Cheatsheet InfoTheory

Nicolas Wehrli

June 2024

1 Foundations

Definitions

Information of an outcome x

$$h(x) = -\log(p(x))$$

Cross-Entropy between p and q

$$H(p;q) = -\sum_x p(x)\log q(x)$$

Shannon Entropy

$$H(p) = H(p;p)$$

Notation
We identify outcomes x with integers $1,...,m$ and associate probabilities $p(x) \geq 0$.
$H(\frac{1}{m})$ for $H(p)$ with $p(x) = \frac{1}{m}$ (uniform)
$H(X) = H(p) = \mathbb{E}(-\log(p(X)))$ where p is the pdf of X

Jensen's Inequality
Let f be convex and $g : [m] \to \mathbb{R}$ be an arbitrary function that assigns a value to each outcome.

$$f\left(\sum_x p(x)g(x)\right) \leq \sum_x p(x)f(g(x)), \forall p(x) \geq 0, \sum_x p(x) = 1$$

alternatively

$$f(\mathbb{E}(g(X))) \leq \mathbb{E}(f(g(X)))$$

Applying this inequality to relate **Cross-Entropy** and **Entropy**, we get the following properties.

$$H(p;q) \geq H(p)$$

Definining **KL divergence** or **Relative Entropy** as

$$D(p||q) = \sum_x p(x)\log\frac{p(x)}{q(x)}$$

we get

$$H(p;q) = H(p) + D(p||q)$$

Further investigating KL divergence, we find

$$D(p||q) \geq 0 \tag{1}$$

$$D(p||q) = 0 \iff p = q \tag{2}$$

A further consequence of (1) is that the uniform distribution **maximizes** entropy.

$$H\left(\frac{1}{m}\right) = \max_p H(p)$$

Definitions - Conditional distributions

Conditional information

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

Conditional Entropy

$$H(X|Y=y) = -\sum_x p(x|y)\log p(x|y)$$

$$H(X|Y) = \sum_y p(y)H(X|Y=y)$$

Monotonicity of Conditioning

$$H(X|Y) \leq H(X)$$

Joint Entropy

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

Chain Rule

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

Subadditivity

$$H(X,Y) \leq H(X) + H(Y)$$

with equality if $X \perp Y$.

Multiple Conditioning

$$H(X|Y,Y') \leq H(X|Y)$$

Generalized to $X_1,...,X_n$ we get

$$H(X_1,...,X_n) = \sum_{i=1}^n H(X_i|X_1,...,X_{i-1}) \leq \sum_{i=1}^n H(X_i)$$

Mutual Information

$$I(X;Y) := H(X) - H(X|Y) = H(Y) - H(Y|X)$$

We further have

$$I(X;Y) = D(P(X,Y)||P(X)P(Y))$$

with $I(X;Y) = 0$ if $X \perp Y$.

Conditional Mutual Information

$$I(X;Y|Z) := H(X|Z) - H(X|Y,Z)$$

Conditional Independence
If $X \perp Y|Z$

$$I(X;Y|Z) = 0 \text{ and } I(X;Y) \leq I(X;Z)$$

We can deduct that for any function ϕ on outcomes of X

$$I(\phi(X);Y) \leq I(X;Y)$$

2 Compression

Definition - Code

A code C is a mapping from outcomes to codewords

$$C : \{1,...,m\} \to \{0,1\}^*$$

- If there is no codeword that is a prefix of another codeword, the code is a **prefix code**.
- Prefix codes retain injectivity when concatenating codewords.

Sets of codewords fulfilling the prefix property can be uniquely represented by the leaves of a binary tree. Since a leaf node has no children the prefix property is guaranteed.

Kraft's Inequality
If $\{c_1,...,c_m\}$ are codewords of a prefix code, then

$$\sum_x 2^{-l_x} \leq 1, \text{ where } l_x = |c_x| \tag{3}$$

Conversely, given $\{l_1,...,l_m\} \subset \mathbb{N}$ satisfying (3), there exists a prefix code with those codeword lengths.

