COMP2610 / COMP6261 Information Theory
Lecture 7: Relative Entropy and Mutual Information

Quanling Deng

Computational Mathematics Group
School of Computing
College of Engineering & Computer Science
The Australian National University
Canberra, Australia

Australian National University

Announcements

Assignment 1

- Available via Wattle
- Worth 10% of Course total
- Due Monday 29 August 2022, 5:00 pm
- Answers could be typed or handwritten

You can use latex LaTeX primer:
http://tug.ctan.org/info/lshort/english/lshort.pdf

Course Reps

Course Reps: Consider nomination yourself as a course Rep
by Thushara Abhayapala - Wednesday, 10 August 2022, 4:31 PM

Dear all,

We would like to appoint 2-4 class representatives for COMP2610/6210. It would be good to cover different student cohorts - eg., COMP 2610, COMP6261, On-Campus and Remote.

No previous experience is needed. You need to be able collect feedback from students and pass it to the Teaching team to improve the course. Timely feedback is always welcome so that we can make immediate adjustments.

If you are interested, could you please e-mail me asap.

Thushara

CECS Class Representatives

Class Student Representation is an important component of the teaching and learning quality assurance and quality improvement processes within the ANU College of Engineering and Computer Science (CECS).

The role of Class Representatives is to provide ongoing constructive feedback on behalf of the student cohort to Course Conveners and to Associate Directors (Education) for continuous improvements to the course.

Roles and responsibilities:

- Act as the official liaison between your peers and convener.
- Be creative, available and proactive in gathering feedback from your classmates.
- Attend regular meetings, and provide reports on course feedback to your course convener

1

Why become a class representative?

- **Ensure students have a voice** to their course convener, lecturer, tutors, and College.
- **Develop skills sought by employers**, including interpersonal, dispute resolution, leadership and communication skills.
- **Become empowered**. Play an active role in determining the direction of your education.
- **Become more aware of issues influencing your University** and current issues in higher education.
- **Course design and delivery.** Help shape the delivery of your current courses as well as future improvements for following years.

Note: Class representatives will need to be comfortable with their contact details being made available via Wattle to all students in the class.

Want to be a class representative? Nominate today!

Please nominate yourself to your course convener **asap**

You are free to nominate yourself whether you are currently on-campus or studying remotely.

For more information regarding roles and responsibilities, contact:

ANUSA CECS representatives: sa.cecs@anu.edu.au

Last time

- Information content and entropy: definition and computation
- Entropy and average code length
- Entropy and minimum expected number of binary questions
- Joint and conditional entropies, chain rule

Information Content: Review

Let X be a random variable with outcomes in $\mathcal{X}$

Let $p(x)$ denote the probability of the outcome $x \in \mathcal{X}$

The (Shannon) information content of outcome x is

$$h(x) = \log_2 \frac{1}{p(x)}$$

As $p(x) \to 0$, $h(x) \to +\infty$ (rare outcomes are more informative)

Entropy: Review

The entropy is the average information content of all outcomes:

$$H(X) = \sum_{x} p(x) \log_2 \frac{1}{p(x)}$$

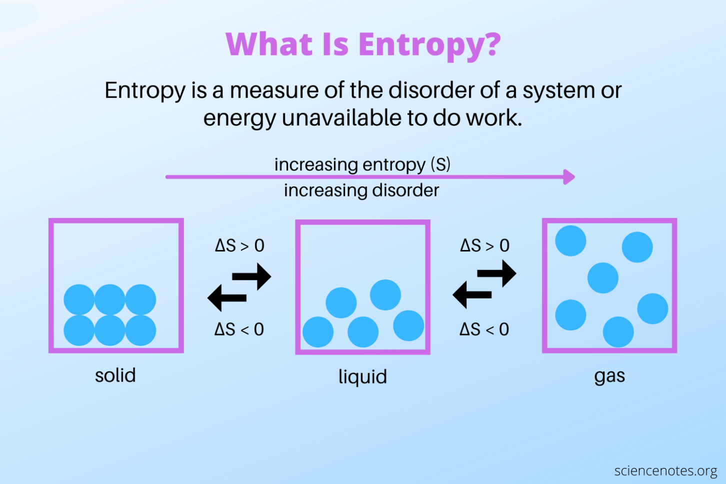
Entropy is minimised if **p** is peaked, and maximized if **p** is uniform:

$$0 \leq H(X) \leq \log |\mathcal{X}|$$

Entropy is related to minimal number of bits needed to describe a random variable

Entropy in another view

- A measurement of the degree of randomness of energy in a system
- Lower entropy means more ordered and less random; vice versa



This time

- The decomposability property of entropy

- Relative entropy and divergences

- Mutual information

Outline

Decomposability of Entropy

Example 1 (Mackay, 2003)

