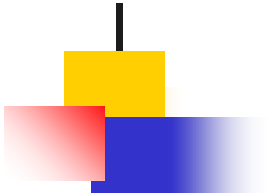
Fundamentals of Machine Learning

INFORMATION THEORY

Amit K Roy-Chowdhury

Acknowledgments: Adapted from slides at https://probml.github.io/pml-book/teaching1.html by Prof. Saw Shier Nee

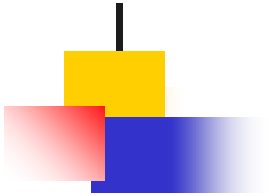


Outline

Quantifying information using entropy

- Cross entropy
- KL divergence / Relative entropy
- Joint entropy
- Mutual Information

Further Reading: Ch. 2, Elements of Information Theory, Cover and Thomas.



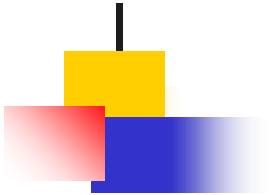
Entropy

Quantifying information is the foundation of the field of information theory.

The intuition behind quantifying information is the idea of measuring how much surprise there is in an event. Those events that are rare (low probability) are more surprising and therefore have more information than those events that are common (high probability).

- **Low Probability Event**: High Information (*surprising*).
- **High Probability Event**: Low Information (*unsurprising*).

Rare events are more uncertain or more surprising and require more information to represent them than common events.



Entropy

We can calculate the amount of information there is in an event using the probability of the event. This is called "*Shannon information*," "*self-information*," or simply the "*information*," and can be calculated for a discrete event *x* as follows:

$$\mathbb{H}(X) \triangleq -\sum_{k=1}^{K} p(X=k)\log_2 p(X=k) = -\mathbb{E}_X\left[\log p(X)\right]$$

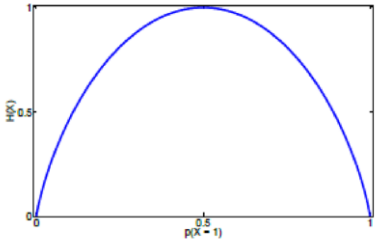
Where *log()* is the base-2 logarithm and *p(x)* is the probability of the event *x*.

The choice of the base-2 logarithm means that the units of the information measure is in bits (binary digits). This can be directly interpreted in the information processing sense as the number of bits required to represent the event.

Entropy - Examples

- Uniform Distribution $\mathbb{H}(X) = -\sum_{k=1}^{K} \frac{1}{K} \log(1/K) = -\log(1/K) = \log(K)$

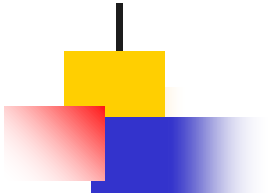
- Binomial Distribution



$$\begin{aligned}\mathbb{H}(X) &= -[p(X=1)\log_2 p(X=1) + p(X=0)\log_2 p(X=0)] \\ &= -[\theta \log_2 \theta + (1-\theta)\log_2(1-\theta)]\end{aligned}$$

- Gaussian (d dimensions) $h(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})) = \frac{1}{2}\ln|2\pi e\boldsymbol{\Sigma}| = \frac{1}{2}\ln[(2\pi e)^d|\boldsymbol{\Sigma}|] = \frac{d}{2} + \frac{d}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{\Sigma}|$

(1 dimension) $h(\mathcal{N}(\mu, \sigma^2)) = \frac{1}{2}\ln\left[2\pi e\sigma^2\right]$



Entropy - Examples

Four symbols: {a, b, c, d}; P(a) = 0.4, P(b) = 0.1, P(c) = 0.3, P(d) = 0.1, P(e) = 0.1

Huffman Code: Uses frequencies of symbols to create a variable length code.

Huffman Code: a: 0, b: 1110, c: 10, d: 110, e: 1111 (not required for this course)

Entropy: -(0.4 x log(0.4) + 0.1 x log(0.1) + 0.3 x log(0.3) + 0.1 x log(0.1) + 0.1 x log(0.1))
= 2.05 bits/symbol

Average Code Length: 0.4 x 1 + 0.1 x 4 + 0.3 x 2 + 0.1 x 3 + 0.1 x 4 = 2.1 bits/symbol
(Huffman code is within 1 bit of the entropy)

Cross Entropy

Calculates the number of bits required to represent or transmit an average event from one distribution compared to another distribution

The intuition for this definition comes if we consider a target or underlying probability distribution P and an approximation of the target distribution Q. Then the cross-entropy of Q from P is average bits of information needed to identify an event drawn from the estimated probability distribution q, rather than the true distribution p

The cross-entropy between two probability distributions, such as Q from P, can be stated formally as:

- H(P, Q)

$$H(P, Q) = -\sum^{\{x \in X\}} P(x) \times \log(Q(x))$$

Where H() is the cross-entropy function, P may be the target distribution and Q is the approximation of the target distribution.

