Chapter 2. Entropy, Relative Entropy and Mutual Information

Chang-Su Kim

The contents herein are based on the book "Elements of Information Theory" by Cover and Thomas and only for the course EKE674, Korea University.

I. Entropy

The entropy of a random variable is the measure of uncertainty or information in the random variable.

⋆ Let X be a discrete random variable with alphabet $\mathcal{X}$ and probability mass function $p(x) = Pr\{X = x\}$, $x \in \mathcal{X}$. Then, the entropy $H(X)$ of X is defined as

$$H(x) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) \quad \text{(bits)},$$

where the log base is 2.

Remarks:

- The entropy does not depend on the actual values taken by the random variable X, but only on the probabilities.
- $H(X) = E \log \frac{1}{p(X)} = -E \log p(X)$.
- $1 \log 1 = 0$ and $0 \log 0 = 0$.
- $H(X) \geq 0$.

⋆ The entropy of a binary random variable

$$X = \begin{cases} 0 & \text{with probability } p \\ 1 & \text{with probability } 1-p \end{cases}$$

is given by

$$H(X) \doteq H(p) = -p \log p - (1-p) \log(1-p)$$

Remarks:

- $H(p)$ is the highest when $p = 0.5$.
- $H(p)$ is the smallest when $p = 0$ or 1.
- $H(p)$ is symmetric with respect to $p = 0.5$.

Let us consider a pair of random variables (X, Y) with a joint distribution $p(x, y)$.

⋆ Joint entropy: the information or uncertainty in X and Y

$$\begin{aligned} H(X,Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) \\ &= -E\log p(X,Y) \end{aligned}$$

⋆ Conditional entropy: the uncertainty in Y when X is known

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

⋆ Chain rules

$$H(X,Y) = H(X) + H(Y|X) \quad (1)$$

$$H(X,Y) = H(Y) + H(X|Y) \quad (2)$$

The meaning of (1): The uncertainty in X and Y is equal to the sum of the uncertainty in X and the uncertainty in Y when X is known.

II. Relative Entropy and Mutual Information

⋆ The relative entropy (or Kullback-Leibler distance) between two probability mass functions $p(x)$ and $q(x)$ is defined by

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

Remarks

- Distance should satisfy three properties:
  1. $d(\alpha, \beta) \geq 0$ with equality iff $\alpha = \beta$. (positiveness)
  2. $d(\alpha, \beta) = d(\beta, \alpha)$. (symmetry)
  3. $d(\alpha, \beta) \leq d(\alpha, \gamma) + d(\gamma, \beta)$. (triangle inequality)
- $D(p||q)$ satisfies only the first property.

⋆ Mutual information $I(X;Y)$ is the amount of information that X contains about Y or that Y contains about X.

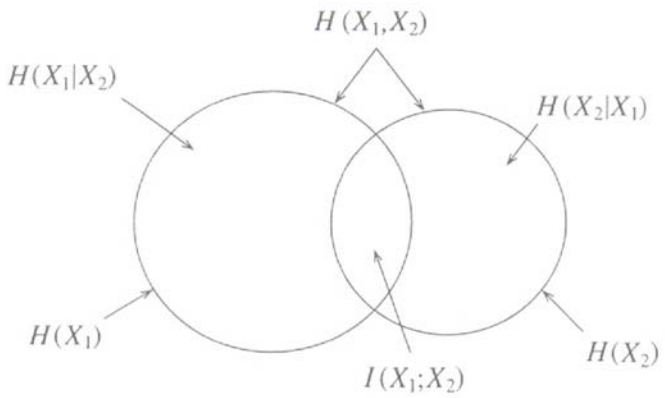
$$\begin{aligned} I(X;Y) &= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\ &= D(p(x,y)||p(x)p(y)) \end{aligned}$$

- $p(x)p(y)$ is also a probability mass function on (X,Y).
- $I(X;Y)$ is non-negative.
- $I(X;Y) = I(Y;X)$.

⋆ Properties of mutual information:

1. $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$.
2. $I(X;Y) = H(X) + H(Y) - H(X,Y)$
3. $I(X;X) = H(X)$

⋆ Information diagram rules by Raymond Yeung:



1. Sets correspond to I or H.
2. $\cup \Leftrightarrow$,
3. $\cap \Leftrightarrow$;
4. $- \Leftrightarrow$ $|$

III. More Than Three Random Variables

⋆ Chain rule:

$$H(X_1, X_2, \cdots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \cdots, X_1)$$

⋆ Conditional mutual information:

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

⋆ Chain rule for mutual information:

$$I(X_1, X_2, \cdots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_{i-1}, \cdots, X_1)$$

IV. Nonnegativity of Mutual Information

A function f is called convex if

$$f(\lambda x_1 + (1-\lambda)x_2) \le \lambda f(x_1) + (1-\lambda) f(x_2)$$

for every x_1, x_2 and $0 \le \lambda \le 1$. It is called strictly convex, if the above equality holds only if $\lambda = 0$ or 1. A function f is called concave if $-f$ is convex. For example, x^2 and e^x are strictly convex, whereas $\log x$ is strictly concave.

$\star$ Jensen's inequality:
If f is a convex function and X is a random variable, then

$$Ef(X) \ge f(EX)$$

Moreover, if f is strictly convex, the quality implies that $X = EX$ with probability 1, *i.e.*, X is a constant.

⋆ Information inequality:

$$D(p||q) \geq 0$$

with equality if and only if

$$p(x) = q(x) \text{ for all } x.$$

⋆ Nonnegativity of mutual information:

$$I(X;Y) \geq 0$$

with equality if and only if X and Y are independent.

⋆ Uniform distribution maximizes entropy:

$$H(X) \leq \log |\mathcal{X}|.$$

⋆ Conditioning reduces entropy:

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

⋆ Independence bound on entropy:

$$H(X_1, X_2, \cdots, X_n) \leq \sum_{i=1}^{n} H(X_i)$$

with equality if and only if X_i's are independent.

V. Data Processing Inequality

Random variables X, Y, Z are said to form a Markov chain, denoted by $X \to Y \to Z$, if the conditional distribution of Z depends only on Y and is conditionally independent of X. Specifically, $X \to Y \to Z$ if

$$p(x, y, z) = p(x)p(y|x)p(z|y).$$

Results:

- $X \to Y \to Z$ iff X and Z are conditionally independent given Y, *i.e.*

$$p(x, z|y) = p(x|y)p(z|y).$$

- $X \to Y \to Z$ implies $Z \to Y \to X$. Thus, we write $X \leftrightarrow Y \leftrightarrow Z$
- If $Z = f(Y)$, then $X \to Y \to Z$.

⋆ Data processing inequality:
If $X \to Y \to Z$, then $I(X;Y) \geq I(X;Z)$.
In particular, if $Z = f(Y)$, we have $I(X;Y) \geq I(X;g(Y))$