Chapter 3: Information measure

3.1. Information measure

3.2. Review probability (reading at home)

3.3. Amount of information

3.4. Entropy

3.5. Mutual information

3.1.Information measure (cont.)

- Measure?

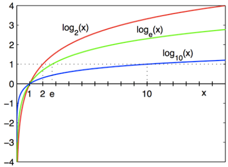
→Measure of a quantity is a numerical value that allows determining the magnitude of that quantity

- The measure must be:
  - Not negative
  - Reality reflection: the more significant the quantity to be measured, the greater the value of the measure
  - Linear: the measure of 2 entities must be equal to the sum of 2 measures if two entities are independent

3.1.Information measure (cont.)

- Uncertainty is inversely proportional to appearance probability $f(\frac{1}{p(x)})$
- To ensure the linearity
  - Two independent information x1, x2 with probability p(x1), p(x2) then
    - Uncertainty of x1 is $f(\frac{1}{p(x1)})$
    - Uncertainty of x2 is $f(\frac{1}{p(x2)})$
    - x1, x2 appear simultaneously: $p(x1, x2) = p(x1,)\ p(x2)$
    - Joint amount of information $f(\frac{1}{p(x1,x2)}) = f(\frac{1}{p(x1)p(x2)})$
    - Linearity: $f(\frac{1}{p(x1)p(x2)}) = f(\frac{1}{p(x1)}) + f(\frac{1}{p(x2)})$

      → f must be log function
  - $0 \leq p(x) \leq 1$ → $\log(\frac{1}{p(x)}) \geq 0$



→ $\log(\frac{1}{p(x)})$ is measure of uncertainty

3.1.Information measure (cont.)

- For a discrete source with finite alphabet $X = \{x_1, x_2, \ldots, x_m\}$ where the probability of each symbol is given by $P(X = x_k) = p_k$

$$I(x_k) = \log \frac{1}{p_k} = -\log p_k$$

- If logarithm is base 2, information unit is given in bit (binary unit).
- If logarithm is base e, information unit is given in nat (natural unit).
- If logarithm is base 10, information unit is given in Hartley

3.1.Information measure (cont.)

- It represents the *surprise* of seeing the outcome (a highly probable outcome is not surprising).

Event	**Probability**	**Surprise**
one equals one	1	0 bits
wrong guess on a 4-choice question	3/4	0.415 bits
correct guess on true-false question	1/2	1 bit
correct guess on a 4-choice question	1/4	2 bits
seven on a pair of dice	6/36	2.58 bits
win a Jackpot	≈1/76 million	≈26 bits

3.2. Review of probability (homework)

- Probability: a measure of the probability of an event occurring
- Determining whether an event can occur or not is done by tests
- Based on a triplet

$$(\Omega, F, P)$$

Where:

- Ω: sample space
    - Set of all possible outcomes
- F: σ−algebra
    - Set of all possible events or combinations of outcomes
- P: probability function
    - Any set function
        - Domain is Ω
        - Range is the closed unit interval [0,1]

3.2. Review of probability (cont.)

- P must obey the following rules:
  - $P(\Omega) = 1$
  - Let A be any event in F, then $P(A) \geq 0$
  - Let A and B be two events in F such that $A \cap B = \emptyset$, then $P(A \cup B) = P(A) + P(B)$.
  - Probability of complement: $P(A) = 1 - P(A)$.
  - $P(A) \leq 1$.
  - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

3.2. Review of probability (cont.)
Bayes rule

- If *A* and *B* are events:

$$P(A|B) = \frac{P(B/A)P(A)}{P(B)}$$

3.2. Review of probability (cont.)
Total Probability Theorem

- A set of B_i, $i = 1, \ldots, n$ of events is a partition of Ω when:
  - $\bigcup_{i=1}^{n} B_i = \Omega$.
  - $B_i \cap B_j = \emptyset$, if $i \neq j$.
- Theorem: If A is an event and B_i, $i = 1, \ldots, n$ of is a partition of Ω, then:

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$$

3.2. Review of probability
Conditional probability (Cont.)

- Let A and B be two events, with $P(A) > 0$. The conditional probability of B given A is defined as:

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

- Hence, $P(A \cap B) = P(B|A)P(A) = P(A|B)P(B)$
- If $A \cap B = \emptyset$ then $P(B|A) = 0$
- If $A \subset B$, then $P(B|A) = 1$

