Information Theory

Shan-Hung Wu
shwu@cs.nthu.edu.tw

Department of Computer Science,
National Tsing Hua University, Taiwan

NetDB-ML, Spring 2014

Outline

1. **Entropy**
   - Definitions and Interpretations
   - The Maximum Entropy Configuration
   - Conditional Entropy

2. **Relative Entropy and Mutual Information**
   - Relative Entropy
   - Mutual Information

Outline

1. Entropy
   - Definitions and Interpretations
   - The Maximum Entropy Configuration
   - Conditional Entropy

2. Relative Entropy and Mutual Information
   - Relative Entropy
   - Mutual Information

知道一個硬幣的結果是正面 和
知道一個骰子的點數是1
後者所得到的資訊量比較大

熵
正相關於
可能狀態的數量

Entropy (1/2) 如何定義Entropy

- Given a discrete random variable x, how much information is received when we observe a specific value for this variable? 知道稀有事件發生，比知道平凡事情發生所得到的情報量更多
    - If we are told that a highly improbable event has just occurred, we will have received more information than if we were told that some very likely event has just occurred, and if we knew that the event was certain to happen we would receive no information.
- Let $h(x)$ be a monotonic function of the probability $p(x)$ and express the information content. We want $h(x)$ to have the properties

知道兩件獨立的事情(A、B)
所獲得的情報量總和
是知道A的情報量+知道B的情報量

$$h(x,y) = h(x) + h(y)$$

if two events x and y are unrelated, i.e. $p(x,y) = p(x)p(y)$.

- From these two relationships, it is easily shown that $h(x)$ must be given by the logarithm of $p(x)$ and so we have

$$h(x) = -\log_2 p(x)$$

where the negative sign ensures that information is positive or zero.

    - The choice of basis for the logarithm is arbitrary.

Entropy (2/2) 熵就是

已知分布，得到其中任一Data Point時，所獲得的資訊量的期望值

- Now suppose that a sender wishes to transmit the value of a random variable to a receiver. The average amount of information transmitted in the process is

一個分布集合的情報總量=

$$\mathrm{H}[x] = -\sum_{x} p(x) \log_2 p(x).$$

單一事件的Entropy

所有事件加總

  - This quantity is called the *entropy* of the random variable x.
  - Note that $\lim_{p\to 0} p \ln p = 0$ and so we shall take $p(x) \ln p(x) = 0$ whenever we encounter a value for x such that $p(x) = 0$.

- Next we show that these definitions indeed possess useful properties.
  - Average code length
  - A measure of disorder

Average Code Length (1/2)

- Consider a random variable x having 8 possible states, each of which is equally likely.
- In order to communicate the value of x to a receiver, we would need to transmit a message of length 3 bits. Notice that the entropy of this variable is given by

$$\mathrm{H}[x] = -8 \times \frac{1}{8} \log_2 \frac{1}{8} = 3 \text{ bits.}$$

一般Encoding

- If the probabilities of the 8 states are $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64})$ instead, the entropy is

$$\mathrm{H}[x] = 2 \text{ bits.}$$

Hoffman Encoding

  - This can be done using, for instance, the following set of code strings: $0, 10, 110, 1110, 111100, 111101, 111110, 111111$. The average code length is then

$$\frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 3 + \frac{1}{16} \times 4 + 4 \times \frac{1}{64} \times 6 = 2 \text{ bits.}$$

Average Code Length (2/2)

- In the above example, we see that the nonuniform distribution has a smaller entropy than the uniform one.
- The relation between entropy and shortest coding length is a general one. The *noiseless coding theorem* (Shannon, 1948) states that the entropy is a lower bound on the number of bits need to transmit the state of a random variable.

A Measure of Disorder (1/2)

- Let's consider a set of N identical objects that are to be divided amongst a set of bins, such that there are n_i objects in the i^{th} bin.
- The total number of ways of allocating the N objects to the bins is given by

  $$W = \frac{N!}{\prod_i n_i!}$$

  which is called the *multiplicity*.
- The entropy is then defined as the logarithm of the multiplicity scaled by an appropriate constant

  $$H = \frac{1}{N}\ln W = \frac{1}{N}\ln N! - \frac{1}{N}\sum_i \ln n_i!.$$

A Measure of Disorder (2/2)

- We now consider the limit $N \to \infty$, in which the fractions n_i/N are held fixed, and apply Stirling's approximation

$$\ln N! \simeq N \ln N - N$$

which gives

$$H = -\lim_{N\to\infty} \sum_i \left(\frac{n_i}{N}\right) \ln \left(\frac{n_i}{N}\right) = -\sum_i p_i \ln p_i$$

where we have used $\sum_i n_i = N$.

