Information Theory

Information theory: provides a theoretical ground for ML in general

Problems we aim at solving:
- how to complete a sequence?
  1 2 3 5 7 11 13 17 ?
  ↳ why is "19" more "probable"? how to justify it?
- how to decide between 2 models for given data?
- how to set a ML problem?
  ↳ which criterion to optimize?
  ↳ how to measure a solution's performance?

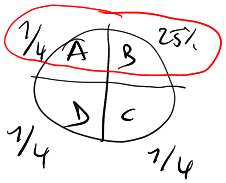
This will lead us to the following problems:
- how to quantify information?
- is there a "natural" distribution over | numbers? / models?
- how to compress (losslessly) data? are there bounds?

Information Theory is also (not covered in this course):
- communication in a noisy channel (Shannon) ⇒ error-correcting codes
- hash functions

I - Information & Entropy

- quantify information
  → game of 20 binary questions
    ↳ level of information = number of questions to ask to identify the object
- pick an object according to a random law ⇒ can do better than dichotomy if I know the distribution
  ↳ fewer questions ↳ take into account probabilities

Ex:



- A, B, C, D: 4 possible objects
- I select one of them randomly according to the law:

$$\begin{cases} p(A) = 1/4 \\ p(B) = 1/4 \\ p(C) = 1/4 \\ p(D) = 1/4 \end{cases}$$

Question 1: is it (A ∪ B)?

yes → Question 2: is it A?
- yes → A
- no → B

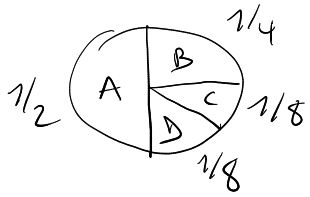
no → D or C
is it C?
- yes → C
- no → D

⇒ in all cases: you need to ask 2 questions to identify the object
↳ on average: 2 questions

⇒ The amount of information I need to provide you to identify the object, knowing that I picked it according to that law is 2 bits of information.

↳ by answering binary questions

Non-uniform example:



green: non-uniform case

orange: compare uniform

Is it A?

- yes (1/2) → A! → 1 bit (1/4)
- no → Is it B?
  - yes (1/4) → B! → 2 bits (1/4)
  - no → is it C?
    - yes (1/8) → C → 3 bits (1/4)
    - no (1/8) → D → 3 bits (1/4)

On average:

$$= \frac{1}{2}\times 1 + \frac{1}{4}\times 2 + \frac{1}{8}\times 3 + \frac{1}{8}\times 3$$

$$= 1.75$$

< 2 in the uniform case

$$= \frac{1}{4}(1+2+3+3)$$

$$= 2{,}25 > 2$$

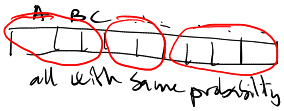
Desired properties:

- continuous function H of the distrib° of a random variable X:

$$H(Q_X) = H((p_1, p_2, p_3, \dots))$$

↳ $H(p_1+\varepsilon, p_2-\varepsilon, p_3 \dots) \simeq H(p_1, p_2, p_3 \dots)$

- symmetry: $H(p_1, p_2, p_3) = H(p_2, p_1, p_3)$
- additivity / independence to internal representation:



H(uniform over n bins) = H(distribution over clusters) + $\sum_i p(\text{cluster } i) \times H$(uniform over bins inside cluster i)

average # of questions asked if proceeding in 2 steps

average # of questions with the direct process (1 step)

or

$$H(p_1, p_2, p_3 \dots) = H(p_1+p_2, p_3, \dots) + (p_1+p_2)\, H\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right)$$

- uniform distribution has maximum entropy & its entropy increases with the number of bins

Entropy definition

Proposition [Khinchin, 1957]: Any function F satisfying the properties above can be written as: $F(p) = -C \sum_i p_i \log p_i$ for some constant $C \in \mathbb{R}^+$

$\vec{p} = (p_1, p_2, p_3 \dots)$

We choose $C=1$, $\log_2$: **Shannon entropy**

↳ unit: bit: number of binary questions needed

Examples:

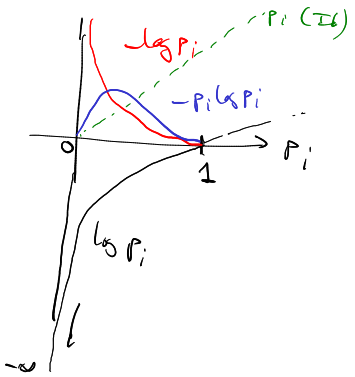
- uniform law over n bins: $\forall i,\ p_i = 1/n$

$$H(p) = -\sum_i p_i \log p_i = \sum_i p_i \times (-\log p_i)$$

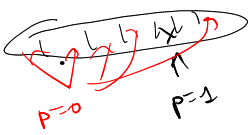
($-\log x$; $-\log p_i \geq 0$)

$$= n \times \frac{1}{n} \times \left(-\log \frac{1}{n}\right)$$

$$= \log_2 n \geq 0$$

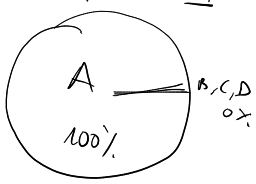


- Dirac peak:



