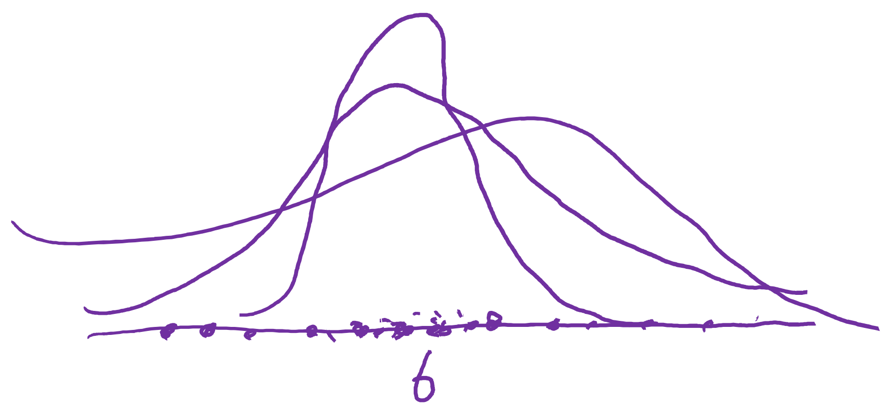
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



$P(X = x \mid \sigma, \mu) = $ —

---

②



$$\boxed{L(\sigma,\mu;X) = f(x_1,\dots,x_{300})}$$

$$\max_{\sigma,\mu} (\sigma,\mu;X) \rightsquigarrow \text{the best value for } \sigma,\theta$$

"i.i.d" $\Rightarrow$ $f(x) = f(x_1)\, f(x_2) \cdots f(x_{n=300})$

$$\max_{\sigma,\mu} L(\sigma,\mu:X) = \prod_{i=1}^{n} f(x_i)$$

$$\cancel{P(x|y) = \frac{\overbrace{P(y|x)}^{\text{likelihood}}\, \underbrace{P(x)}_{\text{prior}}}{P(y)}}$$

