The week ahead

- **Quiz 1,** mean is 76% and average completion time 6.26 min
- **Assignment 1 Early bird special** → 2 complete questions by Wednesday, Sep 2nd
- **Second round of project seminars**, available Thursday, Aug 3rd
- **Open office hours on Thursday, 7pm to 8pm**
  - https://primetime.bluejeans.com/a2m/live-event/qfsqxjec
- **Quiz 2, Friday, Sep 4th 6am until Sep 5th 6am**
  - Information theory and optimization

Coming up soon

- Labor day, Sep 7th → NO CLASS
- Project team composition **due Sep 8th**
- **Assignment 1 due Sep 9th**

CS4641B Machine Learning

Highlights: Linear algebra and probability theory

Rodrigo Borela ▸ rborelav@gatech.edu

Linear algebra

- **Norms:** measuring vector lengths
- **Covariance and correlation:** understanding relationships between features
- **SVD:** data compression and dimensionality reduction

$$\mathbf{X} = \begin{bmatrix} 2 & 3 & \cdots & 3 \\ 5 & 6 & \cdots & 6 \\ 3 & 5 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 2 & 7 & \cdots & 4 \end{bmatrix}$$

datapoint

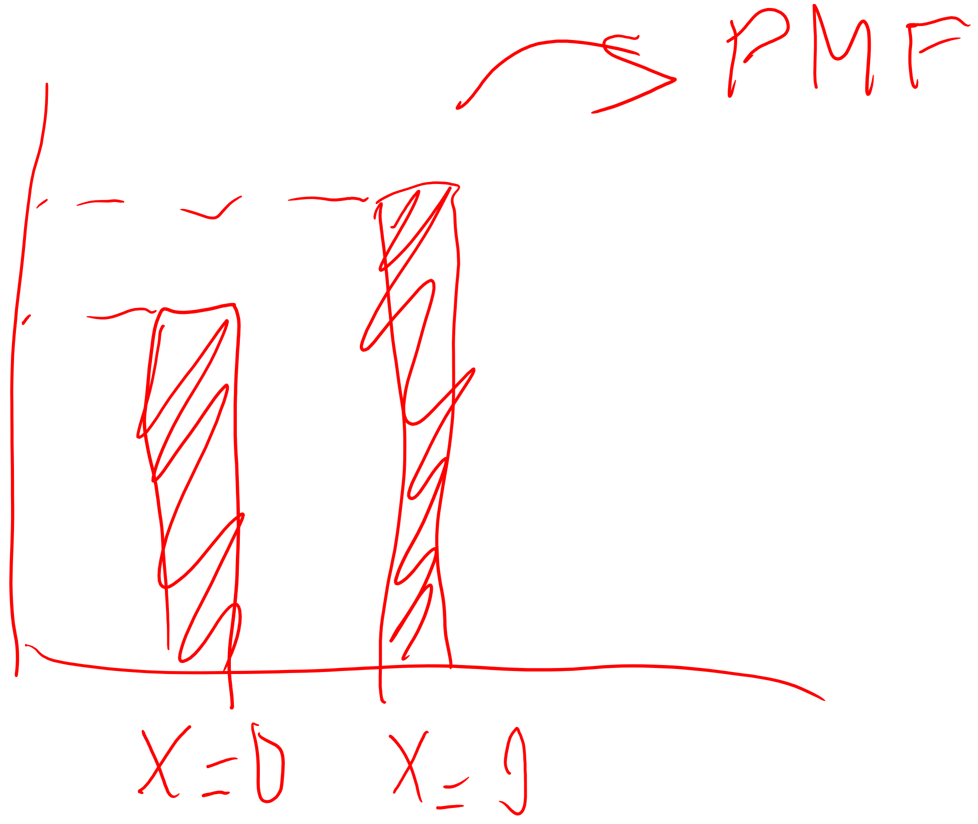
$$\mathbf{x}_1^{\mathrm{T}} = [2 \quad 3 \quad \cdots \quad 3]$$

feature (characteristic, measurement, etc.)

Probability theory

- Discrete variable
  - Example: Coin flip (integer)
  - Discrete probability distribution (e.g. Bernoulli)
  - Probability mass function
  - Probability value

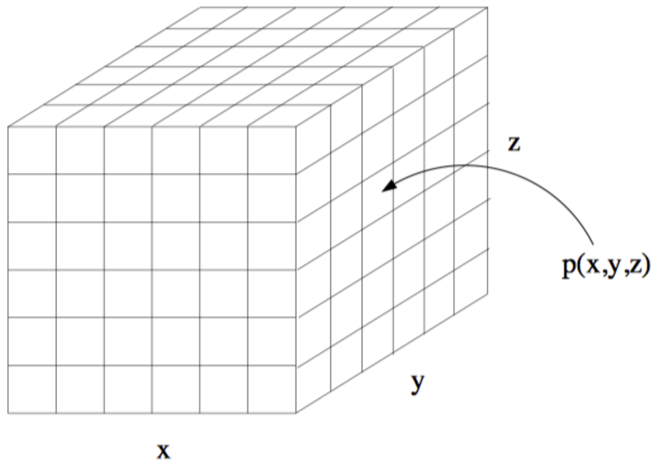
$$\sum_{x \epsilon A} p(x) = 1$$



- Continuous variable
  - Example: Temperature (real number)
  - Continuous probability distribution (e.g. Gaussian)
  - Probability density function
  - Density or likelihood value