Outline

1. Entropy
   - Definitions and Interpretations
   - The Maximum Entropy Configuration
   - Conditional Entropy
2. Relative Entropy and Mutual Information
   - Relative Entropy
   - Mutual Information

The Maximum Entropy Configuration (1/4)

- The maximum entropy configuration can be found by maximizing H using a Lagrange multiplier to enforce the normalization constraint on the probabilities. Thus we maximize

$$\widetilde{H} = -\sum_i p(x_i) \ln p(x_i) + \lambda \left(\sum_i p(x_i) - 1 \right)$$

  from which we find that all of the $p(x_i)$ are equal and are given by $p(x_i) = 1/M$ where M is the total number of states x_i.

- To verify that the stationary point is indeed a maximum, we can evaluate the second derivative of the entropy, which gives

$$\frac{\partial \widetilde{H}}{\partial p(x_i) \partial p(x_j)} = -I_{ij} \frac{1}{p_i}$$

  where I_{ij} are the elements of the identity matrix.

The Maximum Entropy Configuration (2/4)

- In the case of discrete distributions, we saw that the maximum entropy configuration corresponded to an equal distribution of probabilities across the possible states of the variable.
- For a continuous variable x, we have

$$\mathrm{H}[x] = -\int p(x) \ln p(x) \, \mathrm{d}x$$

  which is called the *differential entropy*.
- In order for the maximum for $\mathrm{H}[x]$ to be well defined, it will be necessary to constrain the first and second moments of $p(x)$ as well as preserving the normalization constraint:

$$\int_{-\infty}^{\infty} p(x) \, \mathrm{d}x = 1$$

$$\int_{-\infty}^{\infty} x p(x) \, \mathrm{d}x = \mu$$

$$\int_{-\infty}^{\infty} (x-\mu)^2 p(x) \, \mathrm{d}x = \sigma^2$$

The Maximum Entropy Configuration (3/4)

- Then we maximize

$$
\begin{aligned}
&-\int p(x)\ln p(x)\,\mathrm{d}x + \lambda_1\left(\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x - 1\right) \\
&+\lambda_2\left(\int_{-\infty}^{\infty} xp(x)\,\mathrm{d}x - \mu\right) + \lambda_3\left(\int_{-\infty}^{\infty}(x-\mu)^2 p(x)\,\mathrm{d}x - \sigma^2\right).
\end{aligned}
$$

- Using the calculus of variations, we set the derivative of this functional to zero giving

$$p(x) = \exp\left\{-1+\lambda_1+\lambda_2 x+\lambda_3(x-\mu)^2\right\}.$$

- The Lagrange multipliers can be found by back substitution of this result into the constraint equations, leading finally to the result

$$p(x) = \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

and so the distribution that maximizes the differential entropy is the Gaussian.

The Maximum Entropy Configuration (4/4)

- If we evaluate the differential entropy of the Gaussian, we obtain

$$\mathrm{H}[x] = \frac{1}{2}\left\{1 + \ln\left(2\pi\sigma^2\right)\right\}.$$

  - The entropy increases as the distribution becomes broader, i.e., as σ^2 increases.
  - Note that the differential entropy, unlike the discrete entropy, can be negative, because $\mathrm{H}[x] < 0$ for $\sigma^2 < 1/(2\pi)$.

Outline

1. **Entropy**
   - Definitions and Interpretations
   - The Maximum Entropy Configuration
   - Conditional Entropy
2. **Relative Entropy and Mutual Information**
   - Relative Entropy
   - Mutual Information

Conditional Entropy

- Suppose we have a joint distribution $p(x,y)$ from which we draw pairs of values of x and y.
- If a value of x is already known, then the additional information needed to specify the corresponding value of y is given by $-\ln p(y|x)$.
- Thus the average additional information needed to specify y can be written as

  $$\mathrm{H}[y|x] = -\iint p(y,x)\ln p(y|x)\,\mathrm{d}y\mathrm{d}x$$

  which is called the *conditional entropy* of y given x.
- It is easily seen, using the product rule, that the conditional entropy satisfies the relation

  $$\mathrm{H}[x,y] = \mathrm{H}[y|x] + \mathrm{H}[x].$$

  - $\mathrm{H}[x,y]$ is called the *joint entropy* of x and y.
  - Thus the information needed to describe x and y is given by the sum of the information needed to describe x alone plus the additional information required to specify y given x.

Outline

1. Entropy
   - Definitions and Interpretations
   - The Maximum Entropy Configuration
   - Conditional Entropy

2. Relative Entropy and Mutual Information
   - Relative Entropy
   - Mutual Information

Relative Entropy (1/2)

- Consider some unknown distribution $p(x)$, and suppose that we have modeled this using an approximating distribution $q(x)$.
- If we use $q(x)$ to construct a coding scheme for the purpose of transmitting values of x to a receiver, then the average additional amount of information required to specify the value of x as a result of using $q(x)$ instead of the true distribution $p(x)$ is given by

$$\begin{aligned}\mathrm{KL}(p\|q) &= -\int p(x)\ln q(x)\,\mathrm{d}x - \left(-\int p(x)\ln p(x)\,\mathrm{d}x\right) \\ &= -\int p(x)\ln\left\{\frac{q(x)}{p(x)}\right\}\mathrm{d}x.\end{aligned}$$

  - This is known as the *relative entropy* or *Kullback-Leibler (KL) divergence* between the distributions $p(x)$ and $q(x)$.
  - Note that $\mathrm{KL}(p\|q) \neq \mathrm{KL}(q\|p)$.