$$\to \cancel{P(x=6 \mid y=BP) = \frac{P(y=BP|x)\,P(x)}{P(y)}}$$

$$P(Y|X) = \frac{P(X|Y)\,P(Y)}{P(X)}$$

Lecture 04
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

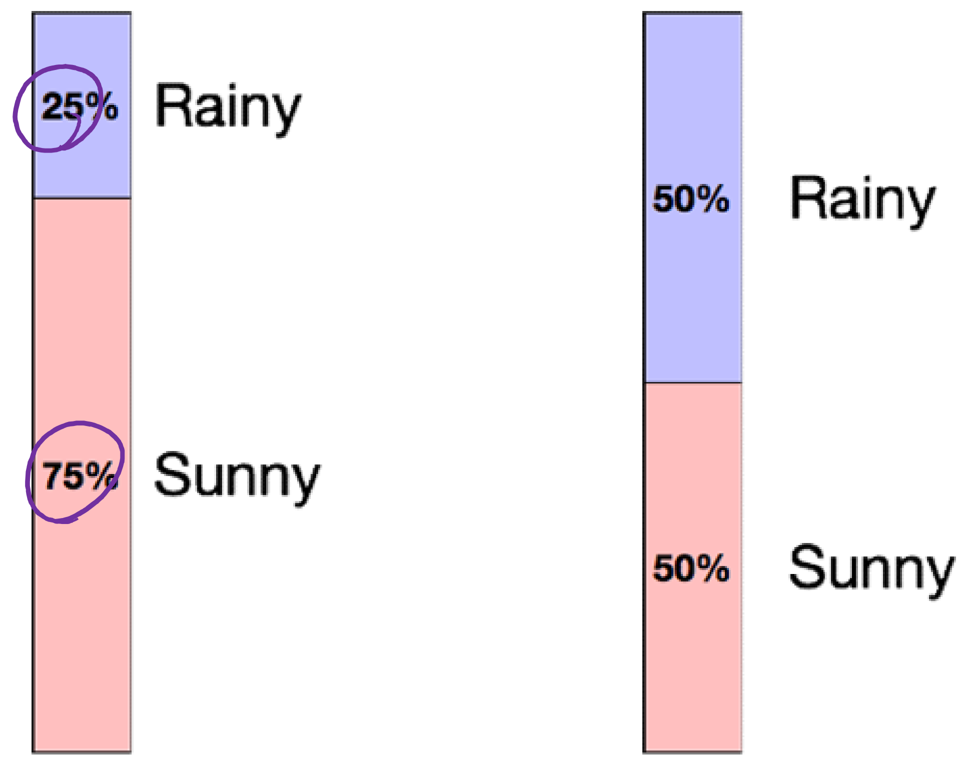
Uncertainty and Information

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

Uncertainty and Information



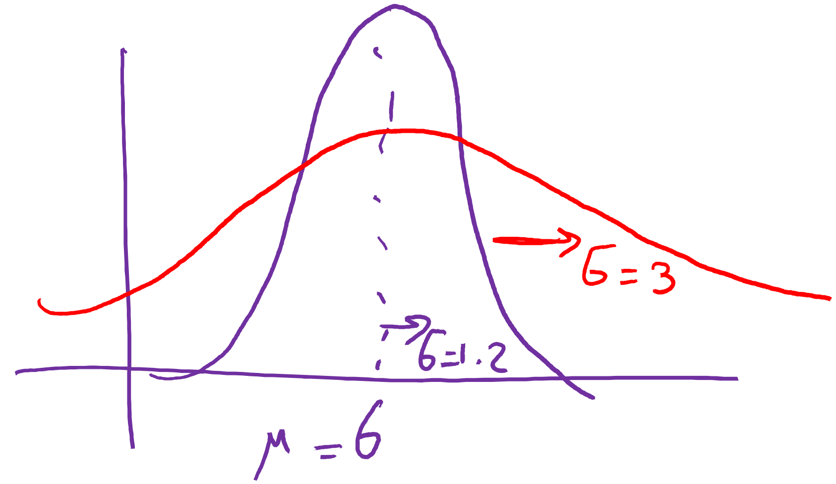
Which day is more uncertain?

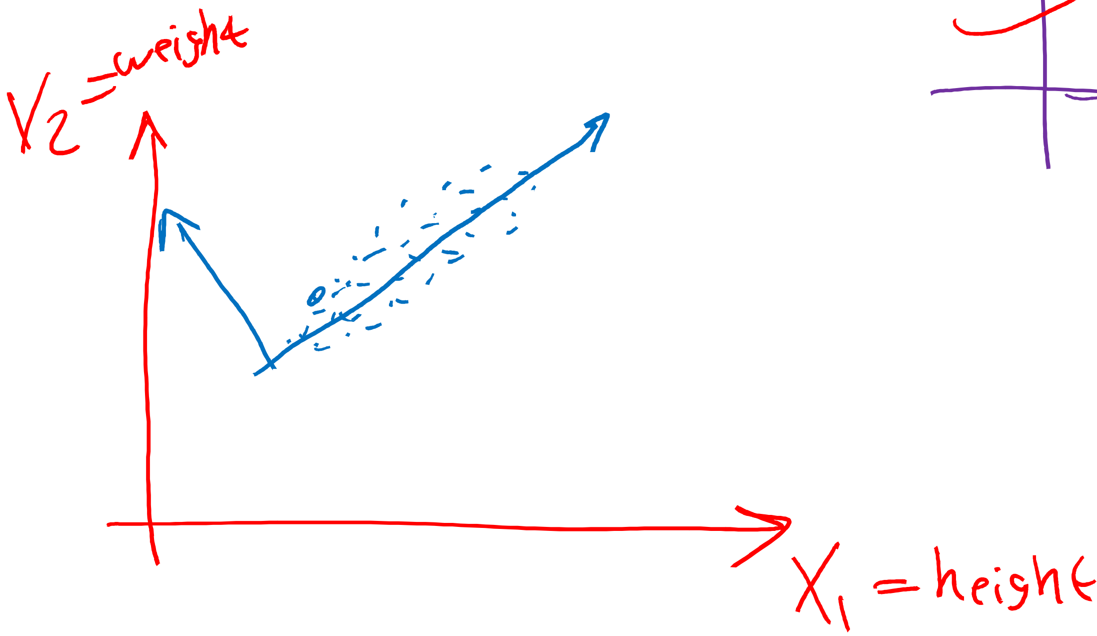
How do we quantify uncertainty?

