Brandeis

COSI 104a Introduction to Machine Learning

Chapter 1 – Math concepts

Instructor: Dr. Hongfu Liu
Email: hongfuliu@brandeis.edu

How to present a sample

- Sample is the basic unit, which contains multiple features or attributes
  - Student ID, name, email, age, …

- Vector
  - In most cases, a sample is presented as a row vector

d-dimension vector

$$\boldsymbol{v} = \vec{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_d \end{bmatrix} \qquad \vec{v}[k] \text{ – the } k\text{-th element in } \boldsymbol{v}$$

Transpose of a vector

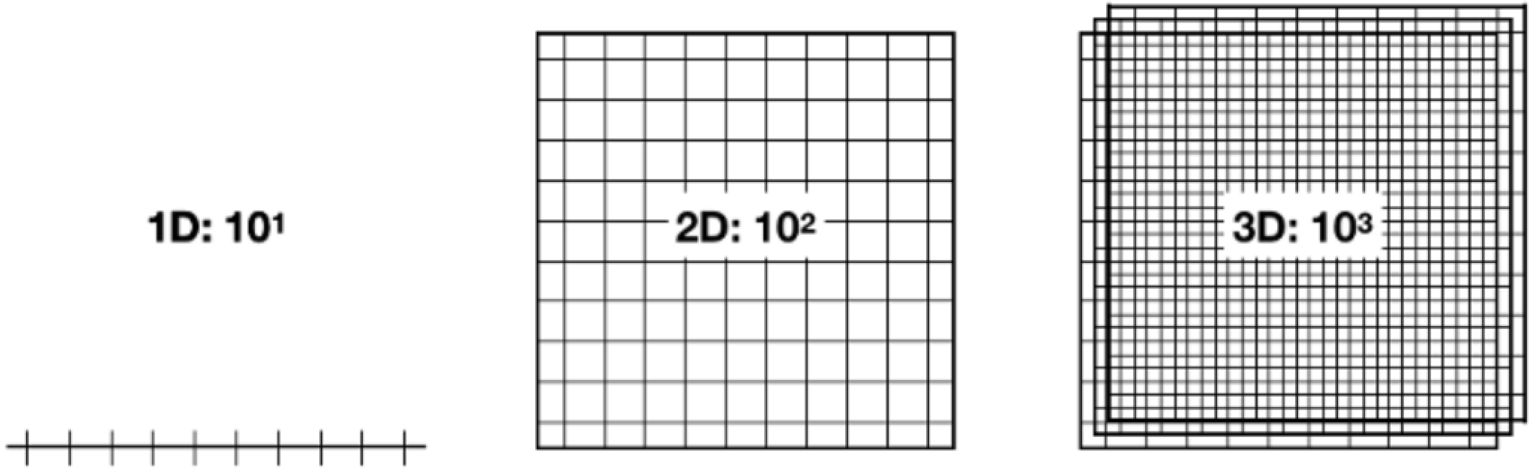
$$\boldsymbol{v}^T = \vec{v}^T = [v_1 \; v_2 \; \dots v_d]$$

Vector operation

- In machine learning, it has little physical meaning to add two samples (Average of a group of samples is meaningful). Instead, it is widely used to calculate the similarity of two samples.

- How?
  - Minus
  - Other ways?

- Curse-dimensionality



The number of features required to keep average distance constant grows exponentially with the number of dimensions.

Vector operation

$$\vec{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_d \end{bmatrix} \quad \vec{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_d \end{bmatrix}$$

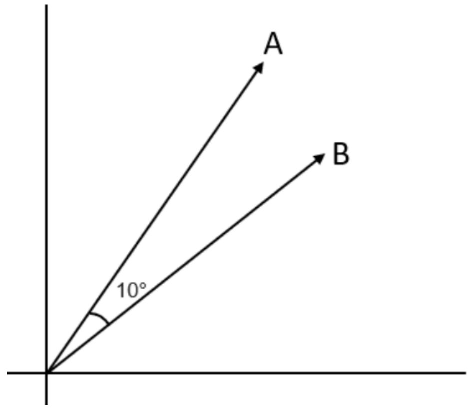
The magnitude/length of a vector

$$\|\vec{v}\| = \sqrt{\sum_{k=1}^{d} v_k^2}$$

The inner product of two vectors $\langle \vec{v}, \vec{u} \rangle = \vec{v}^T \vec{u} = \vec{u}^T \vec{v} = \sum_{k=1}^{d} v_k u_k$