Let $X \in \{1, 2, 3\}$ be a r.v. created by the following process:

1. Flip a fair coin to determine whether $X = 1$

Decomposability of Entropy

Example 1 (Mackay, 2003)

Let $X \in \{1, 2, 3\}$ be a r.v. created by the following process:

1. Flip a fair coin to determine whether $X = 1$

2. If $X \neq 1$ flip another fair coin to determine whether $X = 2$ or $X = 3$

Decomposability of Entropy

Example 1 (Mackay, 2003)

Let $X \in \{1, 2, 3\}$ be a r.v. created by the following process:

1. Flip a fair coin to determine whether $X = 1$
2. If $X \neq 1$ flip another fair coin to determine whether $X = 2$ or $X = 3$

The probability distribution of X is given by:

$$p(X = 1) =$$

$$p(X = 2) =$$

$$p(X = 3) =$$

Decomposability of Entropy

Example 1 (Mackay, 2003)

Let $X \in \{1, 2, 3\}$ be a r.v. created by the following process:

1. Flip a fair coin to determine whether $X = 1$
2. If $X \neq 1$ flip another fair coin to determine whether $X = 2$ or $X = 3$

The probability distribution of X is given by:

$$p(X = 1) = \quad \frac{1}{2}$$

$$p(X = 2) =$$

$$p(X = 3) =$$

Decomposability of Entropy

Example 1 (Mackay, 2003)

Let $X \in \{1, 2, 3\}$ be a r.v. created by the following process:

1. Flip a fair coin to determine whether $X = 1$

2. If $X \neq 1$ flip another fair coin to determine whether $X = 2$ or $X = 3$

The probability distribution of X is given by:

$$\begin{aligned} p(X = 1) &= \frac{1}{2} \\ p(X = 2) &= \frac{1}{4} \\ p(X = 3) &= \end{aligned}$$

Decomposability of Entropy

Example 1 (Mackay, 2003)

Let $X \in \{1, 2, 3\}$ be a r.v. created by the following process:

1. Flip a fair coin to determine whether $X = 1$
2. If $X \neq 1$ flip another fair coin to determine whether $X = 2$ or $X = 3$

The probability distribution of X is given by:

$$
\begin{aligned}
p(X = 1) &= \frac{1}{2} \\
p(X = 2) &= \frac{1}{4} \\
p(X = 3) &= \frac{1}{4}
\end{aligned}
$$

Decomposability of Entropy

Example 1 (Mackay, 2003) — Cont'd

By definition, with $X \sim \mathbf{p}$, overloading H:

$$H(X) = \color{red}{H(\mathbf{p})} = \frac{1}{2}\log 2 + \frac{1}{4}\log 4 + \frac{1}{4}\log 4 = 1.5 \text{ bits.}$$

But imagine learning the value of X *gradually*:

Decomposability of Entropy

Example 1 (Mackay, 2003) — Cont'd

By definition, with $X \sim \mathbf{p}$, overloading H:

$$H(X) = H(\mathbf{p}) = \frac{1}{2}\log 2 + \frac{1}{4}\log 4 + \frac{1}{4}\log 4 = 1.5 \text{ bits.}$$

But imagine learning the value of X *gradually*:

1. First we learn whether $X = 1$:
   - Binary variable with $\mathbf{p}^{(1)} = (\frac{1}{2}, \frac{1}{2})$
   - Hence $H((1/2, 1/2)) = \log_2 2 = 1$ bit.

Decomposability of Entropy

Example 1 (Mackay, 2003) — Cont'd

By definition, with $X \sim \mathbf{p}$, overloading H:

$$H(X) = H(\mathbf{p}) = \frac{1}{2}\log 2 + \frac{1}{4}\log 4 + \frac{1}{4}\log 4 = 1.5 \text{ bits.}$$

But imagine learning the value of X *gradually*:

1. First we learn whether $X = 1$:
   - Binary variable with $\mathbf{p}^{(1)} = (\frac{1}{2}, \frac{1}{2})$
   - Hence $H((1/2, 1/2)) = \log_2 2 = 1$ bit.
2. If $X \neq 1$ we learn the value of the second coin flip:
   - Also binary variable with $\mathbf{p}^{(2)} = (\frac{1}{2}, \frac{1}{2})$
   - Therefore $H((1/2, 1/2)) = 1$ bit.

