Information Theory (1)

LC 1-9, 1-10

Lecture 24, 2008-12-12

Content

- Information
- Entropy
- Source Rate
- Discrete Channel Models
- Joint and Conditional Entropy
- Mutual Information
- Channel Capacity

Basic Concepts

- The purpose of communication systems is to transmit information from a source to receiver. However, what exactly is information, and how do we measure it?
- Example
- At the end of this class, I make one of the following statements to the class:

  A. I will see you next period

  B. Due to some reasons, I will miss next lecture.

  C. Everyone gets an A in the course, and there will be no more homework and projects.
- What is the relative information conveyed to you?

  There is little information conveyed by statement A, since we have a schedule. That is, the probability, P(A) is nearly 1.

  Intuitively, we know that statement B contains more information, and the probability that I miss one lecture P(B) is relatively low.

  Statement C contains a vast amount of information for the entire class, and such a statement has a very low probability of occurrence.
- Information is defined consistent with this intuitive example.

Information

- Definition. The information sent from a digital source when the jth message is transmitted is given by

$$I_j = \log_2\left(\frac{1}{P_j}\right) \text{ bits}$$

Where Pj is the probability of transmitting the jth message.

- From this definition, we see that messages that are less likely to occur (smaller value for pj) provide more information (larger value of Ij). The information does not depend on possible interpretation of the content as to whether or not it makes sense.
- The base of the logarithm determines the units used for the information measure. Thus, for units of “bits”, the base 2 logarithm is used. If the natural logarithm is used, the units are “nats” and for base 10 logarithms, the unit is the “hartley”, named after R. V. Hartley, who first suggest using the logarithm measure in 1928.

Example

- Consider a random experiment with 16 equally likely outcomes. The information associate with each outcome is

$$I_j = \log_2\left(\frac{1}{1/16}\right) = \log_2 16 = 4 \text{ bits}$$

Where j ranges from 1 to 16.

Average Information

- Definition. The **average information** measure of a digital source is

$$H = \sum_{j=1}^{m} P_j I_j = \sum_{j=1}^{m} P_j \log_2 \left(\frac{1}{P_j} \right) \text{ bits}$$

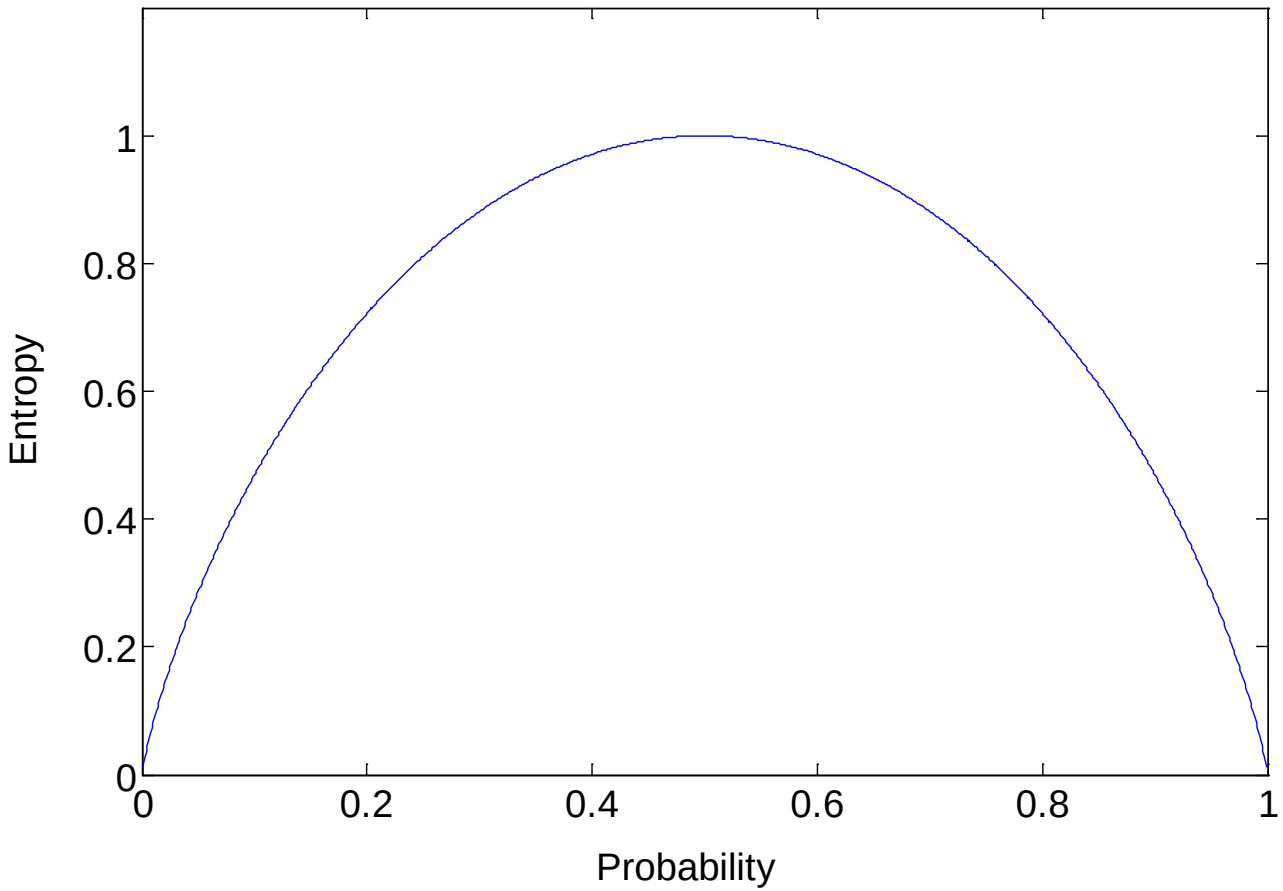
Where m is the number of possible different source message and Pj is the probability of transmitting the jth message. The average information is also called entropy.