$\vec{v}$ and $\vec{u}$ are said to be orthogonal if $\langle \vec{v}, \vec{u} \rangle = 0$

The angle θ between two vectors satisfies: $cos\theta = \frac{\langle \vec{v}, \vec{u} \rangle}{\|\vec{v}\| \|\vec{u}\|}$



Vector operation

- 1. Document Similarity
- A scenario that involves the requirement of identifying the similarity between pairs of a document is a good use case for the utilization of cosine similarity as a quantification of the measurement of similarity between two objects.

- To find the quantification of the similarity between two documents, you need to convert the words or phrases within the document or sentence into a vectorized form of representation.

- The vector representations of the documents can then be used within the cosine similarity formula to obtain a quantification of similarity.

- In the scenario described above, the cosine similarity of 1 implies that the two documents are exactly alike and a cosine similarity of 0 would point to the conclusion that there are no similarities between the two documents.

Step 1: Obtain a vectorized representation of the texts.

Vectorised Representation

Word	Document 1	Document 2
Deep	1	1
Learning	1	1
Can	1	1
Be	1	1
Hard	1	0
Simple	0	1

A vectorized representation of texts in table format. | Image: Richmond Alake

- Document 1: [1, 1, 1, 1, 1, 0] let's refer to this as A
- Document 2: [1, 1, 1, 1, 0, 1] let's refer to this as B

Step 2: Find the Cosine Similarity

`cosine similarity (CS) = (A . B) / (||A|| ||B||)`

- Calculate the dot product between A and B: `1.1 + 1.1 + 1.1 + 1.1 + 1.0 + 0.1 = 4`.
- Calculate the magnitude of the vector A: `√1² + 1² + 1² + 1² + 1² + 0² = 2.2360679775`.
- Calculate the magnitude of the vector B: `√1² + 1² + 1² + 1² + 0²+ 1² = 2.2360679775`.
- Calculate the cosine similarity: `(4) / (2.2360679775*2.2360679775) = 0.80` (80 percent similarity between the sentences in both document).

- Any problem?

Matrix

- A group of samples

$$A_{m\times n} = \underbrace{\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}}_{n \text{ columns}} \left.\vphantom{\begin{bmatrix} a_{11} \\ \vdots \\ a_{m1} \end{bmatrix}}\right\} \begin{matrix} m \\ \text{rows} \end{matrix}$$

$A_{m\times n}[i,j]$ – the element at the i-th row and j-th column of A

Transpose of a matrix $$(A_{m\times n})^T = \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nm} \end{bmatrix}$$

$$(A_{m\times n})^T[j,i] = A_{m\times n}[i,j]$$

Matrix operation

$$A_{m\times n}B_{n\times q} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} b_{11} & \cdots & b_{1q} \\ \vdots & \ddots & \vdots \\ b_{n1} & \cdots & b_{nq} \end{bmatrix} = \begin{bmatrix} c_{11} & \cdots & c_{1q} \\ \vdots & \ddots & \vdots \\ c_{m1} & \cdots & c_{mq} \end{bmatrix}$$

$$\text{where} \quad c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$$

The product of matrix and vector

$$A\vec{v} \overset{?}{=} \vec{v}^T A$$

Matrix operation

$$AB = \underset{2 \times 2}{\begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix}} \underset{2 \times 3}{\begin{bmatrix} 1 & 2 & 0 \\ 0 & 1 & 2 \end{bmatrix}} = \underset{2 \times 3}{\begin{bmatrix} 2 & 5 & 2 \\ 1 & 6 & 8 \end{bmatrix}}$$

match

$$BA = \underset{2 \times 3}{\begin{bmatrix} 1 & 2 & 0 \\ 0 & 1 & 2 \end{bmatrix}} \underset{2 \times 2}{\begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix}} \quad \text{undefined}$$

do not match

Matrix operation

$$\text{Matrix } A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 6 & 9 \end{bmatrix}$$

$$A^{100}$$

- When we use matrix multiple in machine learning?