Decomposability of Entropy

Example 1 (Mackay, 2003) — Cont'd

By definition, with $X \sim \mathbf{p}$, overloading H:

$$H(X) = H(\mathbf{p}) = \frac{1}{2}\log 2 + \frac{1}{4}\log 4 + \frac{1}{4}\log 4 = 1.5 \text{ bits.}$$

But imagine learning the value of X *gradually*:

1. First we learn whether $X = 1$:
   - Binary variable with $\mathbf{p}^{(1)} = (\frac{1}{2}, \frac{1}{2})$
   - Hence $H((1/2, 1/2)) = \log_2 2 = 1$ bit.
2. If $X \neq 1$ we learn the value of the second coin flip:
   - Also binary variable with $\mathbf{p}^{(2)} = (\frac{1}{2}, \frac{1}{2})$
   - Therefore $H((1/2, 1/2)) = 1$ bit.

However, the second revelation only happens half of the time:

$$H(X) = H((1/2, 1/2)) + \frac{1}{2}H((1/2, 1/2)) = 1.5 \text{ bits.}$$

Decomposability of Entropy

Generalization

For a r.v. with probability distribution $\mathbf{p} = (p_1, \ldots, p_{|\mathcal{X}|})$:

$$H(\mathbf{p}) = H((p_1, 1 - p_1)) + (1 - p_1) H\left(\left(\frac{p_2}{1 - p_1}, \ldots, \frac{p_{|\mathcal{X}|}}{1 - p_1}\right)\right)$$

Decomposability of Entropy

Generalization

For a r.v. with probability distribution $\mathbf{p} = (p_1, \ldots, p_{|\mathcal{X}|})$:

$$H(\mathbf{p}) = H((p_1, 1 - p_1)) + (1 - p_1) H\left(\left(\frac{p_2}{1 - p_1}, \ldots, \frac{p_{|\mathcal{X}|}}{1 - p_1}\right)\right)$$

$H((p_1, 1 - p_1))$: entropy for a random variable corresponding to "Is $X = 1$?"

Decomposability of Entropy

Generalization

For a r.v. with probability distribution $\mathbf{p} = (p_1, \ldots, p_{|\mathcal{X}|})$:

$$H(\mathbf{p}) = H((p_1, 1 - p_1)) + (1 - p_1)H\left(\left(\frac{p_2}{1 - p_1}, \ldots, \frac{p_{|\mathcal{X}|}}{1 - p_1}\right)\right)$$

$H((p_1, 1 - p_1))$: entropy for a random variable corresponding to “Is $X = 1$?”

$1 - p_1$: probability of $X \neq 1$

Decomposability of Entropy

Generalization

For a r.v. with probability distribution $\mathbf{p} = (p_1, \ldots, p_{|\mathcal{X}|})$:

$$H(\mathbf{p}) = H((p_1, 1 - p_1)) + (1 - p_1) H\left(\left(\frac{p_2}{1 - p_1}, \ldots, \frac{p_{|\mathcal{X}|}}{1 - p_1}\right)\right)$$

$H((p_1, 1 - p_1))$: entropy for a random variable corresponding to "Is $X = 1$?"

$1 - p_1$: probability of $X \neq 1$

$\frac{p_2}{1-p_1}, \ldots, \frac{p_{|\mathcal{X}|}}{1-p_1}$: conditional probability of $X = 2, \ldots, |\mathcal{X}|$ given $X \neq 1$.

$H\left(\left(\frac{p_2}{1-p_1}, \ldots, \frac{p_{|\mathcal{X}|}}{1-p_1}\right)\right)$: entropy for a random variable corresponding to outcomes when $X \neq 1$.

Decomposability of Entropy

Generalization

Henceforth write $H((p_1, \ldots, p_N))$ as $H(p_1, \ldots, p_N)$.
Do not confuse with joint entropy $H(X_1, \ldots, X_n)$.
In general, we have that for any m between 1 and $|\mathcal{X}| - 1$:

$$\begin{aligned}H(\mathbf{p}) =& H\left(\sum_{i=1}^{m} p_i, \sum_{i=m+1}^{|\mathcal{X}|} p_i\right) \\ &+ \left(\sum_{i=1}^{m} p_i\right) H\left(\frac{p_1}{\sum_{i=1}^{m} p_i}, \ldots, \frac{p_m}{\sum_{i=1}^{m} p_i}\right) \\ &+ \left(\sum_{i=m+1}^{|\mathcal{X}|} p_i\right) H\left(\frac{p_{m+1}}{\sum_{i=m+1}^{|\mathcal{X}|} p_i}, \ldots, \frac{p_{|\mathcal{X}|}}{\sum_{i=m+1}^{|\mathcal{X}|} p_i}\right)\end{aligned}$$

Apply this formula with $m = 1$, $|\mathcal{X}| = 3$, $\mathbf{p} = (p_1, p_2, p_3) = (1/2, 1/4, 1/4)$

Entropy in Information Theory

If a random variable has distribution p, there exists an encoding with an average length of

$$H(p) \text{ bits}$$

and this is the "best" possible encoding

What happens if we use a "wrong" encoding?