- Originally, the concept of entropy was introduced in Thermodynamics, which is an important branch of physics.

Energy vs. Information

- Energy can exist in one region of space or another, can flow from here to there, can be stored for later use, and can be converted from one form to another.
- In the context of thermodynamics, the conversation of energy principle is known as the First law.
- Like energy, information can reside in one place or another, it can be transmitted through space, and it can be stored for later use.
- But unlike energy, information is inherently subjective because it deals with what you know and what you don't know.
- Also, information is not conserved as is energy. The second law states that entropy never decrease as time goes on.

Entropy of a Binary Source



$$H = -p \log_2 p - (1-p) \log_2 (1-p)$$

Source Rate

- Definition. The **source rate** is given by

$$R = \frac{H}{T} \text{ bits/s}$$

Where H is average information and T is the time required to send a message.

Example

- A source with bandwidth 4000 Hz is sampled at the Nyquist rate. Assuming that the resulting sequence can be approximately modeled by a discrete memoryless source with alphabet A={-2, -1, 0, 1, 2} and with corresponding probablities {1/2, ¼, 1/8, 1/16, 1/16}, determine the source rate

Entropy

$$H = \frac{1}{2}\log_2 2 + \frac{1}{4}\log_2 4 + \frac{1}{8}\log_2 8 + \frac{1}{16}\log_2 16 + \frac{1}{16}\log_2 16$$

$$= \frac{1}{2} + \frac{2}{4} + \frac{3}{8} + \frac{4}{16} + \frac{4}{16} = \frac{15}{8} \text{ bits/sample}$$

Sample rate

$$S = 2 \times 4000 = 8000 \text{ samples/sec}$$

Source rate

$$R = H \cdot S = \frac{15}{8} \times 8000 = 15000 \text{ bits/sec}$$

Discrete Channel Models

- If the values that the input and output variables can take finite, or countably infinite, the channel is called a discrete channel.
- In general, the output yi does not only depend on the input at the same time xi but also on the previous inputs (inter-symbol interference).
- Memoryless channel is defined that the channel output at a given time is a function of the channel input at that time and is not a function of previous channel inputs.
- For memoryless discrete channel, the conditional probability pij=p(yj|xi) of obtaining output yj given that the input is xi is called a channel transition probability.

Transition Probabilities Matrix

The transition probabilities matrix

$$[P(Y \mid X)] = \begin{bmatrix} p(y_1 \mid x_1) & p(y_2 \mid x_1) & p(y_3 \mid x_1) \\ p(y_1 \mid x_2) & p(y_2 \mid x_2) & p(y_3 \mid x_2) \end{bmatrix}$$