3.2. Review of probability
Independence between Events

- Two events A and B are statistically independent when

$$P(A \cap B) = P(A)P(B)$$

- Supposing that both $P(A)$ and $P(B)$ are greater than zero, from the above definition we have that:

$$P(A|B) = P(A) \qquad P(B|A) = P(B)$$

3.2. Review of probability
Independence between Events (Cont.)

- N events are statistically independent if the intersection of the events contained in any subset of those N events have probability equal to the product of the individual probabilities
- Example: Three events A, B and C are independent if:

$$P(A \cap B) = P(A)P(B), \quad P(A \cap C) = P(A)P(C), \quad P(B \cap C) = P(B)P(C)$$

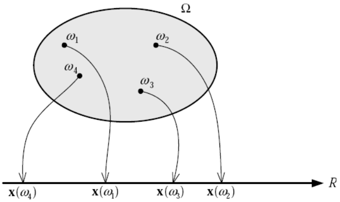
$$P(A \cap B \cap C) = P(A)P(B)P(C)$$

3.2. Review of probability
What is random variables?

- A random variable (rv) is a function that maps each $\omega \in \Omega$ to a real number

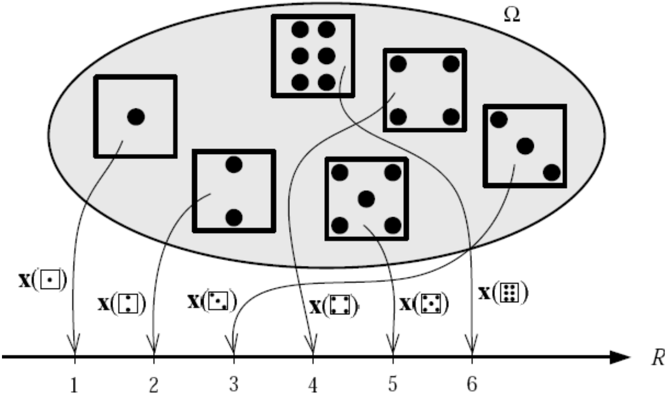
$$X \ : \ \Omega \to \mathrm{R}$$
$$\omega \to X(\omega)$$

E.g:



- A random variable is a *mapping* from the sample space Ω to the set of real numbers.

3.2. Review of probability
What is random variables? (cont.)



3.2. Review of probability
What is random variables? (cont.)

- Random variable is a representation in numeric form of sample space. Thus,
  - We can process events with digital computing devices
  - Values corresponding to the outcomes are sorted

3.2. Review of probability
Cumulative Distribution Function (CDF)

- CDF of a random variable X with a given value x ($F_X(x)$) is the probability for a random variable X has a value that does not exceed x

$$F_X(x) = P(X \le x)$$

- $F_X(x)$ (∞) = 1
- $F_X(x)$ ($-\infty$) = 0
- If $x_1 < x_2$, $F_X(x_2) \ge F_X(x_1)$

3.2. Review of probability
Types of random variable

- Discrete: Cumulative function is a step function (sum of unit step functions)

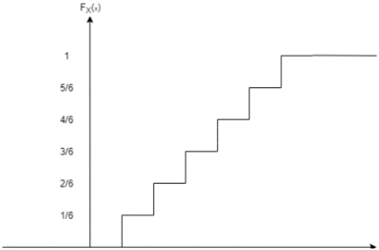
$$F_X(x)=\sum_i P\ (X = x_i)\ u\ (x\ - x_i)$$

where $u(x)$ is unit step function

3.2. Review of probability
Types of random variable (Cont.)

Example: X is the random variable that describes the outcome of the roll of a die $X \in \{1, 2, 3, 4, 5, 6\}$

P(x)=1/6 for all x



- Continous: Cumulative function is a continous function.
- Mixed: Neither discrete nor continous.

3.2. Review of probability
Probability Density Function (PDF)

- It is the derivative of the cumulative distribution function:

$$p_X(x) = \frac{d}{dx} F_X(x)$$

- $\int_{-\infty}^{x} p_X(x)dx = F_X(x)$.
- $p_X(x) \geq 0$.
- $\int_{-\infty}^{\infty} p_X(x)dx = 1$.
- $\int_{a}^{b} p_X(x)dx = F_X(b) - F_X(a) = P(a \leq X \leq b)$.