- e.g. because we make an incorrect assumption on the probability distribution

If the true distribution is p, but we assume it is q, it turns out we will need to use

$$H(p) + D_{\mathrm{KL}}(p||q) \text{ bits}$$

where $D_{\mathrm{KL}}(p||q)$ is some measure of "distance" between p and q

Relative Entropy

Definition

The relative entropy or Kullback-Leibler (KL) divergence between two probability distributions $p(X)$ and $q(X)$ is defined as:

$$D_{\mathrm{KL}}(p \| q) = \sum_{x \in \mathcal{X}} p(x) \left(\log \frac{1}{q(x)} - \log \frac{1}{p(x)} \right)$$
$$= \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} = \mathbb{E}_p \left[\log \frac{p(X)}{q(X)} \right].$$

Relative Entropy

Definition

The relative entropy or Kullback-Leibler (KL) divergence between two probability distributions $p(X)$ and $q(X)$ is defined as:

$$\begin{aligned} D_{\mathrm{KL}}(p \| q) &= \sum_{x \in \mathcal{X}} p(x) \left(\log \frac{1}{q(x)} - \log \frac{1}{p(x)} \right) \\ &= \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} = \mathbb{E}_p \left[\log \frac{p(X)}{q(X)} \right]. \end{aligned}$$

- Note:
  - Both $p(X)$ and $q(X)$ are defined over the same alphabet $\mathcal{X}$
- Conventions on log likelihood ratio:

$$0 \log \frac{0}{0} \stackrel{\text{def}}{=} 0 \qquad 0 \log \frac{0}{q} \stackrel{\text{def}}{=} 0 \qquad p \log \frac{p}{0} \stackrel{\text{def}}{=} \infty$$

Relative Entropy

Properties

- $D_{\mathrm{KL}}(p\|q) \geq 0$ (proof next lecture)

Relative Entropy

Properties

- $D_{\text{KL}}(p \| q) \geq 0$ (proof next lecture)
- $D_{\text{KL}}(p \| q) = 0 \Leftrightarrow p = q$ (proof next lecture)

Relative Entropy

Properties

- $D_{\mathrm{KL}}(p\|q) \geq 0$ (proof next lecture)
- $D_{\mathrm{KL}}(p\|q) = 0 \Leftrightarrow p = q$ (proof next lecture)
- Not symmetric: $D_{\mathrm{KL}}(p\|q) \neq D_{\mathrm{KL}}(q\|p)$

Relative Entropy

Properties

- $D_{\mathrm{KL}}(p \| q) \geq 0$ (proof next lecture)
- $D_{\mathrm{KL}}(p \| q) = 0 \Leftrightarrow p = q$ (proof next lecture)
- Not symmetric: $D_{\mathrm{KL}}(p \| q) \neq D_{\mathrm{KL}}(q \| p)$
- Not satisfy triangle inequality: $D_{\mathrm{KL}}(p \| q) \neq D_{\mathrm{KL}}(p \| r) + D_{\mathrm{KL}}(r \| q)$

Relative Entropy

Properties

- $D_{\mathrm{KL}}(p\|q) \geq 0$ (proof next lecture)
- $D_{\mathrm{KL}}(p\|q) = 0 \Leftrightarrow p = q$ (proof next lecture)
- Not symmetric: $D_{\mathrm{KL}}(p\|q) \neq D_{\mathrm{KL}}(q\|p)$
- Not satisfy triangle inequality: $D_{\mathrm{KL}}(p\|q) \neq D_{\mathrm{KL}}(p\|r) + D_{\mathrm{KL}}(r\|q)$
  - Not a true distance since is not symmetric and does not satisfy the triangle inequality

Relative Entropy

Properties

- $D_{\mathrm{KL}}(p \| q) \geq 0$ (proof next lecture)
- $D_{\mathrm{KL}}(p \| q) = 0 \Leftrightarrow p = q$ (proof next lecture)
- Not symmetric: $D_{\mathrm{KL}}(p \| q) \neq D_{\mathrm{KL}}(q \| p)$
- Not satisfy triangle inequality: $D_{\mathrm{KL}}(p \| q) \neq D_{\mathrm{KL}}(p \| r) + D_{\mathrm{KL}}(r \| q)$
  - Not a true distance since is not symmetric and does not satisfy the triangle inequality
  - Hence, "KL divergence" rather than "KL distance"

Relative Entropy

Properties

- $D_{\text{KL}}(p\|q) \geq 0$ (proof next lecture)
- $D_{\text{KL}}(p\|q) = 0 \Leftrightarrow p = q$ (proof next lecture)
- Not symmetric: $D_{\text{KL}}(p\|q) \neq D_{\text{KL}}(q\|p)$
- Not satisfy triangle inequality: $D_{\text{KL}}(p\|q) \neq D_{\text{KL}}(p\|r) + D_{\text{KL}}(r\|q)$
  - Not a true distance since is not symmetric and does not satisfy the triangle inequality
  - Hence, “KL divergence” rather than “KL distance”
  - Funny notation $D_{\text{KL}}(p\|q)$ is to remind us it is not symmetric.