- A prefix code is **succinct**, if Kraft's inequality holds with a equality. Else it can be optimized by pruning.
- Succinct codes uniquely define a **dyadic probabilistic model**

$$q(x) = 2^{-l_x}$$

- Expected codeword length of a prefix code C

$$L(C) = \sum_x p(x)l_x = \sum_x p(x)(-\log q(x)) = H(p;q)$$

- Using $H(p;q) = H(p) + D(p||q)$ we can deduce that the **minimal** $L(C)$ for a binary prefix code C is

$$L^* = H(p) + \min_{q:\text{dyadic}} D(p||q)$$

- Thus the closer q is to p, the more optimal the prefix code is. But since p doesn't have to be dyadic there can be an **inherent suboptimality** based on rounding.

Weak Law of Large Numbers

Let $Y_1,...,Y_n$ be iid. random variables with mean μ. Then

$$\overline{Y}_n := \frac{1}{n}\sum_{i=1}^n Y_i \xrightarrow{\mathbb{P}} \mu \iff \lim_{n\to\infty}\mathbb{P}(|\overline{Y}_n - \mu| < \varepsilon) = 1, \forall\varepsilon > 0$$

Typicality - Asymptotic Equipartition
Let $X_1,...,X_n \overset{\text{iid}}{\sim} p$. The ε-typical outcomes are

$$\mathcal{A}_\varepsilon^n = \left\{x \in \{1,...,m\}^n : \left|H(p) + \frac{1}{n}\sum_{i=1}^n \log p(x_i)\right| < \varepsilon\right\}$$

By the law of large numbers for any $p, \varepsilon > 0$ and $\delta > 0$, there exists an n_0, s.t. $\forall n \geq n_0$

$$\mathbb{P}(A_\varepsilon^n) > 1 - \delta$$

in particular for $\delta = \varepsilon$.
For all $p, \varepsilon > 0$ and $n \in \mathbb{N}$, let $x \in \mathcal{A}_\varepsilon^n$, then

$$2^{-n(H(p)+\varepsilon)} \leq p(x) \leq 2^{-n(H(p)-\varepsilon)}$$

$$(1-\varepsilon)2^{n(H(p)-\varepsilon)} \leq |\mathcal{A}_\varepsilon^n| \leq 2^{n(H(p)+\varepsilon)}$$

$$\implies |\mathcal{A}_\varepsilon^n| \approx 2^{nH(p)} \text{ and for } x \in \mathcal{A}_\varepsilon^n : p(x) \approx 2^{-nH(p)}$$

We define the AEP Code to encode whole sequences

$$\mathrm{AEP}_\varepsilon^n = \begin{cases} 0B^n(x) & \text{if } x \notin \mathcal{A}_\varepsilon^n \\ 1C^n(x) & \text{otherwise} \end{cases}$$

where we **enumerate** over the typical and atypical sequences.
Then the average codeword length **amortized** over the encoding of the sequence x of n outcomes is

$$\frac{1}{n}|C_\varepsilon^n(x)| \leq \frac{1}{n}(1 + \log|\mathcal{A}_\varepsilon^n|) \leq H(p) + \frac{1}{n} + \varepsilon$$

$$\frac{1}{n}|B^n(x)| \leq \frac{1}{n}(1 + \log m^n) \leq \log m + \frac{1}{n}$$

This result is theoretically optimal but practically not very useful.

Huffman Codes
Let X have outcomes $\{1, ..., m\}$ ordered (wlog) st. $p(1) \geq ... \geq p(m)$.
The Huffman contraction X' of X is defined as

$$X' = \min\{m-1, X\}$$

We define the Huffman Code C for X recursively from C' for the H. contraction X'

$$C(x) = \begin{cases} x-1 & \text{if } m = 2 \\ C'(x)0 & \text{if } x = m-1 \wedge m > 2 \\ C'(x-1)1 & \text{if } x = m \wedge m > 2 \\ C'(x) & \text{otherwise} \end{cases}$$

Let C be a length-optimal code, then

$$p(x) > p(x') \implies l_x \leq l_{x'}$$

$$\forall c \in \mathrm{Img}(C) \text{ wt. } |c| \text{ maximal} : \exists c' \in \mathrm{Img}(C).\ c' \text{ sibling of } c$$

Assume p_i ordered as above. Then a length-optimal prefix code C with $l_1 \leq ... \leq l_{m-1} = l_m$ and c_{m-1}, c_m only differing in last bit, is called **canonical**.
Huffman codes are length-optimal.

