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.
By Jensen we have

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

Defining 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.

- Codes for which Kraft's inequality is strict can be optimized by codeword pruning.
- A prefix is succinct, if Kraft's inequality holds with a equality.
- 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

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

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

$$p(x) > p(x') \implies l_x \leq l_{x'}$$
$$\forall c \in \text{Img}(C) \text{ wt. } |c| \text{ maximal} : \exists c' \in \text{Img}(C).\ c' \text{ sibling of } c$$

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