High entropy correlates to high information or the more uncertain

1 2 → 1 2 3

50% 33%





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:

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

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):

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

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.

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

Approach 1:

H	T
0	00

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

We used 9 characters

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

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

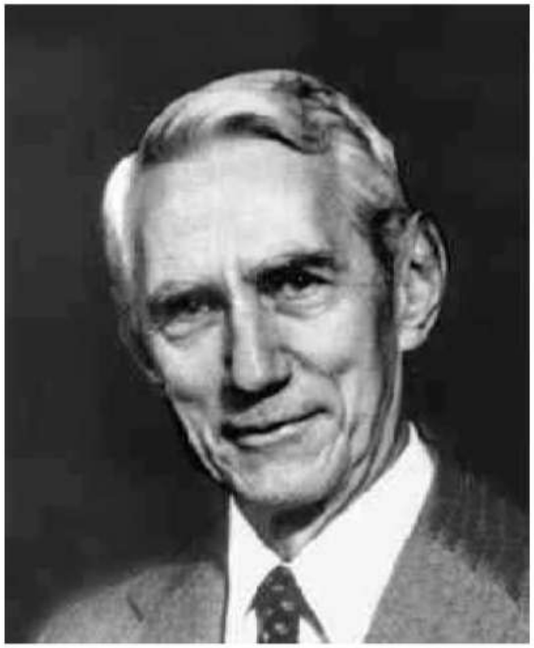
$\boxed{\text{We used 6 characters}}$

Motivation: Compression

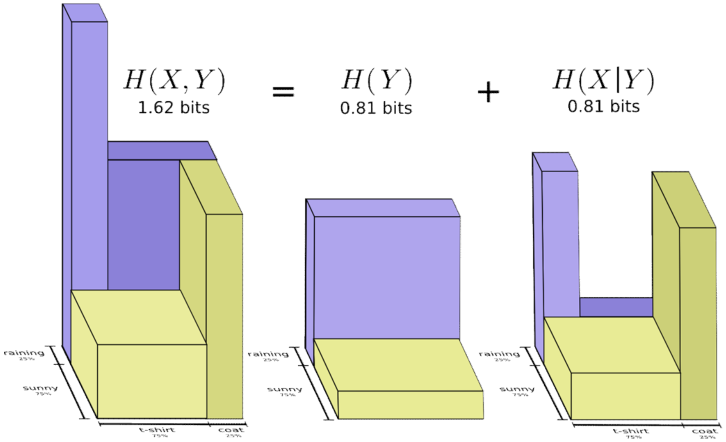
- Frequently occuring events should have short encodings
- We see this in english with words such as "a", "the", "and", etc.
- We want to maximise the information-per-character
- seeing common events provides little information
- seeing uncommon events provides a lot of information

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

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

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

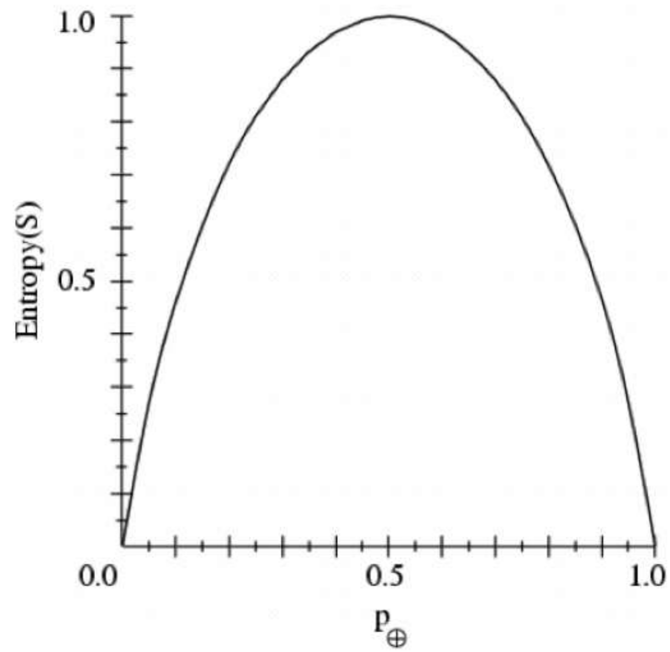
- $H(Y)$ is the expected number of bits needed to encode a randomly drawn value of Y (under most efficient code)