- Matrix inverse

You need to master

- Load a data matrix into Python
- Get a subset of the data by random sampling, or some characteristics
- Get one sample from the matrix according to its index
- Get one feature from the matrix according to its index
- Feature normalization
- Calculate the similarity between two samples

Other concepts

- Probability
- Distribution
- Expectation/Variance
- And so on
- They are really important for advanced machine learning. But…

Random Variables

- A random variable is the result of a stochastic experiment, which can be measured. It can be discrete or continuous.
- Stochastic experiment: a process in which various elementary states (or events, outcomes) are possible.
  - Flip a coin. The state set $\boldsymbol{S} = \{\text{Head, Tail}\}$.
  - Cast a six-face dice. $\boldsymbol{S} = \{1, 2, 3, 4, 5, 6\}$.
  - Binary decision. $\boldsymbol{S} = \{0, 1\}$.
  - The price of a stock.
  - …

If a variable X is discrete, $P(x)$ represents the probability that $X = x$ happens.

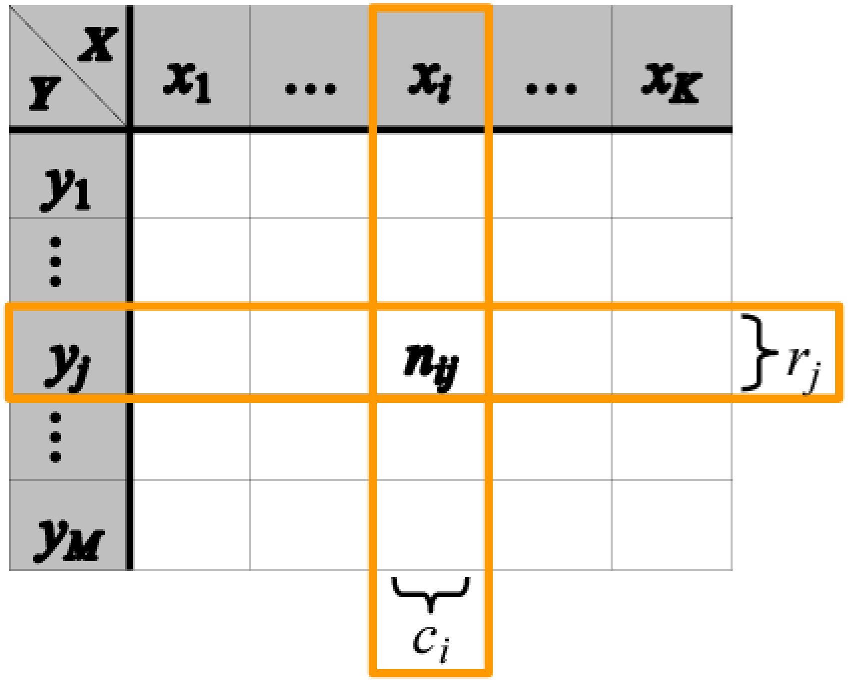
$$\sum_x P(x) = 1$$

Flip a Coin

	Head	*Tail*