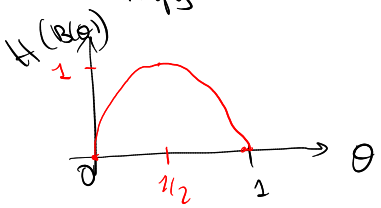
$$H = 0 = \sum \underbrace{0 \log 0}_{0} + 1 \times 0 = 0$$

A B C D previous ex.



- Bernoulli $B(\theta)$
  - toss a coin → head θ
  - → tail $1-\theta$

  - if $\theta = 0$ or 1: Dirac peak → $H = 0$
  - if $\theta = 1/2$: maximum entropy: uniform over 2 bins → $H = \log_2 2 = 1$



Other properties

- $H(X) \geq 0$
- $H(X) = 0$ iff $\exists i\ p_i = 1$ (Dirac peak) no information
- H(any distribution over n bins) $\leq \log_2 n$
- $H(f(X)) \leq H(X)$ for any (deterministic) function f (cf. later)

II - Why "$-\log p$"?

- describe (encode) an integer $\in [\![1, 2^b]\!]$: with b bits
  - ↳ ask $b = \log n$ binary questions
  - $= -\log 1/n$ ↳ binary tree dichotomy

15 : 01111

$[\![1, 8]\!]$

$[\![1, 32]\!]$

- code any event with $-\log p$ bits
  - For $n = 2^b$, uniform law: $\log n = -\log p$ $\quad p = 1/n$
  - if non uniform, $p = 2^{-s}$: see ABCD case before
  - general probability: $\sim \lceil -\log p \rceil$
    - ↳ we'll see later that we can reach $-\log p$

$\frac{1}{8} \leq p \leq \frac{1}{4}$ ⇒ need 3 questions

$p \geq 1/8$

$\frac{1}{p} \leq 8 \quad -\log p \leq 3$

- entropy $H(p)$: $= \mathbb{E}_{x\sim p}[\underbrace{\text{information needed to describe the choice of that event } x}_{-\log p(x)}]$

$=$ average length to encode events from p

- Ex: english text
  - ↳ random indep. characters
  - . ASCII encoding: 256 possible characters $\Rightarrow \log(256) = 8$ bits $\leftrightarrow$ uniform law over 256 symbols
  - . 27 characters are much more probable (↳ 26 alphabet + 1 space) $\Rightarrow \log(27) = 4.8 \leftrightarrow$ uniform ..... 27 ..
  - . use character frequencies in English $\Rightarrow H(\text{English}) = 4.1$

III Extensions : for pairs of variables / of laws

- Conditional entropy.
  - . joint law over (X,Y) :
  - . $H(X,Y) = H((X,Y)) = -\sum_{ij} p(x_i,y_j)\log p(x_i,y_j)$
  - . define $H(Y|X)$ as average (↳ over x) entropy of $p(y|x)$ (distribution over y for fixed x)

↳ $= \sum_x p(x)\, H(Y|x) = \mathbb{E}_{x\sim p}[H(Y|X=x)] = \sum_x p(x) \sum_y p(y|x)\log p(y|x)$

$= -\sum_{xy} p(x,y)\log p(y|x)$

  - . $H(Y|X) \geq 0$
  - . $H(Y|X) \leq H(Y)$
  - . $H(X,Y) = H(X) + H(Y|X)$
    $= H(Y) + H(X|Y)$

Bayes: $p(x,y) = p(x)\,p(y|x)$

$-\log p(x,y) = -\log p(x) - \log(p(y|x))$

$\mathbb{E}_{xy}$ {

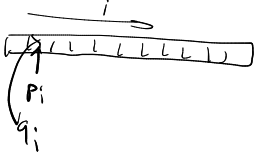
$H(X,Y) = H(X) + H(Y|X)$

Mutual information:

- . $H(X,Y) \leq H(X) + H(Y)$
- . entropy difference
  mutual information: $I(X,Y) = H(X)+H(Y) - H(X,Y)$ — information gain when considering X & Y together
  $= H(X) - H(X|Y)$
  $= H(Y) - H(Y|X)$
  - ↳ symmetric (X vs Y)
  - ↳ $I(X,Y) \geq 0$
  - ↳ $I(X,Y) = 0$ iff $X \perp Y$ (independent)
    $p(X,Y) = p(X)\,p(Y)$