- 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

head	1
tail	5

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

Entropy $= \underbrace{-\,(1/6)\log_2(1/6)}_{\text{head}} \underbrace{-\,(5/6)\log_2(5/6)}_{\text{tail}} = 0.65$

head	2
tail	4

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

Entropy $= -\,(2/6)\log_2(2/6) - (4/6)\log_2(4/6) = \underline{0.92}$

3
3

= 1

Properties of Entropy

$E(x) = \sum p(x) g(x)$

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

$p_i = \frac{1}{k} \Rightarrow \sum_k \frac{1}{k} \log \frac{1}{\frac{1}{k}} =$

$= \cancel{k} \left(\frac{1}{\cancel{k}}\right) \log k$

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 $\{q_1, q_2, \ldots, q_k\}$:
   $$H(P) = \sum_i p_i \cdot \log \frac{1}{p_i} < \sum_i p_i \cdot \log \frac{1}{q_i}$$
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

Joint Entropy

$P(T) = \sum_{M} P(T,M)$

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

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

$0.3 \log \frac{1}{0.3} + 0.5 \log \frac{1}{0.5} \ldots$

- $H(T) = H(0.3, 0.5, 0.2) = 1.48548$
- $H(M) = H(0.6, 0.4) = 0.970951$
- $H(T) + H(M) = \boxed{2.456431}$
- ✓ **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 $H(T, M) < H(T) + H(M)$!!!

Conditional Entropy

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

	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):
  $H(T/M) = \sum_m P(M = m) \cdot \boxed{H(T|M = m)} =$
  $0.6 \cdot H(T|M = low) + 0.4 \cdot H(T|M = high) = 1.350978$

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:

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

Continuous:

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

- Quantify the uncerntainty in Y after seeing feature X_i

- $H(Y)$ is the expected number of bits needed to encode a randomly drawn value of Y
  - given X_i, and
  - average over the likelihood of seeing particular value of x_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
  - $I(X_i, Y) = I(Y, X_i) = H(X_i) - H(X_i|Y)$
  - $I(Y, X_i) = \int \sum_k^K p(x_i, y = k) \log_2 \frac{p(x_i, y=k)}{p(x_i)p(y=k)} \ dx_i$
  - $= \int \sum_k^K p(x_i|y = k)p(y = k) \log_2 \frac{p(x_i|y = k)}{p(x_i)} \ dx_i$

Properties of Mutual Information

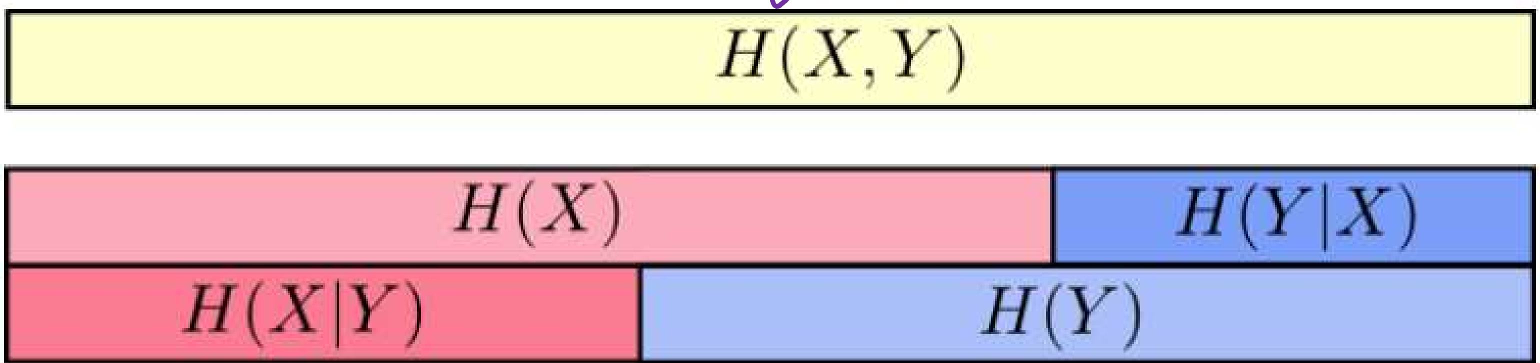
$$
\begin{aligned}
I(X;Y) &= H(X) - H(X/Y) \\
&= \sum_x P(x) \cdot \log \frac{1}{P(x)} - \sum_{x,y} P(x,y) \cdot \log \frac{1}{P(x|y)} \\
&= \sum_{x,y} P(x,y) \cdot \log \frac{P(x|y)}{P(x)} \\
&= \boxed{\sum_{x,y} \underline{P(x,y)} \cdot \log \frac{P(x,y)}{P(x)P(y)}}
\end{aligned}
$$