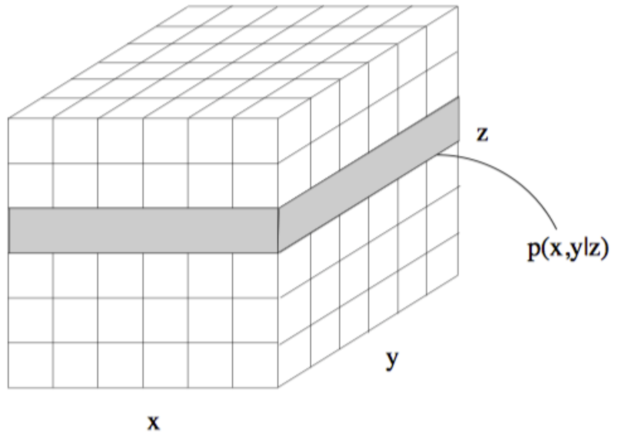
$$\int_x p(x)dx = 1$$

Joint, conditional and marginal distribution



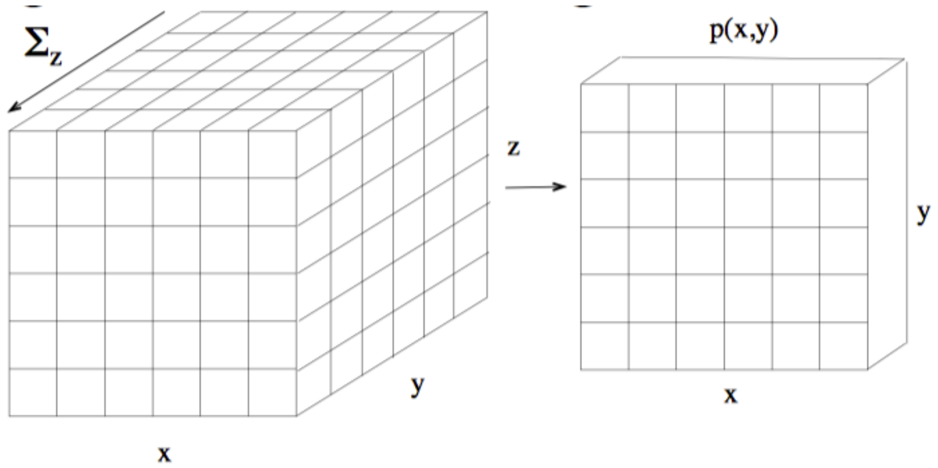
Joint distribution

$p(x, y) = p(X = x \text{ and } Y = y)$, from the product rule $p(x, y) = p(x|y)p(y)$



Conditional distribution

$p(x|y) = p(X = x|Y = y)$, from Bayes' theorem $p(x|y) = \frac{p(x,y)}{p(y)} = \frac{p(y|x)p(x)}{p(y)}$