Relative Entropy

Uniform q

Let q correspond to a uniform distribution: $q(x) = \frac{1}{|\mathcal{X}|}$

Relative entropy between p and q:

$$\begin{aligned}
D_{\mathrm{KL}}(p \| q) &= \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \\
&= \sum_{x \in \mathcal{X}} p(x) \cdot (\log p(x) + \log |\mathcal{X}|) \\
&= -H(X) + \sum_{x \in \mathcal{X}} p(x) \cdot \log |\mathcal{X}| \\
&= -H(X) + \log |\mathcal{X}|.
\end{aligned}$$

Matches intuition as penalty on number of bits for encoding

Relative Entropy

Example (from Cover & Thomas, 2006)

Let $X \in \{0, 1\}$ and consider the distributions $p(X)$ and $q(X)$ such that:

$$p(X = 1) = \theta_p \qquad p(X = 0) = 1 - \theta_p$$
$$q(X = 1) = \theta_q \qquad q(X = 0) = 1 - \theta_q$$

What distributions are these?

Relative Entropy

Example (from Cover & Thomas, 2006)

Let $X \in \{0, 1\}$ and consider the distributions $p(X)$ and $q(X)$ such that:

$$p(X = 1) = \theta_p \qquad p(X = 0) = 1 - \theta_p$$
$$q(X = 1) = \theta_q \qquad q(X = 0) = 1 - \theta_q$$

What distributions are these?

Compute $D_{\mathrm{KL}}(p \| q)$ and $D_{\mathrm{KL}}(q \| p)$ with $\theta_p = \frac{1}{2}$ and $\theta_q = \frac{1}{4}$

Relative Entropy

Example (from Cover & Thomas, 2006) — Cont'd

$$D_{\mathrm{KL}}(p\|q) = \theta_p \log \frac{\theta_p}{\theta_q} + (1-\theta_p) \log \frac{1-\theta_p}{1-\theta_q}$$

Relative Entropy

Example (from Cover & Thomas, 2006) — Cont'd

$$
\begin{aligned}
D_{\mathrm{KL}}(p\|q) &= \theta_p \log \frac{\theta_p}{\theta_q} + (1-\theta_p)\log\frac{1-\theta_p}{1-\theta_q} \\
&= \frac{1}{2}\log\frac{\frac{1}{2}}{\frac{1}{4}} + \frac{1}{2}\log\frac{\frac{1}{2}}{\frac{3}{4}} = 1 - \frac{1}{2}\log 3 \approx 0.2075 \text{ bits}
\end{aligned}
$$

Relative Entropy

Example (from Cover & Thomas, 2006) — Cont'd

$$
\begin{aligned}
D_{\mathrm{KL}}(p\|q) &= \theta_p \log \frac{\theta_p}{\theta_q} + (1-\theta_p) \log \frac{1-\theta_p}{1-\theta_q} \\
&= \frac{1}{2} \log \frac{\frac{1}{2}}{\frac{1}{4}} + \frac{1}{2} \log \frac{\frac{1}{2}}{\frac{3}{4}} = 1 - \frac{1}{2} \log 3 \approx 0.2075 \text{ bits}
\end{aligned}
$$

$$
D_{\mathrm{KL}}(q\|p) = \theta_q \log \frac{\theta_q}{\theta_p} + (1-\theta_q) \log \frac{1-\theta_q}{1-\theta_p}
$$

Relative Entropy

Example (from Cover & Thomas, 2006) — Cont'd

$$D_{\mathrm{KL}}(p\|q) = \theta_p \log \frac{\theta_p}{\theta_q} + (1-\theta_p) \log \frac{1-\theta_p}{1-\theta_q}$$
$$= \frac{1}{2} \log \frac{\frac{1}{2}}{\frac{1}{4}} + \frac{1}{2} \log \frac{\frac{1}{2}}{\frac{3}{4}} = 1 - \frac{1}{2} \log 3 \approx 0.2075 \text{ bits}$$

$$D_{\mathrm{KL}}(q\|p) = \theta_q \log \frac{\theta_q}{\theta_p} + (1-\theta_q) \log \frac{1-\theta_q}{1-\theta_p}$$
$$= \frac{1}{4} \log \frac{\frac{1}{4}}{\frac{1}{2}} + \frac{3}{4} \log \frac{\frac{3}{4}}{\frac{1}{2}} = -1 + \frac{3}{4} \log 3 \approx 0.1887 \text{ bits}$$