3.2. Review of probability
Discrete Random vector

Let Z = [X , Y] be a random vector with sample space Z= $X \times Y$
X and Y are random variables have sample space

The joint probability distribution function of Z is mapping

$$p_Z(z) : \mathcal{Z} \to [0, 1] \text{ satisfying:}$$

$$\sum_{Z \in \mathcal{Z}} p_Z(z) = \sum_{x,y \times \mathcal{Y}} p_{XY}(x, y) = 1$$

3.2. Review of probability
Discrete Random vector (Cont.)

- Marginal Distributions:

$$p_X(x) = \sum_{y \in Y} p_{XY}(x, y)$$

$$p_Y(y) = \sum_{x \in X} p_{XY}(x, y)$$

3.2. Review of probability
Discrete Random vector (Cont.)

- Conditional Distributions:

$$p_{X|Y=y}(\mathrm{x}) = \frac{p_{XY}(x,y)}{p_Y(y)}$$

$$p_{Y|X=x}(\mathrm{y}) = \frac{p_{XY}(x,y)}{p_X(x)}$$

3.2. Review of probability
Discrete Random vector (Cont.)

- Random variables X and Y are independent if and only if

$$p_{XY}(x, y) = p_X(x)p_Y(y)$$

- Consequences:

$$p_{X|Y=y}(x) = p_X(x)$$
$$p_{Y|X=x}(y) = p_Y(y)$$

3.2. Review of probability
Moments of a Discrete Random Variable

- The $n-$th order moment of a discrete random variable X is defined as:

$$E[X^n] = \sum_{x \in \mathcal{X}} x^n p_X(x)$$

- if $n = 1$, we have the mean of X, $m_X = E[X]$.
- The $m-$th order central moment of a discrete random variable X is defined as:

$$E[(X - m_X)^m] = \sum_{x \in \mathcal{X}} (x - m_X)^m p_X(x)$$

- if $m = 2$, we have the variance of X, σ_X^2.

3.2. Review of probability
Moments of a Discrete Random Variable

- The joint moment $n-$th order with relation to X and $k-$th order with relation to Y:

$$m_{nk} = E[X^n Y^k] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} x^n y^k p_{XY}(x, y)$$

- The joint central $n-$th order with relation to X and $k-$th order with relation to Y:

$$\mu_{nk} = E[(X - m_X)^n (Y - m_Y)^k] == \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} (x - m_X)^n (y - m_Y)^k p_{XY}(x, y)$$

3.2. Review of probability
Moments of a Discrete Random Variable

- The correlation of two random variables X and Y is the expected value of their product (joint moment of order 1 in X and order 1 in Y):

$$Corr(X, Y) = m_{11} = E[XY]$$

- The covariance of two random variables X and Y is the joint central moment of order 1 in X and order 1 in Y:

$$Cov(X, Y) = \mu_{11} = E[(X - m_X)(Y - m_Y)]$$

- $Cov(X, Y) = Corr(X, Y) - m_X m_Y$
- Correlation Coefficient:

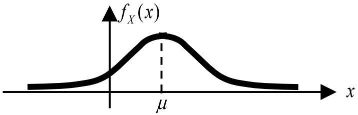
$$\rho_{XY} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} \quad \rightarrow \quad -1 \leq \rho_{XY} \leq 1$$

3.2. Review of probability

Examples of Random variables (Cont.)

- Normal (Gaussian) random variables
  - Density function

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$



  - Bell shape curve
  - Distribution function

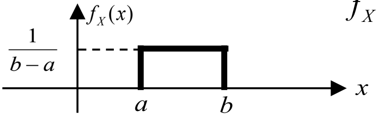
$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-\mu)^2/2\sigma^2} dy = G\left(\frac{x-\mu}{\sigma}\right),$$

$$G(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$$

3.2. Review of probability

Examples of Random variables (Cont.)

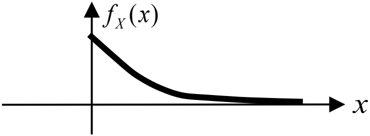
- Uniform random variables: $X \sim U(a,b),\ a<b,$
  - Density function



$$f_X(x)=\begin{cases}\dfrac{1}{b-a}, & a \le x \le b,\\ 0, & \text{otherwise.}\end{cases}$$

- Exponential random variables $X \sim \varepsilon(\lambda)$
  - Density function