3 Prediction

A betting strategy b bets a fraction $b(x)$ on the x-th outcome.
The bookmaker provides odds 1-for-$q(x)$ for each outcome x.

$$S(X) = \frac{b(X)}{q(X)}$$

is a random variable which describes the wealth growth of a gamble.

- Maximizing $\mathbb{E}(S(X))$ over all possible b results in betting all on the highest probable outcome.
- Let $X_1, ..., X_n \sim p$ iid.

$$S_n := S(X_1, ..., X_n) = \prod_{i=1}^n S(X_i)$$

- Any strategy with $b(x) = 0, p(x) > 0$ for some x will almost surely fail for increasing n.
- **Doubling Rate**

$$W(b) = \mathbb{E}(\log S(X)) = \sum_x p(x) \log \frac{b(x)}{q(x)}$$

- Odds 1-for-q are **fair**, if $\sum_x q(x) = 1$
- In general, for fair odds, we have

$$W(b) = D(p||q) - D(p||b)$$

  which is **optimal** for $b = p$, since then $D(p||b) = 0$.
- **Conservation Theorem.** For q uniform, fair and $b = p$.

$$W(b) + H(p) = \log m$$

- With fair odds, withholding part of the budget doesn't gain anything.
- If we have $Q = \sum_x q(x) < 1$, Kelly-betting ($b = p$) remains optimal in expectation. But the **Dutch book**

$$b(x) := \frac{q(x)}{Q} \implies S(X) = \frac{b(X)}{q(X)} = \frac{1}{Q} > 1$$

  has a **guaranteed** doubling rate $W(b) = -\log Q > 0$.

Consider an offered bet, where we can bet $b \in [0,1]$. We receive αb on a win and pay βb on a loss. Then

$$W(b) = p\log(1 + \alpha b) + (1-p)\log(1 - \beta b)$$

We can find an optimal strategy using analysis (taking care of border cases). **Kelly Criterion.**

$$b^* = \min\{1, \max\{0, b\}\}, \qquad b = \frac{p}{\beta} - \frac{1-p}{\alpha}$$

4 Processes

A semi-infinite sequence of random variables $X_1, X_2, ...$ is a **stochastic process**.

- A process is **stationary** if, for any $n \in \mathbb{N}$ and any $\Delta \geq 0$

$$\mathbb{P}(X_1, ..., X_n) = \mathbb{P}(X_{1+\Delta}, ..., X_{n+\Delta})$$

- **Conditional Entropy Rate.**

$$H(X) = \lim_{t\to\infty} H(X_{t+1}|X_t, ..., X_1)$$

- X stationary $\implies H(X)$ well-defined.
- **Entropy Rate.**

$$H'(X) = \lim_{t\to\infty} \frac{1}{t} H(X_1, ..., X_t)$$

A **Markov Chain** is a stochastic process for which

$$X_{t+1} \perp X_{t-1}, ..., X_1 | X_t$$

Let X be a Markov Chain and $\pi := P(X_1)$.

- then by the independence from past and future

$$\mathbb{P}(X_1, ..., X_t) = \mathbb{P}(X_1)\mathbb{P}(X_2|X_1)\mathbb{P}(X_3|X_2)\cdots\mathbb{P}(X_t|X_{t-1})$$

- X is **time-homogeneous**, if

$$\mathbb{P}(X_{t+1}|X_t) = \mathbb{P}(X_2|X_1), \qquad \forall t \geq 1$$

- A time-homogeneous Markov Chain is fully characterized by its initial distribution and the **transition matrix** P with

$$P_{ij} := \mathbb{P}(X_2 = i | X_1 = j)$$

  then

$$\mathbb{P}(X_{i+r} = b | X_i = a) = (P^r)_{ba}$$

- time-homogeneous M.C. **stationary** $\iff \pi$ stationary $\iff P\pi = \pi$
- **Entropy Rate** of a stationary time-homogeneous M.C.

$$H'(X) = H(X) = \sum_a \pi_a \left(-\sum_b P_{ba} \log P_{ba} \right)$$

  Note that $H' = H$ if X is **stationary**.
- A M.C. is **ergodic**, iff. $\exists t \geq 1$ s.t. $(P^t)_{ij} > 0, \forall i, j$.
- An M.C. **ergodic** $\iff$ has a unique stationary distribution