Mutual Information

Definition

Let X, Y be two r.v. with joint distribution $p(X, Y)$ and marginals $p(X)$ and $p(Y)$:

Definition

The *mutual information* $I(X; Y)$ is the relative entropy between the joint distribution $p(X, Y)$ and the product distribution $p(X)p(Y)$:

$$\begin{aligned} I(X; Y) &= D_{\mathrm{KL}}\left(p(X, Y) \| p(X)p(Y)\right) \\ &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \end{aligned}$$

Mutual Information

Definition

Let X, Y be two r.v. with joint distribution $p(X, Y)$ and marginals $p(X)$ and $p(Y)$:

Definition

The *mutual information* $I(X;Y)$ is the relative entropy between the joint distribution $p(X, Y)$ and the product distribution $p(X)p(Y)$:

$$
\begin{aligned}
I(X;Y) &= D_{\mathrm{KL}}\left(p(X,Y) \| p(X)p(Y)\right) \\
&= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}
\end{aligned}
$$

Non-negativity: $I(X;Y) \geq 0$
Symmetry: $I(Y;X) = I(X;Y)$

Intuitively, how much information, on average, X conveys about Y.

Relationship between Entropy and Mutual Information

We can re-write the definition of mutual information as:

$$I(X;Y) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

Relationship between Entropy and Mutual Information

We can re-write the definition of mutual information as:

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

Relationship between Entropy and Mutual Information

We can re-write the definition of mutual information as:

$$
\begin{aligned}
I(X;Y) &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\
&= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x|y)}{p(x)} \\
&= -\sum_{x \in \mathcal{X}} \log p(x) \sum_{y \in \mathcal{Y}} p(x,y) - \left(-\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log p(x|y) \right)
\end{aligned}
$$

Relationship between Entropy and Mutual Information

We can re-write the definition of mutual information as:

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

Relationship between Entropy and Mutual Information

We can re-write the definition of mutual information as:

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

The average reduction in uncertainty of X due to the knowledge of Y.

Self-information: $I(X;X) = H(X) - H(X|X) = H(X)$

Mutual Information:

Properties

- Mutual Information is non-negative:

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

Mutual Information:

Properties

- Mutual Information is non-negative:

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

- Mutual Information is symmetric:

$$I(X; Y) = I(Y; X)$$

Mutual Information:

Properties

- Mutual Information is non-negative:

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

- Mutual Information is symmetric:

$$I(X;Y) = I(Y;X)$$

- Self-information:

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

Mutual Information:

Properties

- Mutual Information is non-negative:

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

- Mutual Information is symmetric:

$$I(X;Y) = I(Y;X)$$

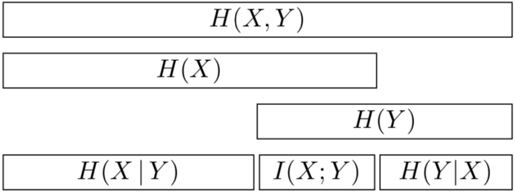
- Self-information:

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

- Since $H(X,Y) = H(Y) + H(X|Y)$ we have that:

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

Breakdown of Joint Entropy



(From Mackay, p140; see his exercise 8.8)

Mutual Information

Example 1 (from Mackay, 2003)

Let X, Y, Z be r.v. with $X, Y \in \{0, 1\}$, $X \perp\!\!\!\perp Y$ and:

$$p(X = 0) = p \qquad p(X = 1) = 1 - p$$
$$p(Y = 0) = q \qquad p(Y = 1) = 1 - q$$
$$Z = (X + Y) \mod 2$$

Mutual Information

Example 1 (from Mackay, 2003)

Let X, Y, Z be r.v. with $X, Y \in \{0, 1\}$, $X \perp\!\!\!\perp Y$ and:

$$p(X = 0) = p \qquad p(X = 1) = 1 - p$$
$$p(Y = 0) = q \qquad p(Y = 1) = 1 - q$$
$$Z = (X + Y) \mod 2$$

(a) if $q = 1/2$ what is $P(Z = 0)$? $P(Z = 1)$? $I(Z; X)$?

Mutual Information

Example 1 (from Mackay, 2003)

Let X, Y, Z be r.v. with $X, Y \in \{0, 1\}$, $X \perp\!\!\!\perp Y$ and:

$$\begin{aligned} p(X = 0) &= p & \quad p(X = 1) &= 1 - p \\ p(Y = 0) &= q & \quad p(Y = 1) &= 1 - q \end{aligned}$$

$$Z = (X + Y) \mod 2$$

(a) if $q = 1/2$ what is $P(Z = 0)$? $P(Z = 1)$? $I(Z; X)$?