Relative entropy : Kullback-Leibler divergence

- . consider $X\sim P$, $Y\sim Q$
- $KL(P\|Q) = \sum_i p_i \log \frac{p_i}{q_i}$



- $KL(P\|Q) \geq 0 \quad \forall P, Q$ known as Gibbs inequality
- $KL(P\|P) = 0$
  $KL(P\|Q) = 0$ iff $P = Q$

} KL as a sort of "distance" to compare 2 distributions P and Q

- not symmetric: $KL(Q\|P) \neq KL(P\|Q)$ in general $\Rightarrow$ not a "distance"

- $KL(P\|Q) =$ cross-entropy(P,Q) $-$ entropy(P)

$\mathbb{E}_P[-\log q]$ $\quad$ $H(P) = \mathbb{E}_P[-\log p]$

$KL \geq 0$ means cross-entropy $\geq$ entropy

Rewriting mutual information

p: joint law over $X \times Y$
p: X
q: Y

$$I(X,Y) = H(X) + H(Y) - H(X,Y)$$
$$= -\sum_x p(x)\log p(x) - \sum_y q(y)\log q(y) + \sum_{xy} p(x,y)\log p(x,y)$$

$p(x,y) = p(x)\,p(y|x) = q(y)\,p(x|y)$

$$= -\sum_{xy} p(x,y)\log p(x) - \sum_{xy} p(x,y)\log p(y) + \text{"} \cdots \text{"}$$

$\sum_y p(x,y) = p(x)$ $\quad$ $p(y) = \sum_x p(x,y)$

$$= -\sum_{xy} p(x,y)\log\frac{p(x)p(y)}{p(x,y)}$$

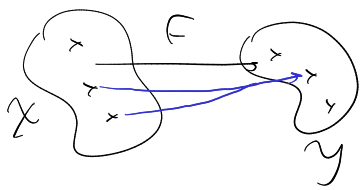
$$= +KL\big(p(x,y)\,\|\,p(x)p(y)\big)$$

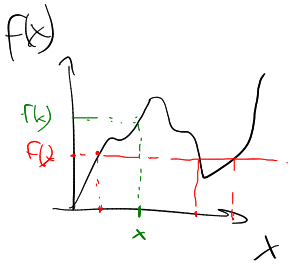
joint law $\quad$ marginals

how far the joint law $p(x,y)$ is from a product of independent laws
i.e. how far x & y are from being independent
$x \perp y \Leftrightarrow p(x,y) = p(x)p(y)$

Information can only be lost by processing

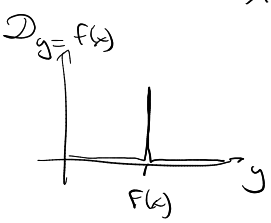
- consider a function f: deterministic





$$H((X, f(X))) = H(X) + \underbrace{H(f(X)|X)}_{=0} = H(f(X)) + \underbrace{H(X|f(X))}_{\geq 0}$$

$H(f(X)) \leq H(X)$



$y|x$
$f(x)|x$

III-B Sequences of non-independent variables

$X_1, X_2, \ldots X_n$ sequence

$$H(X_1, X_2, \ldots X_n) = H(X_1) + H(X_2|X_1) + H(X_3|X_1,X_2) + \cdots + H(X_n|X_1,X_2,\ldots X_{n-1})$$

(Bayes)

$$p(x_1 \cdots x_n) = p(x_1)\,p(x_2|x_1)\,p(x_3|x_1,x_2)\cdots$$

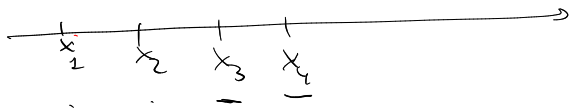
- $n \to +\infty$:

entropy rate: $H(X) = \lim_{n\to\infty} \frac{H(X_1,X_2,\ldots X_n)}{n}$

Examples:

- i.i.d. (identically distributed: $\text{Law}(X_i) = \text{Law}(X_j)$; independent: $X_i \perp X_j$) : $H(\mathcal{X}) = \lim_{n\to\infty} \frac{H(X_1) + H(X_2) + H(X_3) + \dots H(X_n)}{n} = \lim_{n\to+\infty} \frac{n * H(X_1)}{n} = H(X_1)$

- Markov chain : $p(X_n | X_1 \dots X_{n-1}) = p(X_n | X_{n-1})$ order-1 M.C.



