http://people.rennes.inria.fr/Aline.Roumy/roumy_teaching.html

Notations:

- Upper case letters $X, Y, \ldots$ refer to random variables (r.v.)
- Calligraphic letters $\mathcal{X}, \mathcal{Y}, \ldots$ refer to alphabets
- $|\mathcal{A}|$ is the cardinality of the set $\mathcal{A}$
- $X^n = (X_1, X_2, \ldots, X_n)$ is an n-sequence of random variables or a random vector (r.vec.)
  $X_i^j = (X_i, X_{i+1}, \ldots, X_j)$
- Lower case $x, y, \ldots$ and $x^n, y^n, \ldots$ mean scalars/vectors realization
- $X^n \sim p(x^n)$ means that the discrete r.vec. X^n has joint probability mass function (pmf) $p(x^n)$
- $p(y^n|x^n)$ is the conditional pmf of Y^n given X^n, defined if $p(x^n) > 0$.

Entropy

Definition 1 (Entropy)**.** *The **entropy** of a **discrete** random variable $X \sim p(x)$:*

$$H(X) = -\sum_{x \in X} p(x) \log p(x) = -\mathbb{E}_X \log p(X)$$

Notation: $\log := \log_2$ **Convention:** $0 \log 0 := 0$

Theorem 1. *$H(X)$ in bits/source sample is the average length of the shortest description of the rv X.*

Properties

E3 $H(X) \geq 0$ with equality iff X deterministic The degenerate distribution (i.e. constant) has zero entropy.

E4 $H(X) \leq \log|\mathcal{X}|$. with equality iff X uniform The uniform distribution maximizes entropy.

Example 1 (Binary entropy function:)**.** $X \sim \mathcal{B}(p), 0 \leq p \leq 1$, $H(X) = h_b(p) = -p \log p - (1-p)\log(1-p)$

Joint and conditional entropy

Definition 2 (Conditional independence)**.** *Let X, Y, Z be r.v. X is independent of Z conditioning on Y, denoted by $X \perp\!\!\!\perp Z|Y$, if*

$$\forall (x,y,z),\ p(x,y,z)p(y) = p(x,y)p(y,z) \quad \Leftrightarrow \quad \forall (x,y,z), p(x,y,z) = \begin{cases} p(x,y)p(z|y) & \text{if } p(y) > 0 \\ 0 & \text{otherwise} \end{cases}$$

$X \perp\!\!\!\perp Z|Y \quad \Leftrightarrow \quad X \to Y \to Z$ forms a Markov chain $\quad \Leftrightarrow \quad Z \to Y \to X$ forms a Markov chain.

Definition 3. *For discrete random variables $(X, Y) \sim p(x, y)$,*
*the **Conditional entropy for a given x** is:*

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

*and the **Conditional entropy** is:*

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

*and the **Joint entropy** is:*

$$H(X, Y) = -\mathbb{E}_{XY} \log p(X, Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(xy) \log p(xy)$$

Properties

JCE1 $H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$

JCE2 $H(X, Y) \leq H(X) + H(Y)$ with equality iff X and Y are independent (denoted $X \perp\!\!\!\perp Y$).

JCE3 Conditioning reduces entropy: $H(Y|X) \leq H(Y)$ with equality iff $X \perp\!\!\!\perp Y$

JCE4 Chain rule for entropies (formule des conditionnements successifs).
Let X^n be a discrete random vector

$$\begin{aligned} H\left(X^n\right) &= H\left(X_1\right)+H\left(X_2 \mid X_1\right)+\ldots+H\left(X_n \mid X_{n-1}, \ldots, X_1\right) \\ &= \sum_{i=1}^n H\left(X_i \mid X_{i-1}, \ldots, X_1\right)=\sum_{i=1}^n H\left(X_i \mid X^{i-1}\right) \leq \sum_{i=1}^n H(X_i) \end{aligned}$$

with notation $H(X_1|X^0) = H(X_1)$.

JCE5 $H(X|Y) \geq 0$ with equality iff $X = f(Y)$ a.s. (i.e. w.p. 1. i.e. on the support of $p(y)$.)

JCE6 $H(X|X) = 0$ and $H(X,X) = H(X)$

JCE7 Data processing inequality. Let X be a discrete random variable and $g(X)$ be a function of X, then

$$H(g(X)) \leq H(X)$$

with equality iff $g(x)$ is injective on the support of $p(x)$.

JCE8 Fano's inequality: Let $(X,Y) \sim p(x,y)$ and $P_e = \mathbb{P}\{X \neq Y\}$, then

$$H(X|Y) \leq h_b\left(P_e\right) + P_e \log(|\mathcal{X}|-1) \leq 1 + P_e \log|\mathcal{X}|$$

JCE9 $H(X|Z) \geq H(X|Y,Z)$ with equality iff $X \perp\!\!\!\perp Y|Z$.

JCE10 $H(X,Y|Z) \leq H(X|Z) + H(Y|Z)$ with equality iff $X \perp\!\!\!\perp Y|Z$.

Mutual Information

Definition 4. *For discrete random variables $(X,Y) \sim p(x,y)$, the* ***Mutual Information*** *is:*

$$\begin{aligned} I(X;Y) &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\ &= H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y) \end{aligned}$$

Properties

MI2 $I(X;Y)$ is symmetric: $I(X;Y) = I(Y;X)$

MI3 $I(X;X) = H(X)$

MI4 $I(X;Y) = D(p(x,y)||p(x)p(y))$

MI5 $I(X;Y) \geq 0$ with equality iff $X \perp\!\!\!\perp Y$

MI6 $I(X;Y) \leq \min(H(X), H(Y))$ with equality iff X=f(Y) a.s. or Y=f(X) a.s.

Conditional Mutual Information

Definition 5. *For discrete random variables $(X,Y,Z) \sim p(x,y,z)$, the* ***Conditional Mutual Information*** *is:*

$$\begin{aligned} I(X;Y|Z) &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \sum_{z \in \mathcal{Z}} p(x,y,z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)} \\ &= H(X|Z) - H(X|Y,Z) = H(Y|Z) - H(Y|X,Z) \end{aligned}$$

Properties

CMI1 $I(X;Y|Z) \geq 0$ with equality iff $X \perp\!\!\!\perp Y|Z$

CMI2 **Chain rule**

$$I\left(X^n; Y\right) = \sum_{i=1}^n I\left(X_i; Y \mid X^{i-1}\right)$$

CMI3 If $X \to Y \to Z$ form a Markov chain, then $I(X;Z|Y) = 0$

CMI4 Corollary: If $X \to Y \to Z$, then $I(X;Y) \geq I(X;Y|Z)$

CMI5 Corollary: **Data processing inequality:**
If $X \to Y \to Z$ form a Markov chain, then $I(X;Y) \geq I(X;Z)$

CMI6 There is **no order relation** between $I(X;Y)$ and $I(X;Y|Z)$

References

[1] C.E. Shannon, "A mathematical theory of communication", *Bell Sys. Tech. Journal*, 27: 379–423, 623–656, 1948

[2] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications. Wiley, New York, 2006.

[3] A. E. Gamal and Y-H. Kim. *Lecture Notes on Network Information Theory*. arXiv:1001.3404v5, 2011.

[4] R. W. Yeung. *Information Theory and Network Coding*. Springer, August 2008.