Marginal distribution

$p(x) = p(X = x)$, from the sum rule $p(x) = \sum_y p(x, y)$

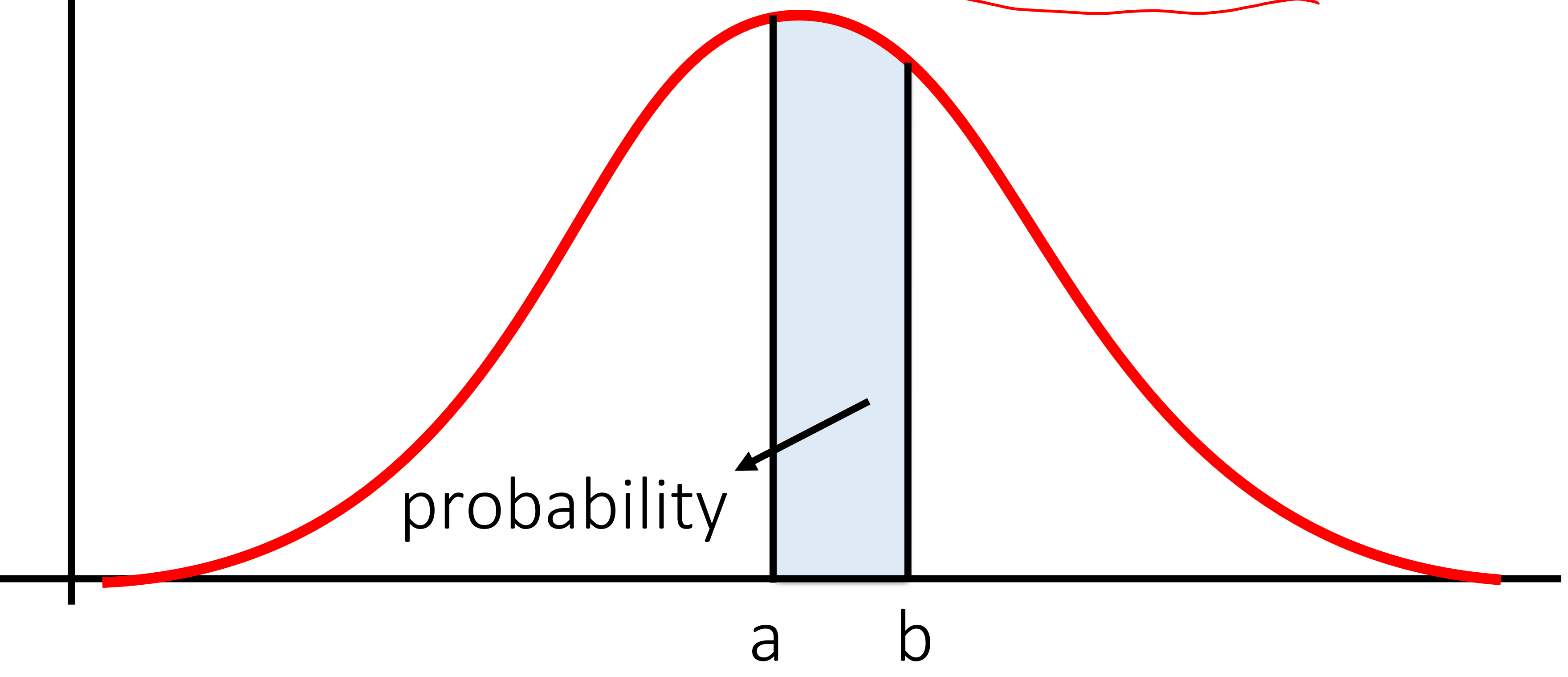
Probability vs likelihood

Gaussian density function: $\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

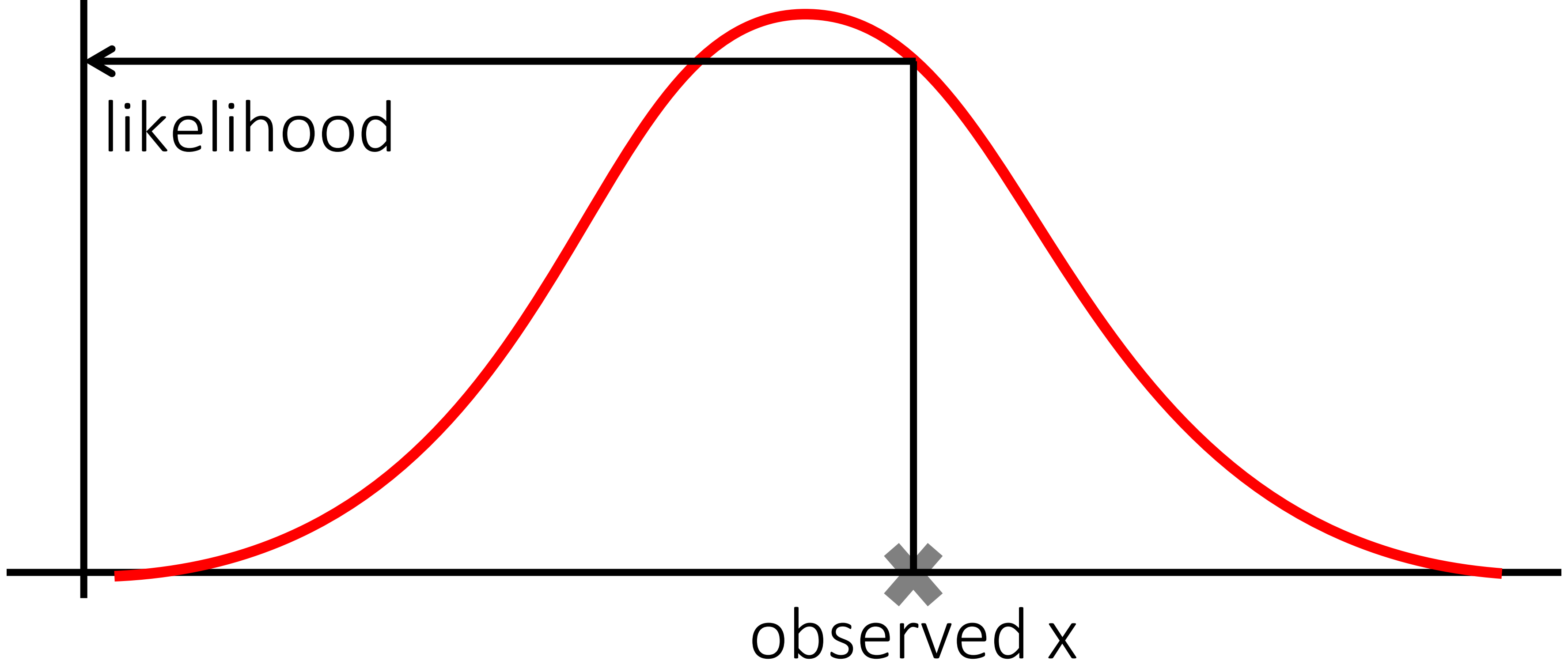
MEAN

VARIANCE

What is the probability of $x \in [a,b]$?