$X_2 = 2X_1 \pm 1$ (1/2, 1/2)

$X_3 = 2X_2 \pm 1$

$= 4X_1 (\pm 1) + \pm 1$

$p(X_3|X_2) = p(X_3|X_1, X_2)$

$$H(\mathcal{X}) = \lim_{n\to+\infty} \frac{H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) \dots}{n}$$

$H(X_3|X_2) = H(X_2|X_1)$

$$= \lim_{n\to\infty} \frac{H(X_1) + (n-1) H(X_2|X_1)}{n}$$

$$= H(X_2|X_1)$$

$\frac{1}{n} H(X_1) + \frac{n-1}{n} H(X_2|X_2)$

$\frac{1}{n} \to 0$, $\frac{n-1}{n} \to 1$

III-C Continuous distributions

From discrete to continuous distributions:

- continuous space $\mathcal{X}$ with probability density p
- $H(X_\delta) = -\sum_{\text{bins } i} p_\delta(i) \log p_\delta(i)$

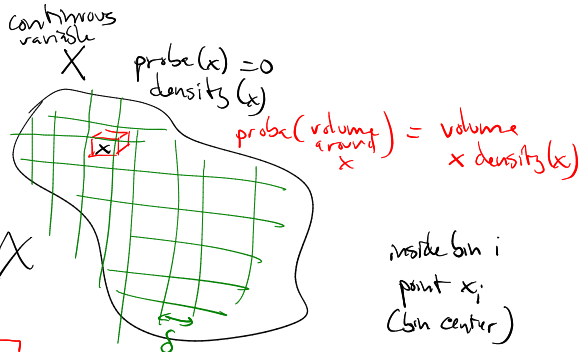
$$= -\sum_{\text{bins } i} \delta^D p(x_i) \log \left(\delta^D p(x_i) \right)$$

$$= -\delta^D \sum_i p(x_i) \log p(x_i) - \delta^D \sum_i p(x_i) D \log \delta$$

$\sum_i p(x_i) \to 1$

$\delta \to 0$: $-\int_{\mathcal{X}} p \log p - D \log \delta$ ($-D\log\delta \to +\infty$)

- Differential entropy: $\boxed{H(X) := -\int_{\mathcal{X}} p(x) \log p(x) \, dx}$



discretization step

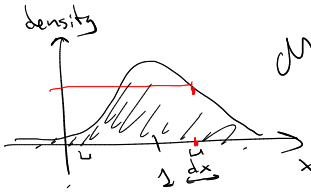
$\Downarrow$ back to discrete setting

X_δ bin (discrete random variable)

$p_\delta(\text{bin})$

$p_\delta(\text{bin } i) = \text{volume of bin } i \times p(x_i)$

δ^D



$\mathcal{N}(1, 0.5)$ Gaussian distribution

$p(x = 1,230\,470\,0000) = 0$

$p(x \in [1,23 \dots \pm 0.001]) = p(x) dx$

$dp(x)$

$\int_{\mathcal{X}} p(x) dx = 1$ $\int_{\mathcal{X}} dp(x) = 1$ $\int_{\mathcal{X}} p = 1$

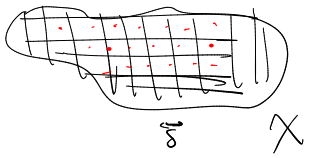
⚠ $H(X)$ is not necessarily ≥ 0 ⚠

↳ $KL(P\|Q) = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} dx$

↳ same properties as before: $KL(P\|Q) \geq 0$ and minimum (0) is reached when $P = Q$

because: $KL(P\|Q) = \mathbb{E}_P[-\log Q] - H(P)$

Details about integrals



$\sum_{\text{sets } i} \text{volume}(i) \; f(\text{center of set } i) \longrightarrow \int_{\mathcal{X}} f(x)\,dx$

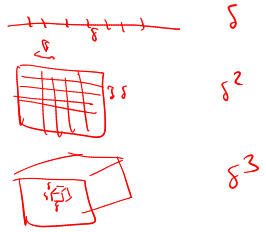
$\delta \to 0$ discretization step, set volume δ^D

D: dimension of the space $\mathcal{X}$

D=1: $x \in \mathbb{R}$ — δ

D=2: $x \in \mathbb{R}^2$ — δ^2

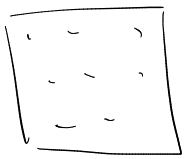
D=3 — δ^3



III-D Example

EEG head set
→ type letters just by focusing on them → no muscle involved
→ control a mouse just by thinking left/right

↳ issue: very noisy signals
→ maximize information retrieved by optimizing protocol



→ optimal way to flash letters in order to identify the letter we focus on

Noisy-channel coding theorem [Shannon, 1948]

- transmitter: X → noisy channel → Y : receiver
- capacity: $C = \sup_{P_X} I(X;Y)$
- For any $\varepsilon > 0$ and for any transmission rate $R \leq$ channel capacity C,
  $\exists$ encoding & decoding scheme that transmits data at rate R
  with error probability $\leq \varepsilon$ for a sufficiently large block length
- For any rate $R \geq C$, $p(\text{error}) \to 1$