$P(X)$	0.49	0.51

Probability

Total is N



- **Joint Probability**

$$P(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

- **Marginal Probability**

$$P(X = x_i) = \frac{c_i}{N}$$

$$P(Y = y_j) = \frac{r_j}{N}$$

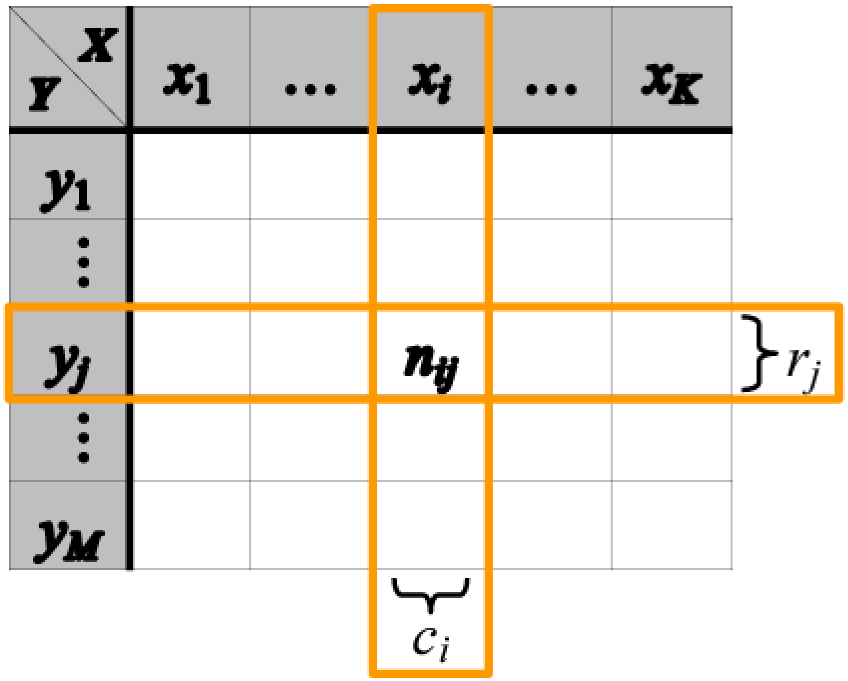
- **Conditional Probability**

$$P(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

$$P(X = x_i | Y = y_j) = \frac{n_{ij}}{r_j}$$

Probability

Total N



- **Product Rule**

$$P(X,Y) = P(Y|X)\,P(X)$$

$$P(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

$$= \frac{n_{ij}}{c_i} \times \frac{c_i}{N} = P(Y|X)\,P(X)$$

- **Sum Rule** $P(X) = \sum_{Y} P(X,Y)$

$$P(X = x_i) = \frac{c_i}{N} = \frac{\sum_{j=1}^{M} n_{ij}}{N} = \sum_{j=1}^{M} P(X = x_i, Y = y_j)$$

Probability

- **Bayes' Theorem (Rule)**

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

- **Chain Rule**

$$P(X, Y, Z, \ldots) = P(X)P(Y|X)P(Z|X, Y)P(\ldots|X, Y, Z)$$

Probability

- **Independent**

$$P(X,Y) = P(X)P(Y)$$

- **Conditional Independent**

$$P(X,Y|Z) = P(X|Z)P(Y|Z)$$

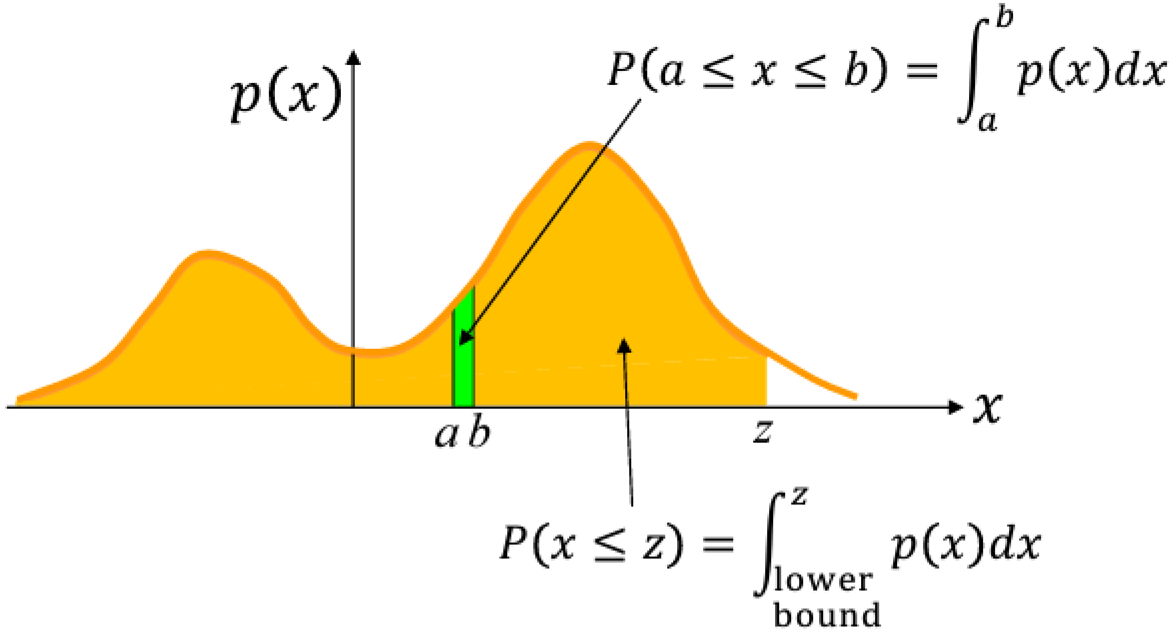
Without knowing Z $\quad P(X,Y) = P(X)P(Y)$?