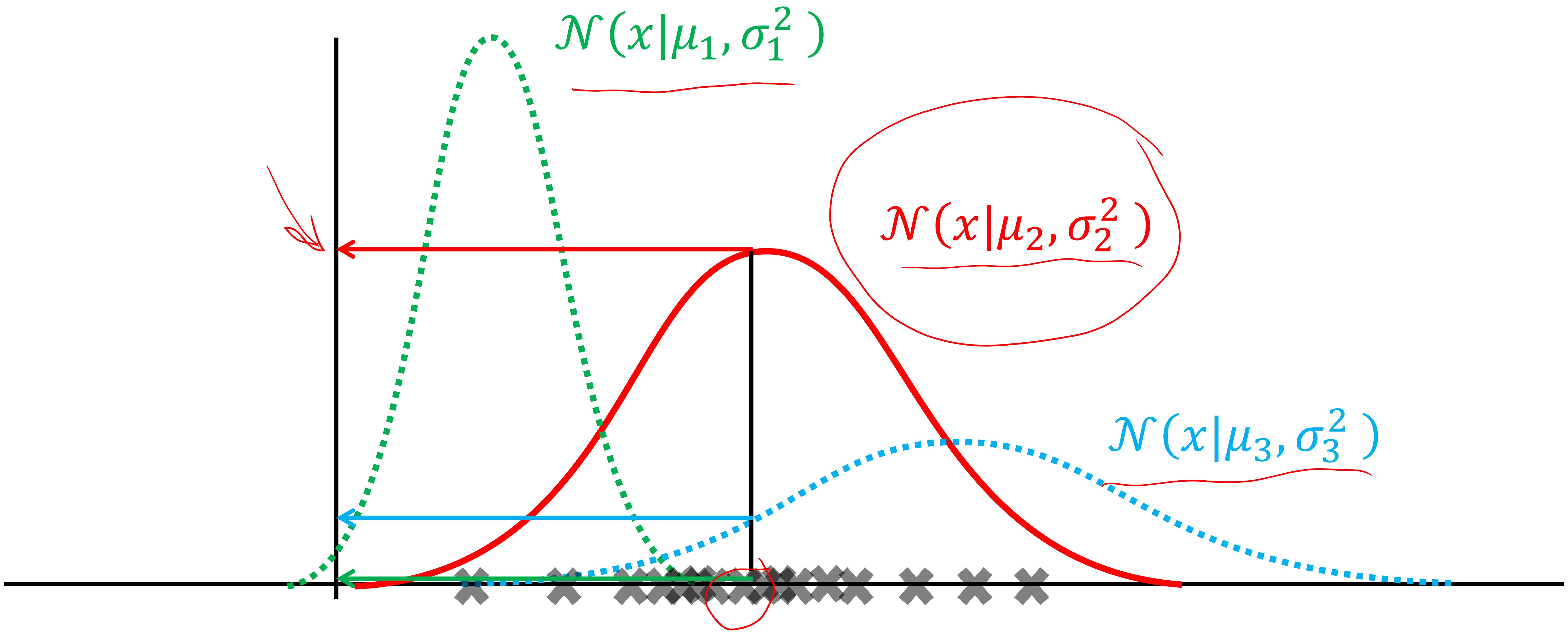


How likely is x for a distribution with parameters μ,σ^2?



Maximum likelihood estimation

- What are the parameters that best explain the data I have observed?



Maximum likelihood estimation

1. Write the likelihood function for our dataset using i.i.d. assumption

$$L(\mathcal{D}|\,\theta) = p(x_1, x_2, x_3, \dots, x_n)$$

applying the i.i.d. assumption

LIKELIHOOD X_1

$$= p(x_1)p(x_2)\dots p(x_n)$$

DATASET

PARAMETERS

$$L(\mathcal{D}|\,\theta) = \prod_{i=1}^{n} f(x_i|\theta) \rightarrow L(\mathcal{D}|\,\mu,\sigma^2) = \prod_{i=1}^{n} \mathcal{N}(x_i|\mu,\sigma^2\,)$$

PROB DENSITY FUNCTION

Maximum likelihood estimation

GAUSSIAN

2. Compute the logarithm to of the likelihood function

$$logL(\mathcal{D}|\,\theta) = l(\mathcal{D}|\,\theta) = \sum_{i=1}^{n} \log f(x_i|\theta) \rightarrow l(\mathcal{D}|\,\mu, \sigma^2) = \sum_{i=1}^{n} \log \mathcal{N}(x_i|\mu, \sigma^2)$$

3. Maximize the log-likelihood with respect to each parameter

$$\frac{\partial l}{\partial \mu} = 0 \rightarrow \mu_{ML}$$ (the mean that maximizes the likelihood)

$$\frac{\partial l}{\partial \sigma^2} = 0 \rightarrow \sigma_{ML}^2$$ (the variance that maximizes the likelihood)

CS4641B Machine Learning

Lecture 05: Information theory

Rodrigo Borela ▸ rborelav@gatech.edu

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

Outline

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

Complementary reading: Bishop PRML – Chapter 1, Sections 1.6 through 1.6.1

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.