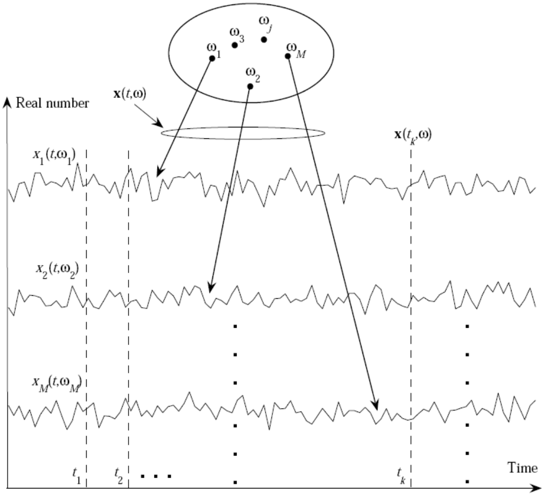
$$f_X(x)=\begin{cases}\dfrac{1}{\lambda}e^{-x/\lambda}, & x \ge 0,\\ 0, & \text{otherwise.}\end{cases}$$



3.2. Review of probability
Random process

- Random process is a set of potential random variables (realization, member) to describe a physical object
- Example:
  - To describe the movements of an animal, it is necessary to have different cameras
    - Sample space of each camera is a random variable

3.2. Review of probability
Random process



A mapping from a sample space to *a set of time functions.*

3.2. Review of probability
Random process

- To describe the random process, we need to have a set of values
- It is necessary to have the probability of appearing simultaneously at one time on a time function of the realization

$$f_{\mathbf{x}(t_1),\mathbf{x}(t_2)}(x_1, x_2; t_1, t_2)$$

3.2. Review of probability
Random process

- Stationary random process:
  - probability distribution function does not depend on original time (invariant to shift in the time).
- Ergodic random process: stationary random process and statistically average values equal average values over time
  - One realization can represent the whole ergodic random process

3.3. Amount of information

- Remind:
  - A source has mathematically model is a random variable

    → a source corresponds with a random variable
  - Information can be thought of as the resolution of uncertainty. It is abstract concept that describes understanding of objects in social life, in nature.

  → Amount of information equals uncertainty

  → Calculate amount of information through uncertainty

3.3. Amount of information (Cont.)

- To calculate amount of information of a message, it is necessary to know the message
  - In many ways, we cannot know the message
  - We can only identify the number of information in the message (length of message)

  → estimate amount of information of message using average amount of information of a source
    - denoted by $I(X) = E\{I(x_k)\}$

3.4. Entropy

3.4.1. Definition

3.4.2. Entropy of binary source

3.4.3. Joint entropy

3.4.4. Conditional entropy

3.4.5. Relationship between entropies

3.4.6. Example

3.4.7. Relative Entropy: Kullback-Leibler Distance

3.4.1. Definition

- Expected value of information from a source. It also be considered as quantity of uncertainty of a source.
- Denoted by H(X)

$$H(X) = E[I(x_k)] = \sum_{x \in \mathcal{X}} p_x(x) I(x_k)$$
$$= -\sum_{x \in \mathcal{X}} p_x(x) \log p_x(x)$$

- Properties of entropy:
  - $0 \leq H(X) \leq H(X)max$
  - $H(X)max = \log|X|$ when X has uniform distribution
  - $|X|$: cardinality of set X

3.4.1. Definition (Cont.)

- E.g. 1 : Source X = {a,b}, P(X) ={0.5,0.5}
  - Entropy H(X) = - 0.5 $log_2$0.5 - 0.5 $log_2$0.5 = 1 bit/information
- E.g.2: Source X = {a,b}, P(X) ={0.75,0.25}
  - Entropy H(X) = - 0.75 $log_2$0.75 - 0.25 $log_2$0.25 = 0.81 bits/information
- E.g.3: Source X = {a,b,c,d}, P(X) = {0.25,0.25,0.25,025}
  - Entropy H(X) = - 0.25 $log_2$0.25 - 0.25 $log_2$0.25 - 0.25 $log_2$0.25 - 0.25 $log_2$0.25 = 2 bits/information

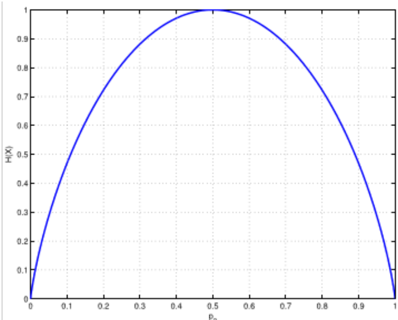
→With entropy, two sources can be compared:

- Source has higher entropy will transmit information with higher rate

3.4.2. Entropy of binary source

- Let X be a binary source with p_1 and p_2 being the probability of symbol $x_0$and x_1, respectively

$$
\begin{aligned}
H(X) &= -p_0 \log p_0 - p_1 \log p_1 \\
&= -p_0 \log p_0 - (1-p_0)\log(1-p_0)
\end{aligned}
$$



3.4.3. Joint entropy

- The joint entropy of a pair of random variables X and Y is given by:

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

3.4.4. Conditional entropy

- Average amount of information of a random variable given the occurence of other

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

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

3.4.5. Relationship between entropies

- The entropy of a pair of random variables is equal to the entropy of one of them plus the conditional entropy:

$$
\begin{aligned}
H(X, Y) &= H(X) + H(Y|X) \\
&= H(Y) + H(X|Y)
\end{aligned}
$$

- Corollary:

$$
\begin{aligned}
H(X, Y|Z) &= H(X|Z) + H(Y|X, Z) \\
&= H(Y|Z) + H(X|Y, Z)
\end{aligned}
$$

3.4.5. Relationship between entropies (Cont.)

$$H(X_1, X_2, \ldots, X_M) = \sum_{j=1}^{M} H(X_j | X_1, \ldots, X_{j-1})$$

- M: number of random variables
- Conditional random variables X_i appear before X_j

3.4.6. Examples

- Source X,Y = $\begin{pmatrix} x0, y0 & x0, y1 \\ x1, y0 & x1, y1 \end{pmatrix}$ with probability P(X,Y)= $\begin{pmatrix} 0.25 & 0.25 \\ 0.25 & 0.25 \end{pmatrix}$
- Joint entropy H(X,Y)

H(X,Y) $= -$ P(x0,y0)logP(x0,y0) $-$ P(x0,y1)logP(x0,y1) $-$ P(x1,y0)logP(x1,y0) $-$ P(x1,y1)logP(x1,y1) = $4 \times 0.25 log_2 0.25 = 2$ bits/information

- Entropy H(X)

H(X) = -P(x0)logP(x0) $-$ P(x1)logP(x1)

  - P(x0) = P(x0,y0) $+$ P (x0,y1) $=$ 0.5 (marginal probability)
  - P (x1) = 1 $-$ P(x0) $=$ 0.5

→H(X) = $-\ 2 \times 0.5 log_2 0.5 = 1$ bit/information

- Entropy H(Y)

H(Y) = $-$ P(y0)logP(y0) $-$ P(y1)logP(y1)

  - P(y0) $=$ P(x0,y0) $+$ P (x1,y0) $=$ 0.5 (marginal probability)
  - P (y1) $=$ 1 $-$P(y0) = 0.5

→H(Y) $=\ -2 \times 0.5 log_2 0.5 = 1$ bit/information

3.4.6. Examples (Cont.)

- Conditional entropy H(X|Y)

H(X|Y) = - P(x0,y0)logP(x0|y0) - P(x1,y0)logP(x1|y0)

- P(x1,y0)logP(x1|y0) - P(x1,y1)logP(x1|y1)

P(x0|y0) = $\frac{P(x0,y0)}{P(y0)} = \frac{0.25}{0.5}$ = 0.5

P(x0|y1) = $\frac{P(x0,y1)}{P(y1)} = \frac{0.25}{0.5}$ = 0.5

P(x1|y0) = $\frac{P(x1,y0)}{P(y0)} = \frac{0.25}{0.5}$ = 0.5

P(x1|y1) = $\frac{P(x1,y1)}{P(y1)} = \frac{0.25}{0.5}$ = 0.5

→ H(X|Y) = - 4 x 0.25 $log_2$0.5 = 1 bit/information

3.4.6. Examples (Cont.)

- Conditional entropy

H(Y|X) = - P(x0,y0)logP(y0|x0) - P(x1,y0)logP(y0|x1)

- P(x0,y1)logP(y1|x0) - P(x1,y1)logP(y1|x1)