The **p**robability **d**istribution **f**unction (PDF) of a variable X, usually denoted as $p(x)$, describes the likelihood of the possible values that a random variable can take.

Properties:

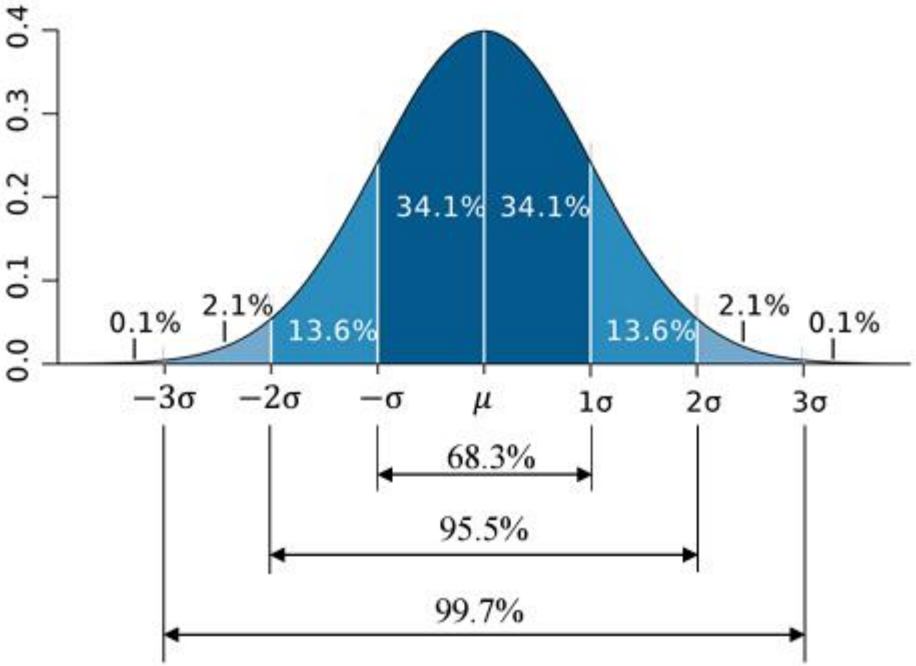
$$p(x) \geq 0$$

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



Gaussian Distribution

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

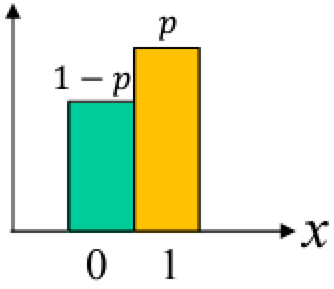


Bernoulli Distribution

Example: Flip a coin with probability p coming up head

A single random variable X which takes value 1 with probability p and value 0 with probability $1-p$.

$$x \sim Bern(p) = p^x(1-p)^{1-x}$$



Expectation

$$x \sim P(x)$$

$$E[f] = \sum_x P(x) f(x) \qquad E[f] = \int p(x) f(x) dx$$

Given **i**dentically **i**ndependently **d**istributed data $\{x_n\}_{n=1,\ldots,N}$

$$E[f] \cong \frac{1}{N} \sum_{n=1}^{N} f(x_n)$$

Condition on y: $E_x[f|y] = \sum_x P(x|y) f(x)$

Variance

Variance: the expectation of the squared deviation from mean (i.e., the degree of spread out from the average value.

$$var[f] = E[(f(x) - E[f(x)])^2] = E[f(x)^2] - E[f(x)]^2$$

Covariance: the joint variability of two random variables

$$cov[x, y] = E[(x - E[x])(y - E[y])]$$
$$= E[xy] - E[x]E[y]$$

x	0	1	2	3	4
$f(x)$	1/5	1/5	1/5	1/5	1/5

- Standard deviation

https://online.stat.psu.edu/stat500/lesson/3/3.2/3.2.1

Important quantity in

- Coding theory
- Statistical physics
- Machine learning

$$H[X] = -\sum_{x} P(x) \log_2 P(x)$$

Coding theory: X is discrete with 8 possible states. At least how many bits are needed to transmit the state of X if all states equally likely?