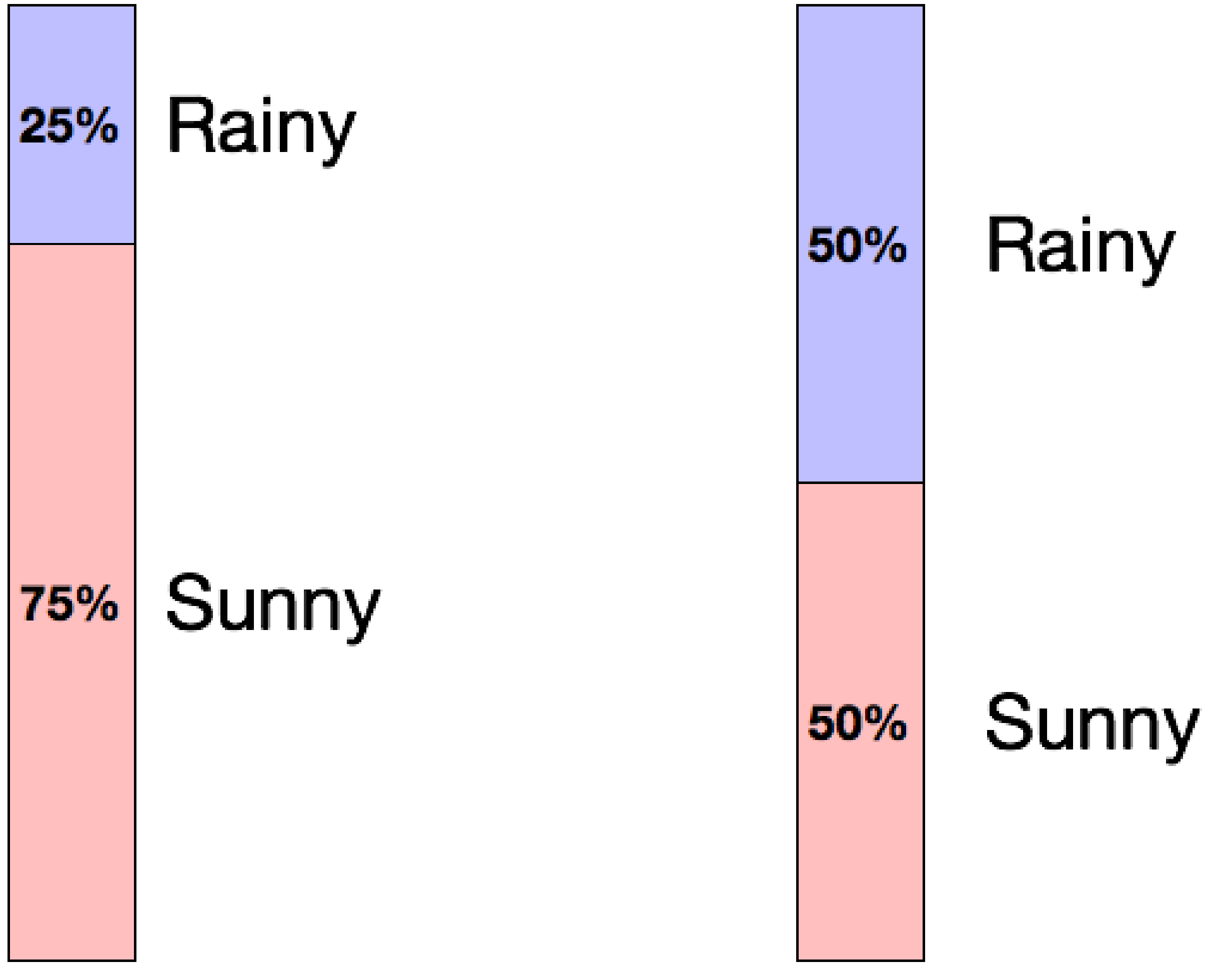
- You need information to be able to get knowledge

  Data/fact → information → knowledge

- Information ≠knowledge

  Concerned with abstract possibilities, not their meaning

Uncertainty and Information



Which day is more uncertain? How do we relate uncertainty and information?

Information

- Define a measure of information based on the probability of an event happening
- 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 beautiful Los Angeles, California and you are told it did not rain yesterday → **not a lot of information since it rarely rains in SoCal**
- We can associate our measure of information with probability of an event occurring. Let X be a random variable with distribution $p(x) = p(X = x)$:

$$I(x) = h(x) = -\log_2 p(x)$$

Example: is a picture worth **1,000** words?

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

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

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

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

- 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! #shook

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

Example: compression

- 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?