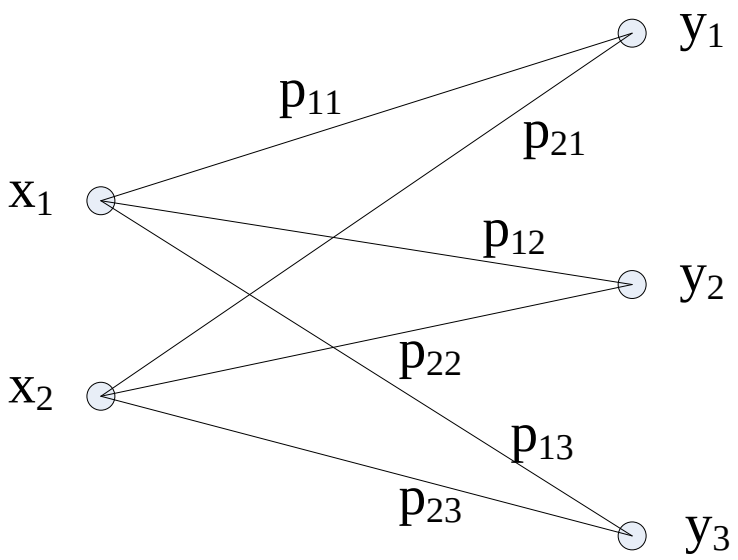
Input probabilities

$$[P(X)] = [\, p(x_1) \quad p(x_2) \,]$$

Output probabilities

$$[P(Y)] = [\, p(y_1) \quad p(y_2) \quad p(y_3) \,]$$

$$[P(Y)] = [P(X)][P(Y \mid X)]$$



Joint Entropy

- Definition. The **joint entropy** of two discrete random variables (X, Y) is defined by

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

For the case of n random variables

$$\mathbf{X} = (X_1, X_2, ..., X_n)$$

$$H(\mathbf{X}) = -\sum_{x_1, x_2, ..., x_n} p(x_1, x_2, ..., x_n) \log_2 p(x_1, x_2, ..., x_n)$$

Conditional Entropy

- Definition. The **conditional entropy** of the random variables X given the random variable Y is defined by

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

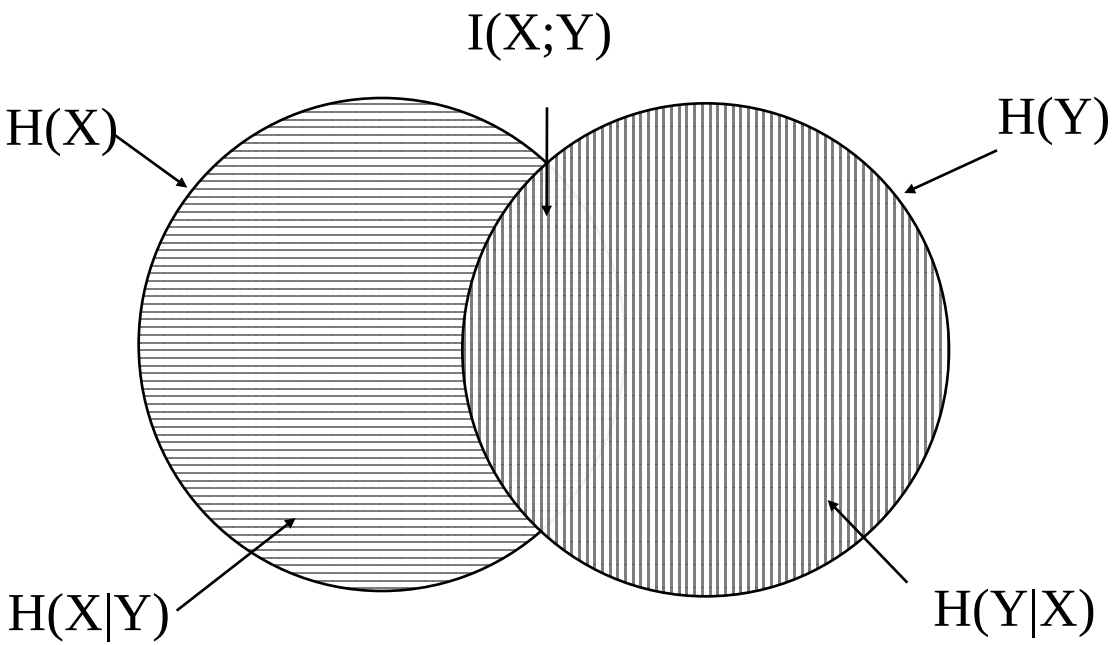
- Theorem H(X,Y)=H(Y)+H(X|Y)

$$\begin{aligned} H(X,Y) &= -\sum_{x,y} p(x,y) \log_2 p(x,y) \\ &= -\sum_{x,y} p(x,y) \log_2 \left[p(y) p(x \mid y) \right] \\ &= -\sum_{x,y} p(x,y) \log_2 p(y) - \sum_{x,y} p(x,y) \log_2 p(x \mid y) \\ &= -\sum_{y} p(y) \log_2 p(y) - \sum_{x,y} p(x,y) \log_2 p(x \mid y) \\ &= H(Y) + H(X \mid Y) \end{aligned}$$