Properties of Average Mutual Information:

- Symmetric (but $H(X) \neq H(Y)$ and $H(X/Y) \neq H(Y/X)$)
- Non-negative (but $H(X) - H(X/y)$ may be negative!)
- Zero iff X, Y independent

CE and MI: Visual Illustration

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



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

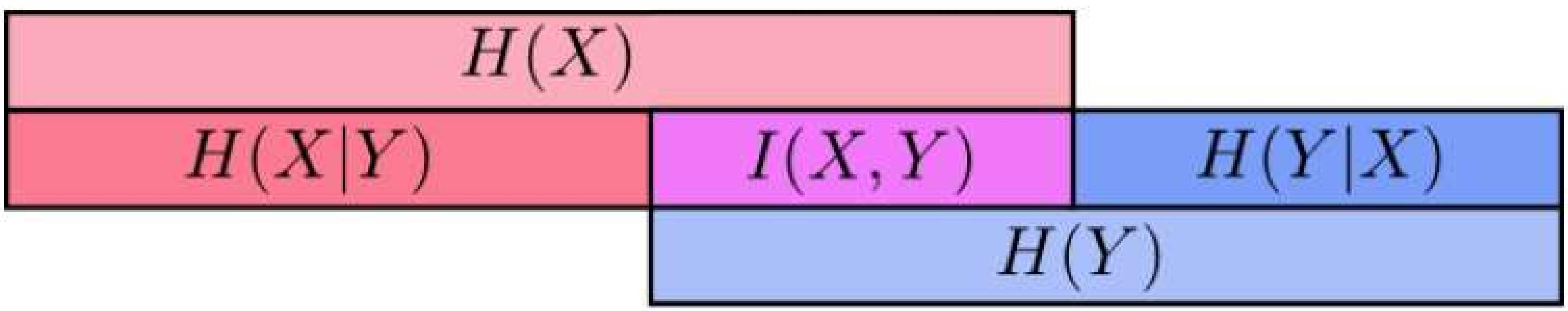


Image Credit: Christopher Olah.

Outline

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

Cross Entropy

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

$$\boxed{H(p,q) = -\sum_{x\in\mathcal{X}} p(x)\,\log q(x)}$$

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.

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

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

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

KL Divergence is a distance measurement

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

$$\sum_s P(s)\log\frac{Q(s)}{P(s)} \le \log\sum_s \cancel{P(s)}\frac{Q(s)}{\cancel{P(s)}} \quad \text{by Jensen}$$

