Chapter 3: Information measure

3.1. Amount of information

3.2. Entropy

3.1. Amount of information

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

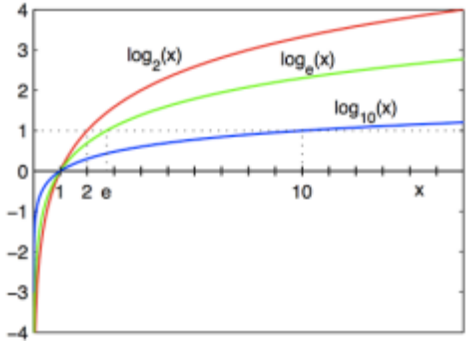
    → a source corresponds with a random variable
  - Information can be thought of as the resolution of uncertainty. It is abstract concept that describes understanding of objects in social life, in nature.

    → Amount of information equals uncertainty

    → Calculate amount of information through uncertainty

3.1. Amount of information (Cont.)

- It is a measure that quantifies the *uncertainty* of an event with given probability - Shannon 1948.
- Measure theory:
  - Uncertainty is inversely proportional to appearance probability $f(\frac{1}{p(x)})$
  - To ensure the linearity, uncertainty is calculate by log function
    - Two independent information x1, x2 with probability p(x1), p(x2) then
      - Amount of information of x1 is $f(\frac{1}{p(x1)})$
      - Amount of information 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 or amount of information

- Amount of information of *x* is denoted *I(x)*

3.1. Amount of information (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. Amount of information (Cont.)

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

Event	**Probability**
one equals one	1
wrong guess on a 4-choice question	3/4
correct guess on true-false question	1/2
correct guess on a 4-choice question	1/4
seven on a pair of dice	6/36
win a Jackpot	$\approx -1/76$ million

3.1. Amount of information (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

3.1. 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 $I(X) = E\{I(x_k)\}$

3.2. Entropy

3.2.1. Definition

3.2.2. Entropy of binary source

3.2.3. Joint entropy

3.2.4. Conditional entropy

3.2.5. Relationship between entropies

3.2.6. Example

3.2.7. Relative Entropy: Kullback-Leibler Distance

3.2.1. 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

3.2.1. 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

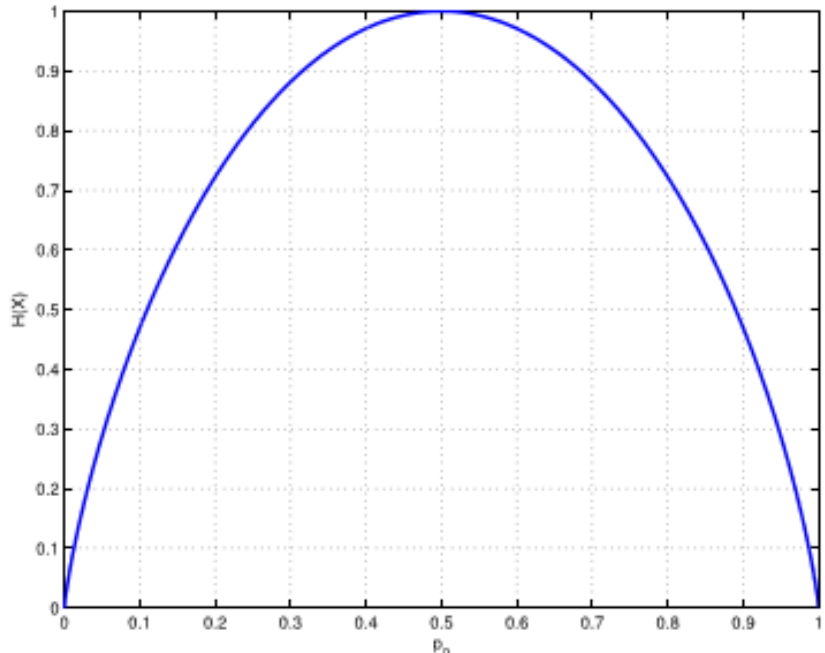
→With entropy, two sources can be compared:

- Source has higher entropy will transmit information with higher rate

3.2.2. 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

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



3.2.3. Joint entropy

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

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

3.2.4. Conditional entropy

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

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

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

3.2.5. Relationship between entropies

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

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

- Corollary:

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

3.2.5. 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

3.2.6. 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

3.2.6. 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

3.2.6. 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

3.2.7. 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 $p_X(x)$

3.3. 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

3.3. Mutual information (cont.)

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

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

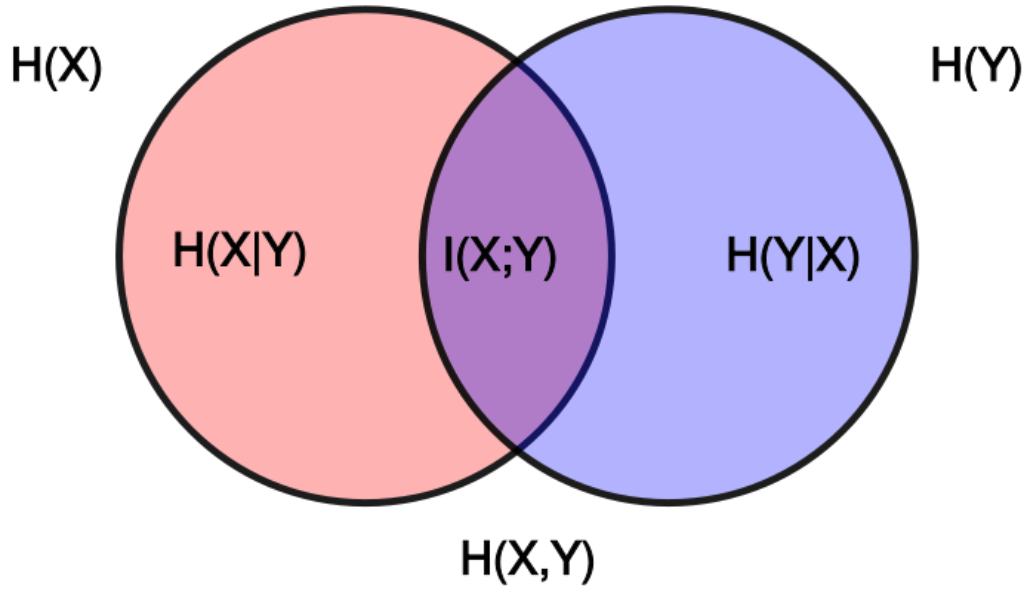
- Symmetry of the relation above: $I(X;Y) = H(Y) - H(Y|X)$
- Sum of entropies:

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

- "Self" Mutual Information:

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

3.3. Mutual information (cont.)



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

3.3. 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.)

$$H(X|Y) = -P(x0,y0)\log P(x0|y0) - P(x0,y1)\log P(x0|y1) - P(x0,y2)\log P(x0|y2)$$
$$- P(x1,y0)\log P(x1|y0) - P(x1,y1)\log P(x1|y1) - P(x1,y2)\log P(x1|y2)$$
$$- P(x2,y0)\log P(x2|y0) - P(x2,y1)\log P(x2|y1) - P(x2,y2)\log P(x2|y2)$$

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

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

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

Solution (cont.)

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

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

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

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

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

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

Solutions (Cont)

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

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

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