Mutual Information

- Definition. The **mutual information** between two discrete random variables X and Y is denoted by I(X;Y) and defined by

$$I(X;Y) = H(X) - H(X \mid Y)$$



Entropy, conditional entropy, and mutual information

Example

- Let X and Y be binary random variables with P(x=0,y=0)=1/3, P(x=1,y=0)=1/3, P(x=0,y=1)=1/3, Find I(X;Y)

$$P(x=0) = P(x=0, y=0) + P(x=0, y=1) = 2/3,$$

$$P(y=0) = 2/3, \quad P(x=1) = 1/3, \quad P(y=1) = 1/3$$

$$H(X) = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} = -\frac{2}{3} + \log_2 3 = 0.919$$

$$H(Y) = 0.919$$

$$H(X,Y) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{1}{3}\log_2\frac{1}{3} - \frac{1}{3}\log_2\frac{1}{3} = \log_2 3 = 1.585$$

$$H(X \mid Y) = H(X,Y) - H(Y) = 0.666$$

$$I(X;Y) = H(X) - H(X \mid Y) = 0.253$$

Channel Capacity

- Is it possible to invent a system with no bit error at the output even when we have noise introduced into the channel?
- **Shannon noisy channel-coding theorem**. Reliable transmission without bit error is possible even over noisy channel as long as the transmission rate is less than **channel capacity** $C=\max[I(X;Y)]$.
- The maximum is with respect to the source probabilities are fixed by the channel.

Example

- Binary symmetric channel

$$I(X;Y) = H(Y) - H(Y \mid X)$$

$$H(Y \mid X) = -\sum_{i=1}^{2}\sum_{j=1}^{2} p(x_i, y_j)\log_2 p(y_j \mid x_i)$$

$$= -\alpha(1-\varepsilon)\log_2(1-\varepsilon) - \alpha\varepsilon\log_2\varepsilon$$

$$-(1-\alpha)(1-\varepsilon)\log_2(1-\varepsilon) - (1-\alpha)\varepsilon\log_2\varepsilon$$

$$= -(1-\varepsilon)\log_2(1-\varepsilon) - \varepsilon\log_2\varepsilon$$

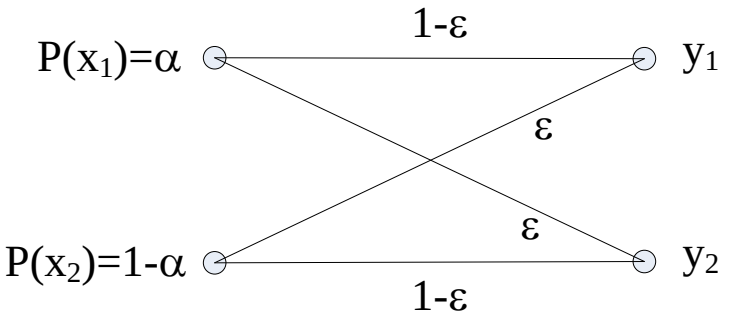
Thus

$$I(X;Y) = H(Y) + (1-\varepsilon)\log_2(1-\varepsilon) + \varepsilon\log_2\varepsilon$$

which is maximum when H(Y) is maximum.

$$Max[H(Y)] = 1$$

$$C = Max[I(X;Y)] = 1 + (1-\varepsilon)\log_2(1-\varepsilon) + \varepsilon\log_2\varepsilon$$



Gaussian Channel Capacity

- Shannon showed that (for the case of signal plus white Gaussian noise) a channel capacity C (bits/s) could be calculated such that if the rate of information R (bits/s) was less than C, the probability of error would approach zero. The equation for C is

$$C = B\log_2(1+\frac{S}{N})bits/s$$

where B is the channel bandwidth in Hz, and S/N is the signal-to-noise power ratio (watts/watts, not dB) at the input to the digital receiver. Systems approach this bound usually incorporate error-correction coding.

Example

- B=3000Hz, SNR=39dB

39dB=7943

$C=3000\log_2(1+7943)=38.867$ bps

Homework

- LC 1-6, 1-7, 1-9, 1-14