Example: compression

- 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

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

Motivation: Compression

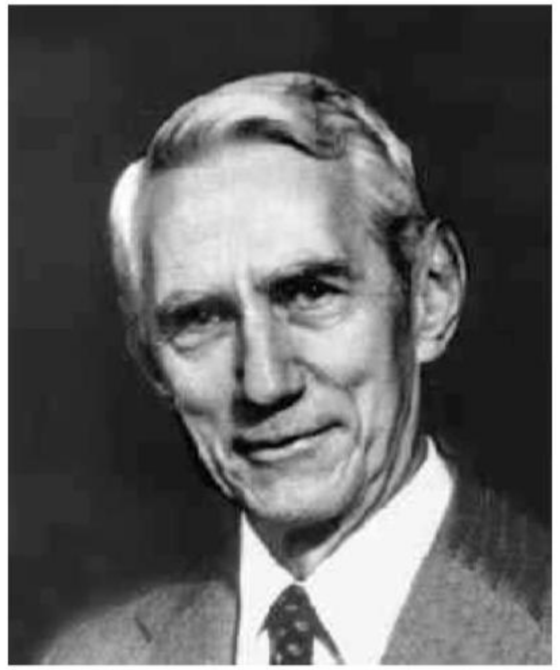
- Frequently occurring events should have short encodings
- We see this in English with words such as “a”, “the”, “and”, etc.
- We want to maximize the information-per-character
- Seeing common events provides little information
- Seeing uncommon events provides a lot of information

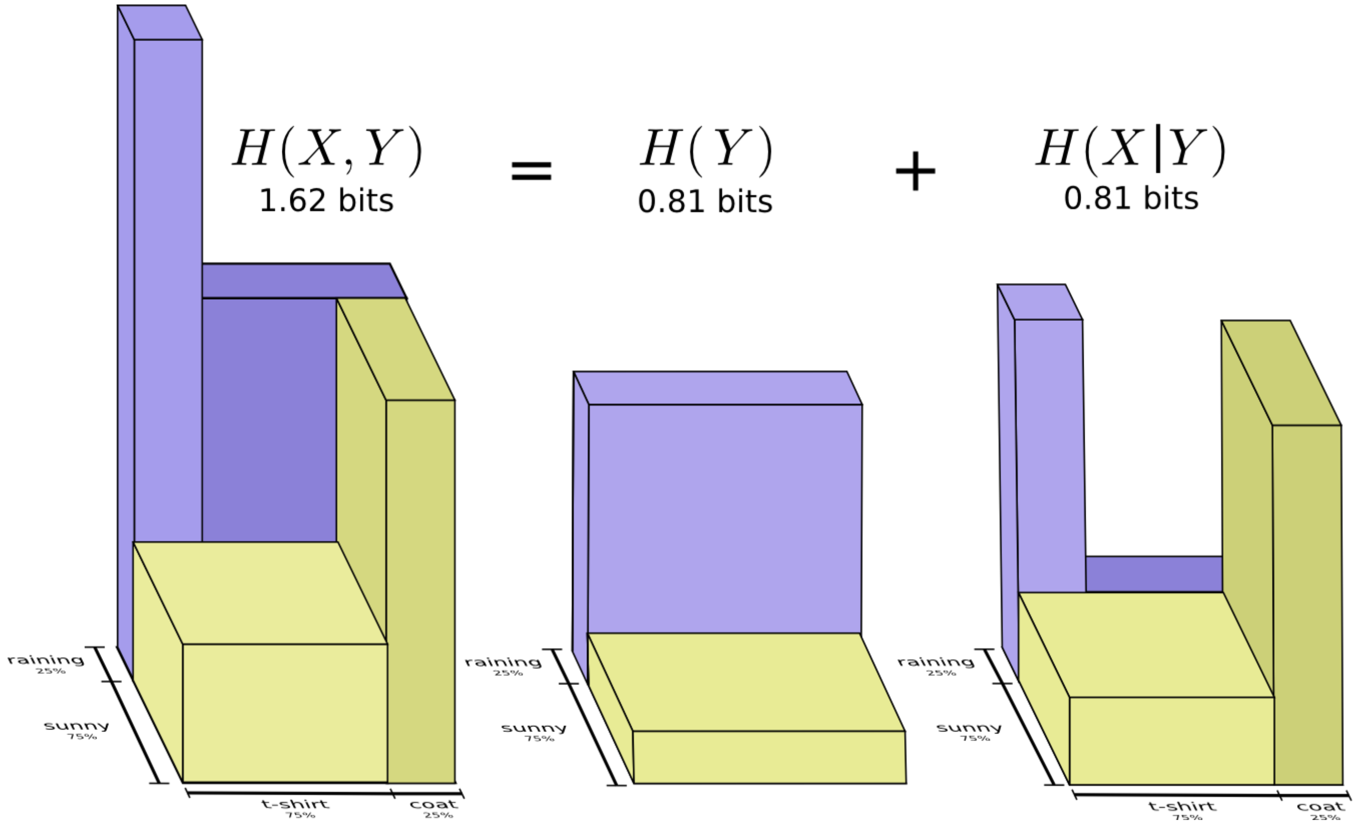
Outline

- Motivation
- Entropy
- Conditional entropy and mutual information
- Cross-Entropy and KL-Divergence

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



Entropy

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

$$H(x) = E[h(x)] = \sum_{x} h(x)p(x) = -\sum_{x} p(x)\log_2 p(x)$$

Considering a random variable X with k possible states:

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

- Information theory:
  - Most efficient code assigns $-\log_2 P(x=k)$ bits to encode the message $x=k$.