P(y0|x0) = $\frac{P(x0,y0)}{P(x0)} = \frac{0.25}{0.5}$ = 0.5

P(y0|x1) = $\frac{P(x0,y1)}{P(x1)} = \frac{0.25}{0.5}$ = 0.5

P(y1|x0) = $\frac{P(x1,y0)}{P(x0)} = \frac{0.25}{0.5}$ = 0.5

P(y1|x1) = $\frac{P(x1,y1)}{P(x1)} = \frac{0.25}{0.5}$ = 0.5

→ H(X|Y) = - 4 x 0.25 $log_2$0.5 = 1 bit/information

- H(X,Y) = H(X) + H(Y|X) = 1 + 1 = 2 bit/information

3.4.7. Relative Entropy: Kullback-Leibler Distance

- Is a measure of the distance between two distributions
- The relative entropy between two probability density functions $p_X(x)$ and $q_X(x)$ is defined as:

$$D(p_X(x)||q_X(x)) = \sum_{x \in \mathcal{X}} p_X(x) \log \frac{p_X(x)}{q_X(x)}$$

- $D(p_X(x)//\ q_X(x)) = 0$ with equality if and only if $p_X(x) = q_X(x)$
- $D(p_X(x)//\ q_X(x)) \neq D(q_X(x)//\ p_X(x))$
  - Higher D, more different between $p_X(x)$ and $q_X(x)$
  - $p_X(x)$: first distribution in domain X
  - $q_X(x)$: relative distribution with $p_X(x)$

3.5. Mutual information

- The mutual information of two random variables X and Y is defined as the relative entropy between the joint probability density $p_{XY}(x, y)$ and the product of the marginals $p_X(x)$ and $p_Y(y)$

$$
\begin{aligned}
I(X;Y) &= D(p_{XY}(x,y) \| p_X(x) p_Y(y)) \\
&= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_{XY}(x,y) \log \frac{p_{X,Y}(x,y)}{p_X(x) p_Y(y)}
\end{aligned}
$$

- $0 \leq I(X;Y) \leq H(X)$
- Mutual information with X as input, Y as output of a channel is the information that X transfers to Y or the information can be transmitted through the channel

3.5. Mutual information (cont.)

- Reducing uncertainty of X due to the knowledge of Y:

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

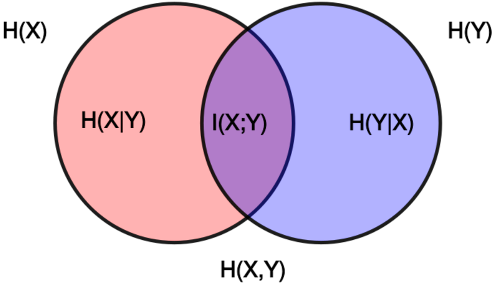
- Symmetry of the relation above: $I(X;Y) = H(Y) - H(Y|X)$
- Sum of entropies:

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

- "Self" Mutual Information:

$$I(X;X) = H(X) - H(X|X) = H(X)$$

3.5. Mutual information (cont.)



$$
\begin{aligned}
I(X;Y) &= H(X) - H(X \mid Y) \\
&= H(Y) - H(Y \mid X) \\
&= H(X) + H(Y) - H(X,Y)
\end{aligned}
$$

3.5. Mutual information (cont.)

Corollary:

- Conditional Mutual Information:

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

- Chain Rule for Mutual Information

$$I(X_1, X_2, \ldots, X_M; Y) = \sum_{j=1}^{M} I(X_j; Y | X_1, \ldots, X_{j-1})$$

Exercises:

1) X,Y = {xi,yj} i = 1..3; j = 1..3

$$P(X,Y) = \begin{pmatrix} 0.1 & 0.2 & 0.1 \\ 0.2 & 0.1 & 0.1 \\ 0.05 & 0.1 & 0.05 \end{pmatrix}$$

Calculate entropies and mutual information?

2) A source can generate only one message that have content: "information theory" is represented in form of a string without space between words, case insensitive. Each character in the message is an information. Probability of each information is calculated using ratio of the number of occurrences of information in the message divided by the total number of information in the message. Calculate entropy of the source and amount of information of the message ?