$KL[P||Q] \geq 0$

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

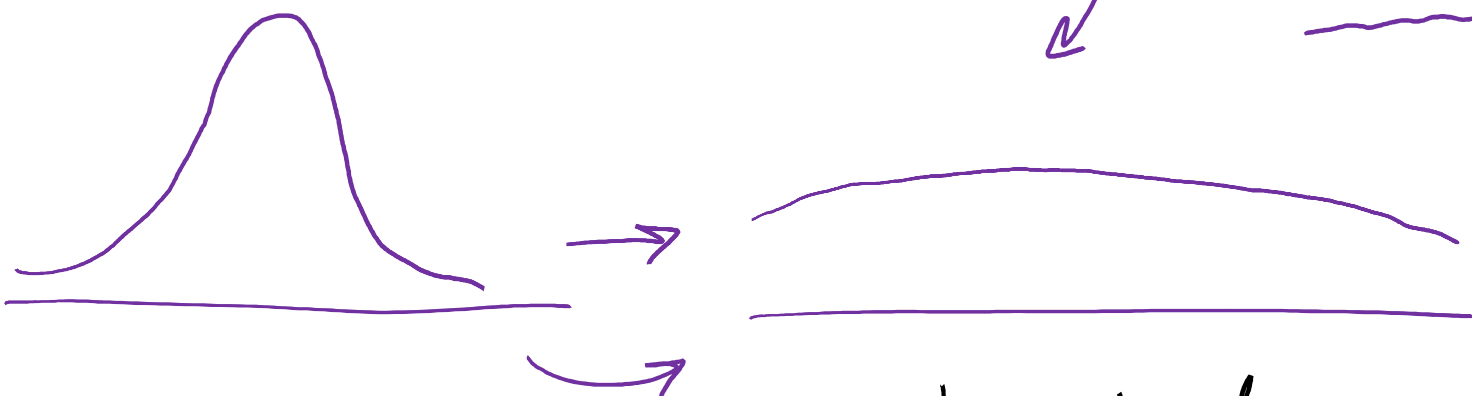
So $\mathbf{KL}[P\|Q] \ge 0$. Equality iff $P = Q$

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

$$I(x) = \log_2 \frac{1}{P(x)} \rightsquigarrow H(x) = \sum_{k=1}^{K} P(x)\, I(x)$$



$$P(Y=BP \mid X=6) = \frac{\overbrace{P(X=6 \mid Y=BP)}^{\text{Likelihood}}\ \overbrace{P(Y=BP)}^{\text{Prior knowledge}}}{P(X)} \to \text{marginalization}$$

Posterior probability

$$P(X=6 \mid Y=BP) \checkmark$$

$$\mu = \frac{\sum x_i}{n} \qquad \sigma = \frac{\sum (x_i-\mu)^2}{n}$$

$$P(X) = \sum_{Y} P(X,Y) = \sum_{Y} P(X \mid Y)\, P(Y)$$

BP & NBP

Common language (jangons)

ML
- PM: probabilistic models $\rightsquigarrow$ GMM, NB
- NPM: non probabilistic models $\rightsquigarrow$ Regression

PM $\rightsquigarrow$ we need to create likelihood function to optimize the praameters

NPM $\rightsquigarrow$ Lagrangian function

$$\underbrace{C}_{\text{per month}}(A, T) = 6A^2 + 3T^2$$

$$\text{s.t} \quad A + T = 100$$

$$\text{Min} \quad C(A, T)$$

$$\underbrace{\text{s.t} \quad A + T = 100}$$

$$g(x) = \underline{A + T - 100}$$

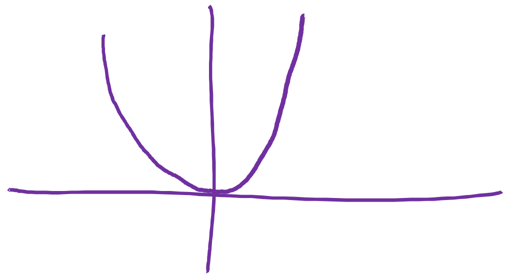
$$\underbrace{L}_{\text{lagrangian function}}(A, T, \delta) = 6A^2 + 3T^2 - \delta(\overbrace{g(x)}^{A+T-100})$$

$$\frac{\partial L}{\partial A} = 0 \Rightarrow 12A - \delta = 0 \Rightarrow A = \frac{\delta}{12} = 33.\bar{3}$$

$$\frac{\partial L}{\partial T} = 0 \Rightarrow 6T - \delta = 0 \Rightarrow T = \frac{\delta}{6} = 66.\bar{6}$$

$$\frac{\partial L}{\partial \delta} = 0 \Rightarrow A + T = 100 \Rightarrow \frac{\delta}{12} + \frac{2\delta}{12} = 100 \Rightarrow \delta = 400$$

$f(x)=x^2$



$f''(x) = 2 > 0 \rightsquigarrow$ Hessian Matrix

min f

s.t $g_i(x)$

$$L = f - \sum_{i}^{M} \delta_i g_i(x)$$

M constraind Eq

$A+T \le 100 \rightsquigarrow A+T=100$ & you follow KKT assumptions

$\delta_i > 0$