KL divergence / Relative Entropy

$$D_{\mathbb{KL}}(p \parallel q) \triangleq \sum_{k=1}^{K} p_k \log \frac{p_k}{q_k}$$

$$D_{\mathbb{KL}}(p \parallel q) \triangleq \int dx\, p(x) \log \frac{p(x)}{q(x)}$$

$$\begin{aligned}\mathbb{KL}(p\|q) &= \sum_{y\in\mathcal{Y}} p(y)\log p(y) - \sum_{y\in\mathcal{Y}} p(y)\log q(y) \\ &= -\mathbb{H}(p) + \mathbb{H}(p,q)\end{aligned}$$

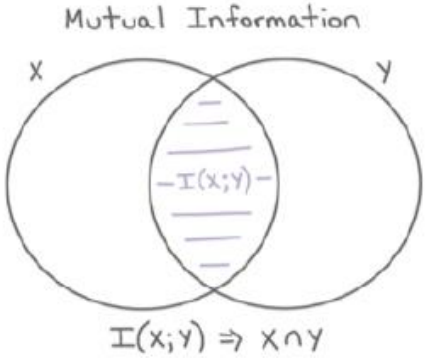
$$\mathbb{H}(p) + \mathbb{KL}(p\|q) = \mathbb{H}(p,q)$$

Mutual Information

How similar two distributions were

The mutual information between rv's X and Y is defined as follows:

$$\mathbb{I}(X;Y) \triangleq \mathbb{KL}\left(p(x,y) \| p(x)p(y)\right) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)}$$



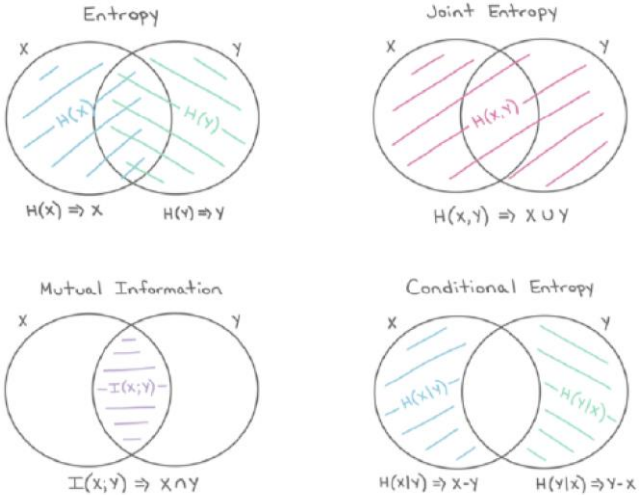
$$NMI(X,Y) = \frac{\mathbb{I}(X;Y)}{\min\left(\mathbb{H}(X), \mathbb{H}(Y)\right)} \leq 1$$

NMI = Normalized Mutual Information

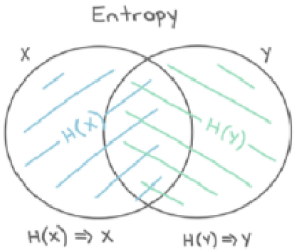
Joint Entropy

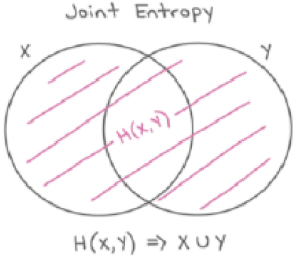
The joint entropy of two random variables X and Y is defined as

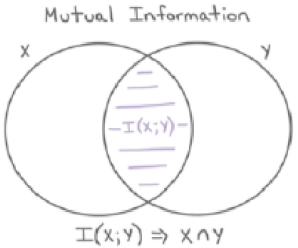
$$\mathbb{H}(X,Y) = -\sum_{x,y} p(x,y) \log_2 p(x,y)$$

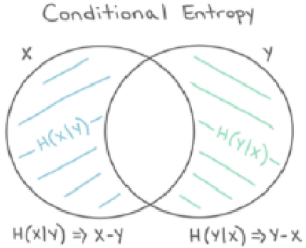


Some Relationships









$$\mathbb{I}(X;Y) = \mathbb{H}(X) - \mathbb{H}(X|Y) = \mathbb{H}(Y) - \mathbb{H}(Y|X)$$

$$\mathbb{I}(X;Y) = \mathbb{H}(X,Y) - \mathbb{H}(X|Y) - \mathbb{H}(Y|X)$$

$$\mathbb{I}(X;Y) = \mathbb{H}(X) + \mathbb{H}(Y) - \mathbb{H}(X,Y)$$

$$0 \leq \mathbb{I}(X;Y) = \mathbb{H}(X) - \mathbb{H}(X|Y)$$

$$\mathbb{H}(X|Y) \leq \mathbb{H}(X)$$

(conditioning reduces entropy)