Solution

1) Joint entropy

$H(X,Y) = - P(x0,y0)\log P(x0,y0) - P(x0,y1)\log P(x0,y1) - P(x0,y2)\log P(x0,y2)$

$- P(x1,y0)\log P(x1,y0) - P(x1,y1)\log P(x1,y1) - P(x1,y2)\log P(x1,y2)$

$- P(x2,y0)\log P(x2,y0) - P(x2,y1)\log P(x2,y1) - P(x2,y2)\log P(x2,y2)$

$H(X) = -P(x0)\log P(x0) - P(x1)\log P(x1) - P(x2)\log P(x2)$

$P(x0) = P(x0,y0)+ P(x0,y1) + P(x0,y2)$

$P(x1) = P(x1,y0)+ P(x1,y1) + P(x1,y2)$

$P(x2) = 1 – P(x1) – P(x2)$

$H(Y) = -P(y0)\log P(y0) - P(y1)\log P(y1) - P(y2)\log P(y2)$

$P(y0) = P(x0,y0)+ P(x1,y0) + P(x2,y0)$

$P(y1) = P(x0,y1)+ P(x1,y1) + P(x2,y1)$

$P(y2) = 1 – P(y1) – P(y2)$

Solution (cont.)

H(X|Y) = - P(x0,y0)logP(x0|y0) - P(x0,y1)logP(x0|y1) - P(x0,y2)logP(x0|y2)

- P(x1,y0)logP(x1|y0) - P(x1,y1)logP(x1|y1) - P(x1,y2)logP(x1|y2)
- P(x2,y0)logP(x2|y0) - P(x2,y1)logP(x2|y1) - P(x2,y2)logP(x2|y2)

P(x0|y0) = $\frac{P(x0,y0)}{P(y0)}$ P(x0|y1) = $\frac{P(x0,y1)}{P(y1)}$ P(x0|y2) = $\frac{P(x0,y2)}{P(y2)}$

P(x1|y0) = $\frac{P(x1,y0)}{P(y0)}$ P(x1|y1) = $\frac{P(x1,y1)}{P(y1)}$ P(x1|y2) = $\frac{P(x1,y2)}{P(y2)}$

P(x2|y0) = $\frac{P(x2,y0)}{P(y0)}$ P(x2|y1) = $\frac{P(x2,y1)}{P(y1)}$ P(x2|y2) = $\frac{P(x2,y2)}{P(y2)}$

Solution (cont.)

$$
\begin{aligned}
H(Y|X) = & - P(x0,y0)\log P(y0|x0) - P(x0,y1)\log P(y1|x0) - P(x0,y2)\log P(y2|x0) \\
& - P(x1,y0)\log P(y0|x1) - P(x1,y1)\log P(y1|x1) - P(x1,y2)\log P(y2|x1) \\
& - P(x2,y0)\log P(y0|x2) - P(x2,y1)\log P(y1|x2) - P(x2,y2)\log P(y2|x2)
\end{aligned}
$$

$P(y0|x0) = \frac{P(x0,y0)}{P(x0)}$ $\quad$ $P(y1|x0) = \frac{P(x0,y1)}{P(x0)}$ $\quad$ $P(y2|x0) = \frac{P(x0,y2)}{P(x0)}$

$P(y0|x1) = \frac{P(x1,y0)}{P(x1)}$ $\quad$ $P(y1|x1) = \frac{P(x1,y1)}{P(x1)}$ $\quad$ $P(y2|x1) = \frac{P(x1,y2)}{P(x1)}$

$P(y0|x2) = \frac{P(x2,y0)}{P(x2)}$ $\quad$ $P(y1|x2) = \frac{P(x2,y1)}{P(x2)}$ $\quad$ $P(y2|x2) = \frac{P(x2,y2)}{P(x2)}$

Solutions (Cont)

$$
\begin{aligned}
I(X;Y) &= H(X) - H(X \mid Y) \\
&= H(Y) - H(Y \mid X) \\
&= H(X) + H(Y) - H(X,Y)
\end{aligned}
$$

I (X;Y) = $\sum_i \sum_j P(x_i, y_j) log \frac{P(x_i|y_j)}{P(x_i)}$

= $\sum_i \sum_j P\left(x_i, y_j\right) \log P(x_i|y_j)$

\- $\sum_i \sum_j P(x_i, y_j) log P(x_i)$

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

Solution (cont.)

2) Message “informationtheory”

X = {i,n,f,o,r,m,a,t,h,e,r,y }

P(X) = {2/17, 2/17, 1/17,3/17,2/17,1/17,1/17,2/17,1/17,1/17,1/17}

H(X)= -3 x (2/17) x log_2 (2/17) – 6 x (1/17) log_2 (1/17) - (3/17) log_2 (3/17) (bit/information)

Amount of information of message is estimated by

I(message) = 17 x H(X) (bits)

- 17 is number of the information in message
- H(X) is average amount of information contained in an information