Example: entropy computation

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

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

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

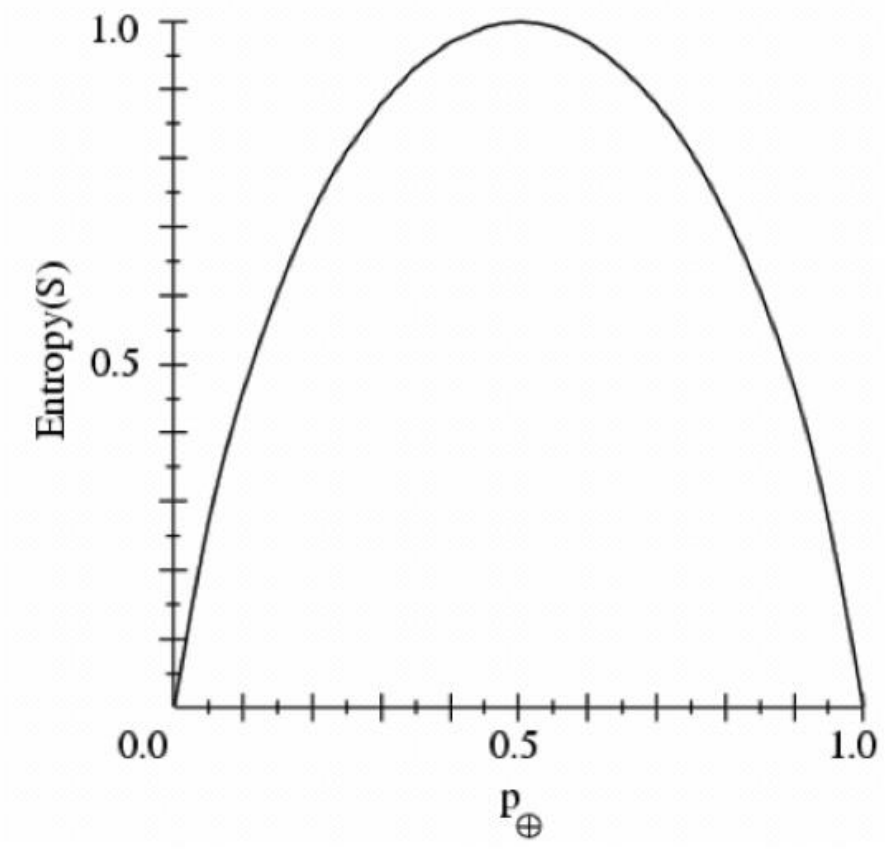
Head	2
Tail	4

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

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 $\boldsymbol{S}$

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



Properties of Entropy

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

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

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

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

- $H(P) \leq \log_2 \mathrm{k}$, with equality iff $p_i = \frac{1}{\mathrm{k}}, \forall i$
- The further P is from uniform, the lower the entropy

Outline

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

Joint Entropy

$p(T = t, M = m)$

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

- $H(T) = H\big(p(cold), p(mild), p(hot)\big) = H(0.3, 0.5, 0.2) = 1.48548$
- $H(M) = H\big(p(low), p(high)\big) = H(0.6, 0.4) = 0.970951$
- $H(T) + H(M) = 2.456431$
- Joint entropy: consider the space of (t, m) events:

$$H(T, M) = \sum_{t,m} p(T = t, M = m) \cdot \log_2 \frac{1}{P(T = t, M = m)}$$

$$H(0.1, 0.4\ , 0.1, 0.2, 0.1, 0.1) = 2.32193$$

Notice that $H(T, M) < H(T) + H(M)$. Does it make sense!?

Conditional Entropy

$p(T = t|\, M = m)$

temperature

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

- Conditional entropy

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

- $H(T|\, M = low) = H\left(\frac{1}{6}, \frac{4}{6}, \frac{1}{6}\right) = 1.25163$
- $H(T|\, M = high) = H\left(\frac{2}{4}, \frac{1}{4}, \frac{1}{4}\right) = 1.5$

- 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

temperature

$p(M = m|\, T = t)$

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

- Conditional entropy
- $H(M|\, T = cold) = H\left(\frac{1}{3}, \frac{2}{3}\right) = 0.918296$
- $H(M|\, T = mild) = H\left(\frac{4}{5}, \frac{1}{5}\right) = 0.721928$
- $H(M|\, T = mild) = H\left(\frac{1}{2}, \frac{1}{2}\right) = 1.0$