(b) For general p and q what is $P(Z = 0)$? $P(Z = 1)$? $I(Z; X)$?

Mutual Information

Example 1 (from Mackay, 2003) — Solution (a)

As $X \perp\!\!\!\perp Y$ and $q = 1/2$ the noise will flip the outcome of X with probability $q = 0.5$ regardless of the outcome of X. Therefore:

$$p(Z = 1) = 1/2 \qquad p(Z = 0) = 1/2$$

Mutual Information

Example 1 (from Mackay, 2003) — Solution (a)

As $X \perp\!\!\!\perp Y$ and $q = 1/2$ the noise will flip the outcome of X with probability $q = 0.5$ regardless of the outcome of X. Therefore:

$$p(Z = 1) = 1/2 \qquad p(Z = 0) = 1/2$$

We have

$$\begin{aligned} H(Z|X) &= -\sum_x p(x) \sum_z p(z|x) \log p(z|X) \\ &= -(1/2) \log(1/2) \sum_x p(x) \\ &= 1 \text{ bit} \end{aligned}$$

Mutual Information

Example 1 (from Mackay, 2003) — Solution (a)

As $X \perp\!\!\!\perp Y$ and $q = 1/2$ the noise will flip the outcome of X with probability $q = 0.5$ regardless of the outcome of X. Therefore:

$$p(Z = 1) = 1/2 \qquad p(Z = 0) = 1/2$$

We have

$$\begin{aligned} H(Z|X) &= -\sum_x p(x) \sum_z p(z|x) \log p(z|X) \\ &= -(1/2)\log(1/2) \sum_x p(x) \\ &= 1 \text{ bit} \end{aligned}$$

Hence:

$$I(X; Z) = H(Z) - H(Z|X) = 1 - 1 = 0$$

Mutual Information

Example 1 (from Mackay, 2003) — Solution (a)

As $X \perp\!\!\!\perp Y$ and $q = 1/2$ the noise will flip the outcome of X with probability $q = 0.5$ regardless of the outcome of X. Therefore:

$$p(Z = 1) = 1/2 \qquad p(Z = 0) = 1/2$$

We have

$$\begin{aligned} H(Z|X) &= -\sum_x p(x) \sum_z p(z|x) \log p(z|X) \\ &= -(1/2)\log(1/2) \sum_x p(x) \\ &= 1 \text{ bit} \end{aligned}$$

Hence:

$$I(X;Z) = H(Z) - H(Z|X) = 1 - 1 = 0$$

Thus for $q = 1/2$, $Z \perp\!\!\!\perp X$.

Mutual Information

Example 1 (from Mackay, 2003) — Solution (a)

As $X \perp\!\!\!\perp Y$ and $q = 1/2$ the noise will flip the outcome of X with probability $q = 0.5$ regardless of the outcome of X. Therefore:

$$p(Z = 1) = 1/2 \qquad p(Z = 0) = 1/2$$

We have

$$\begin{aligned} H(Z|X) &= -\sum_x p(x) \sum_z p(z|x) \log p(z|X) \\ &= -(1/2)\log(1/2) \sum_x p(x) \\ &= 1 \text{ bit} \end{aligned}$$

Hence:

$$I(X; Z) = H(Z) - H(Z|X) = 1 - 1 = 0$$

Thus for $q = 1/2$, $Z \perp\!\!\!\perp X$.

What significance might this have for spies?

Mutual Information

Example 1 (from Mackay, 2003) — Solution (b)

$$\begin{aligned}
\ell \stackrel{\text{def}}{=} p(Z=0) &= p(X=0) \times p(\text{no flip}) + p(X=1) \times p(\text{flip}) \\
&= pq + (1-p)(1-q) \\
&= 1 + 2pq - q - p
\end{aligned}$$

Mutual Information

Example 1 (from Mackay, 2003) — Solution (b)

$$
\begin{aligned}
\ell \stackrel{\text{def}}{=} p(Z=0) &= p(X=0) \times p(\text{no flip}) + p(X=1) \times p(\text{flip}) \\
&= pq + (1-p)(1-q) \\
&= 1 + 2pq - q - p
\end{aligned}
$$

Similarly:

$$
\begin{aligned}
p(Z=1) &= p(X=1) \times p(\text{no flip}) + p(X=0) \times p(\text{flip}) \\
&= (1-p)q + p(1-q) \\
&= q + p - 2pq
\end{aligned}
$$

Mutual Information

Example 1 (from Mackay, 2003) — Solution (b)

$$\begin{aligned}\ell \stackrel{\text{def}}{=} p(Z=0) &= p(X=0) \times p(\text{no flip}) + p(X=1) \times p(\text{flip}) \\ &= pq + (1-p)(1-q) \\ &= 1 + 2pq - q - p\end{aligned}$$

Similarly:

$$\begin{aligned}p(Z=1) &= p(X=1) \times p(\text{no flip}) + p(X=0) \times p(\text{flip}) \\ &= (1-p)q + p(1-q) \\ &= q + p - 2pq\end{aligned}$$

Since $p(Z|X=0) = (q, 1-q)$ and $p(Z|X=1) = (1-q, q)$ we have $H(Z|X=0) = H(Z|X=1) = H(q, 1-q)$.
Averaging over $p(X)$ we have
$H(Z|X) = p(H(q, 1-q)) + (1-p)(H(q, 1-q)) = H(q, 1-q)$.