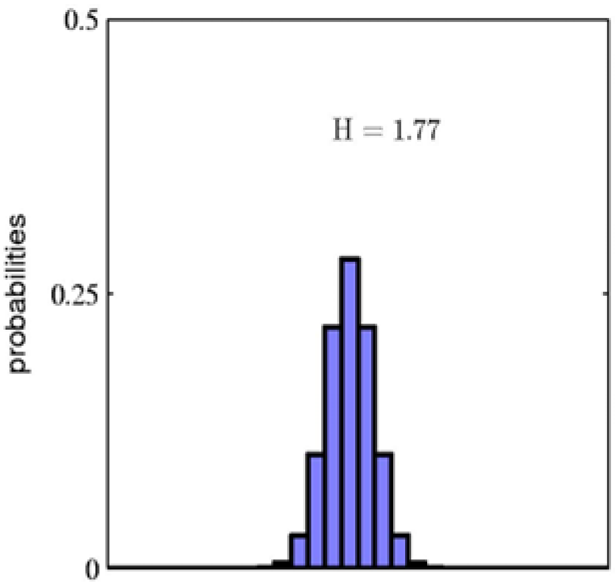
$$\mathrm{H}[X] = -8 \times \frac{1}{8} \log_2 \frac{1}{8} = 3$$

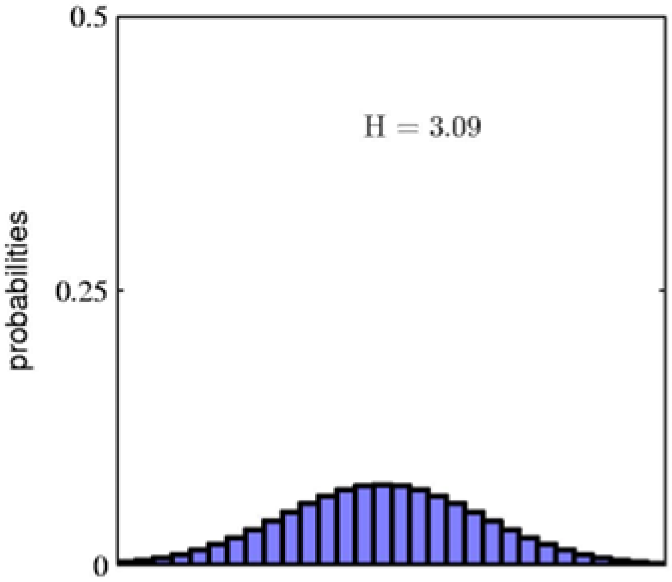
$$H[x] = -\sum_{x} P(x) \log_2 P(x)$$

x	a	b	c	d	e	f	g	h
$p(x)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{64}$	$\frac{1}{64}$	$\frac{1}{64}$	$\frac{1}{64}$
code	0	10	110	1110	111100	111101	111110	111111

$$\begin{aligned} \mathrm{H}[x] &= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{16}\log_2\frac{1}{16} - \frac{4}{64}\log_2\frac{1}{64} \\ &= 2 \text{ bits} \end{aligned}$$

$$\begin{aligned} \text{average code length} &= \frac{1}{2}\times 1 + \frac{1}{4}\times 2 + \frac{1}{8}\times 3 + \frac{1}{16}\times 4 + 4\times\frac{1}{64}\times 6 \\ &= 2 \text{ bits} \end{aligned}$$





When is H maximized?

Bernoulli: $P(\boldsymbol{X}) = p^X(1-p)^{1-X}$

Conditional Entropy

$$\mathrm{H}[Y|X] = -\int\int p(x,y)\log p(y|x)\;dy\,dx$$

$$\mathrm{H}[X,Y] = \mathrm{H}[X] + \mathrm{H}[Y|X] = \mathrm{H}[Y] + \mathrm{H}[X|Y]$$

Mutual Information

$$\mathrm{I}[X,Y] = -\int\int p(x,y)\log\left(\frac{p(x)p(y)}{p(x,y)}\right)dx\,dy$$

$$\mathrm{I}[X,Y] = \mathrm{H}[X] - \mathrm{H}[X|Y] = \mathrm{H}[Y] - \mathrm{H}[Y|X]$$


Q
&
A