- 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 random variable

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

Mutual Information

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

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

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

- Mutual information is symmetric

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

$$I(Y,X) = \int \sum_{i=k}^{K} p(x_i, y = k) \log_2 \frac{p(x_i, y = k)}{p(x_i)p(y = k)} dx_i$$

$$= \int \sum_{i=k}^{K} p(x_i, y = k) \log_2 \frac{p(x_i|y = k)}{p(x_i)} dx_i$$

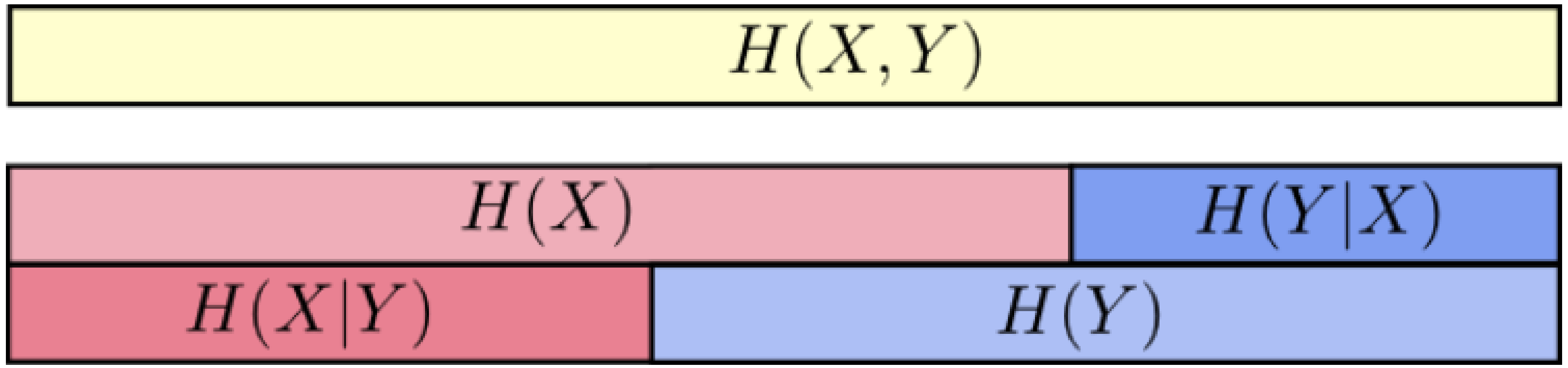
Properties of mutual information

$$I(X,Y) = H(X) - H(X|Y) = \sum_x p(x) \cdot \log \frac{1}{p(x)} - \sum_{x,y} p(x,y) \cdot \log \frac{1}{p(x|y)}$$

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

- Properties of average mutual information:
  - Symmetric (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 $\boldsymbol{X}, \boldsymbol{Y}$ independent**

Conditional entropy and mutual information



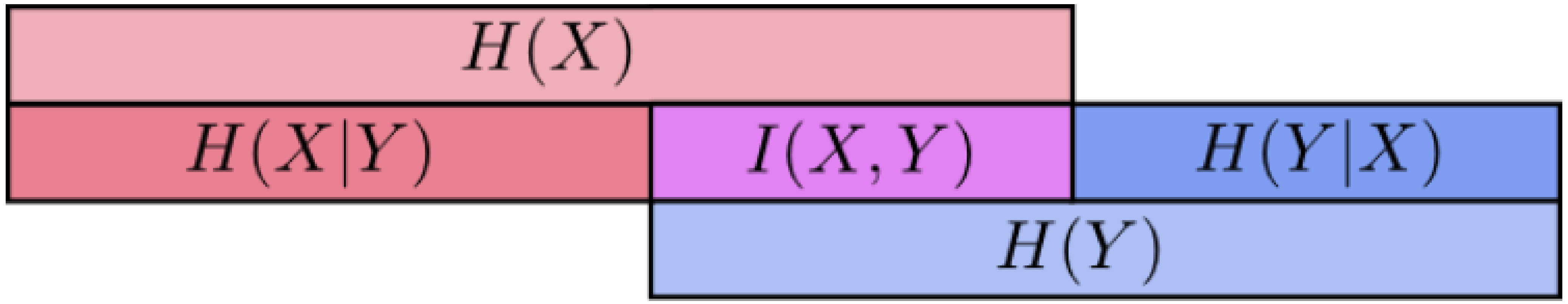


Image Credit: Christopher Olah.

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

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

- This is because

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

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

Kullback-Leibler Divergence

- Another useful information theoretic quantity measures the difference between two distributions
- $KL[p(x)][q(x)] = \sum_{x_i} p(x_i) \log \frac{p(x_i)}{q(x_i)} = \underbrace{\sum_{x_i} p(x_i) \log \frac{1}{q(x)}}_{\text{Cross-entropy}} - H[p] = H(P,Q) - H(P)$
- Excess cost in bits paid by encoding according to q instead of p

$$-KL[p][q] = \sum_{x} p(x) \log \frac{q(x)}{p(x)}$$

$$\sum_{x} p(x) \log \frac{q(x)}{p(x)} \leq \log \sum_{x} p(x) \frac{q(x)}{p(x)}$$

$$\log \sum_{x} q(x) = \log 1 = 0$$

By Jensen Inequality $E[g(x)] \leq g(E[x])$

$g(x) = \log(x)$

- So $KL[p][q] \geq 0$. Equality iff $p = q$, $KL[p][q] = 0$