Mutual Information

Example 1 (from Mackay, 2003) — Solution (b)

$$\begin{aligned}
\ell \stackrel{\text{def}}{=} p(Z=0) &= p(X=0) \times p(\text{no flip}) + p(X=1) \times p(\text{flip}) \\
&= pq + (1-p)(1-q) \\
&= 1 + 2pq - q - p
\end{aligned}$$

Similarly:

$$\begin{aligned}
p(Z=1) &= p(X=1) \times p(\text{no flip}) + p(X=0) \times p(\text{flip}) \\
&= (1-p)q + p(1-q) \\
&= q + p - 2pq
\end{aligned}$$

Since $p(Z|X=0) = (q, 1-q)$ and $p(Z|X=1) = (1-q, q)$ we have $H(Z|X=0) = H(Z|X=1) = H(q, 1-q)$.
Averaging over $p(X)$ we have
$H(Z|X) = p(H(q, 1-q)) + (1-p)(H(q, 1-q)) = H(q, 1-q)$.
Thus:

$$\begin{aligned}
I(Z;X) &= H(Z) - H(Z|X) \\
&= H(\ell, 1-\ell) - H(q, 1-q)
\end{aligned}$$

Joint Mutual Information

Recall that for random variables X, Y,

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

- Reduction in uncertainty in X due to knowledge of Y

More generally, for random variables $X_1, \ldots, X_n, Y_1, \ldots, Y_m$,

$$I(X_1, \ldots, X_n; Y_1, \ldots, Y_m) = H(X_1, \ldots, X_n) - H(X_1, \ldots, X_n | Y_1, \ldots, Y_m)$$

- Reduction in uncertainty in $X_1, \ldots, X_n$ due to knowledge of $Y_1, \ldots, Y_m$

Symmetry also generalises:

$$I(X_1, \ldots, X_n; Y_1, \ldots, Y_m) = I(Y_1, \ldots, Y_m; X_1, \ldots, X_n)$$

Conditional Mutual Information

The conditional mutual information between X and Y given $Z = z_k$:

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

Conditional Mutual Information

The conditional mutual information between X and Y given $Z = z_k$:

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

Averaging over Z we obtain:

The conditional mutual information between X and Y given Z:

$$\begin{aligned} I(X; Y|Z) &= H(X|Z) - H(X|Y, Z) \\ &= \mathbb{E}_{p(X,Y,Z)} \log \frac{p(X, Y|Z)}{p(X|Z)p(Y|Z)} \end{aligned}$$

Conditional Mutual Information

The conditional mutual information between X and Y given $Z = z_k$:

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

Averaging over Z we obtain:

The conditional mutual information between X and Y given Z:

$$\begin{aligned} I(X; Y|Z) &= H(X|Z) - H(X|Y, Z) \\ &= \mathbb{E}_{p(X,Y,Z)} \log \frac{p(X, Y|Z)}{p(X|Z)p(Y|Z)} \end{aligned}$$

The reduction in the uncertainty of X due to the knowledge of Y when Z is given.

Conditional Mutual Information

The conditional mutual information between X and Y given $Z = z_k$:

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

Averaging over Z we obtain:

The conditional mutual information between X and Y given Z:

$$\begin{aligned} I(X;Y|Z) &= H(X|Z) - H(X|Y,Z) \\ &= \mathbb{E}_{p(X,Y,Z)} \log \frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)} \end{aligned}$$

The reduction in the uncertainty of X due to the knowledge of Y when Z is given.

Note that $I(X;Y;Z)$, $I(X|Y;Z)$ are illegal terms while e.g. $I(A,B;C,D|E,F)$ is legal.

Summary

- Decomposability of entropy
- Relative entropy
- Mutual information
- Reading: Mackay §2.5, Ch 8; Cover & Thomas §2.3 to §2.5
- Important: You should be doing lots of exercises from the text!
- Feedback: Please provide feedback — see Wattle page

Next time

Mutual information chain rule

Jensen's inequality

"Information cannot hurt"

Data processing inequality

Acknowledgement

These slides were originally developed by Professor Robert C. Williamson.