Relative Entropy (2/2)

- We can show that the KL divergence satisfies $\mathrm{KL}(p||q) \geqslant 0$ with equality iff $p(x) = q(x)$ using *Jensen's inequality*:

$$f\left(\sum_{i=1}^{M} \lambda_i x_i\right) \leqslant \sum_{i=1}^{M} \lambda_i f(x_i)$$

for a convex function $f(x)$ where $\lambda_i \geqslant 0$ and $\sum_i \lambda_i = 1$.

- If we interpret the λ_i as the probability distribution over a discrete variable x taking the values $\{x_i\}$, then the inequality can be written

$$f(\mathbb{E}[x]) \leqslant \mathbb{E}[f(x)].$$

- For continuous variables, Jensen's inequality takes the form

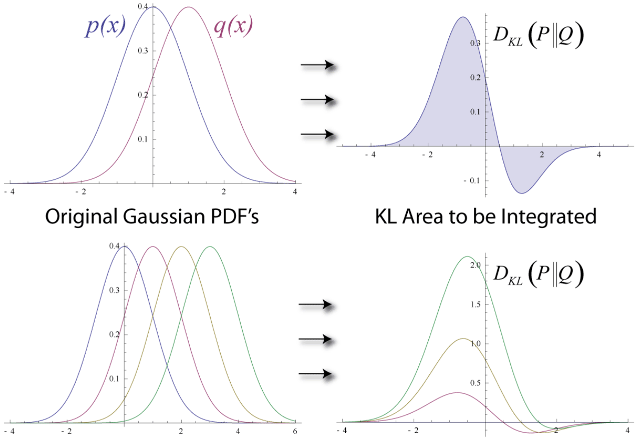
$$f\left(\int x p(x)\,\mathrm{d}x\right) \leqslant \int f(x) p(x)\,\mathrm{d}x.$$

- Since $-\ln x$ is a strictly convex function, we have

$$\mathrm{KL}(p||q) = -\int p(x) \ln\left\{\frac{q(x)}{p(x)}\right\} \mathrm{d}x \geqslant -\ln \int q(x)\,\mathrm{d}x = 0$$

and the equality holds iff $p(x) = q(x)$.

Illustration of the Relative Entropy



Original Gaussian PDF's

KL Area to be Integrated

Figure : Illustration of the relative entropy or KL divergence for two normal distributions. Note that the typical asymmetry is clearly visible.

Outline

1. Entropy
   - Definitions and Interpretations
   - The Maximum Entropy Configuration
   - Conditional Entropy

2. Relative Entropy and Mutual Information
   - Relative Entropy
   - Mutual Information

Mutual Information

- Consider the joint distribution between two variables x and y given by $p(x, y)$. If they are independent, then their joint distribution will factorize into the product of their marginals $p(x, y) = p(x)p(y)$.
- If they are not independent, we can gain some idea of whether they are *close* to being independent by considering the KL divergence between the joint distribution and the product of the marginals, given by

$$\begin{aligned} \mathrm{I}[x, y] &\equiv \mathrm{KL}(p(x, y) \| p(x)p(y)) \\ &= -\iint p(x, y) \ln\left(\frac{p(x)p(y)}{p(x, y)}\right) \mathrm{d}x\mathrm{d}y \end{aligned}$$

  which is called the *mutual information* between the variables x and y.
- Using the sum and product rules of probability, we see that the mutual information is related to the conditional entropy through

$$\mathrm{I}[x, y] = \mathrm{H}[x] - \mathrm{H}[x|y] = \mathrm{H}[y] - \mathrm{H}[y|x].$$

Relation between Entropy and Mutual Information

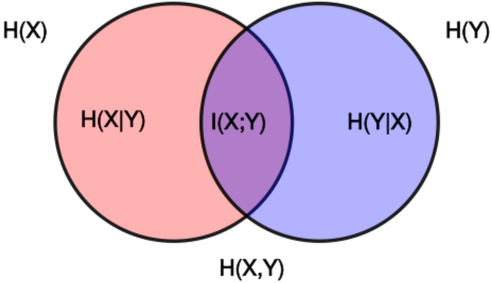


Figure : Individual $(\mathrm{H}[\mathrm{X}], \mathrm{H}[\mathrm{Y}])$, joint $(\mathrm{H}[\mathrm{X}, \mathrm{Y}])$, and conditional entropies for a pair of random variables X, Y with mutual information $\mathrm{I}[\mathrm{X}, \mathrm{Y}]$.