

COMP2610 / COMP6261 - Information Theory

Lecture 7: Relative Entropy and Mutual Information

Assignment Project Exam Help

<https://eduassistpro.github.io>



Australian  
National  
University

Add WeChat edu\_assist\_pro

13 August 2018

# Assignment Project Exam Help

- Information content and entropy: definition and computation

- Ent

<https://eduassistpro.github.io>

- Entropy and minimum expected number of bi

Add WeChat edu\_assist\_pr

- Joint and conditional entropies, chain rule

# Assignment Project Exam Help

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

Let  $p(x)$

The (Shannon)

<https://eduassistpro.github.io>

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

Add WeChat [edu\\_assist\\_pro](#)

As  $p(x) \rightarrow 0$ ,  $h(x) \rightarrow +\infty$  (rare outcomes are

## Entropy: Review

The entropy is the average information content of all outcomes:

1

Entropy is

<https://eduassistpro.github.io>

in.

$$0 \leq H(X) \leq \log$$

Entropy is related to minimal number of bits needed to  
variable

Add WeChat edu\_assist\_pro

This time

# Assignment Project Exam Help

- The decomposability property of entropy

- Rel

<https://eduassistpro.github.io>

Add WeChat edu\_assist\_pr

- Mutual information

## Outline

1 Decomposability of Entropy

2 Relat

3 Mutu

- Definition
- Joint and Conditional Mutual Information

4 Wrapping up

Assignment Project Exam Help

<https://eduassistpro.github.io>

Add WeChat edu\_assist\_pro

# 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  $X \neq 1$

The prob

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

$$p(X = 2) = \frac{1}{4}$$

$$p(X = 3) = \frac{1}{4}$$

# Decomposability of Entropy

Example 1 (Mackay, 2003) — Cont'd

By definition,

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

But imagine

1 First

▶ <https://eduassistpro.github.io>

▶

2

2 If  $X \neq 1$  we learn the value of the second coin flip

▶ Also binary variable with  $p^{(2)} = (1/2, 1/2)$

▶ 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  $\mathbf{p} = (p_1, \dots, p_{|\mathcal{X}|})$ :

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

$$H(p_1, 1 - p_1)$$

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

$$\frac{p_2}{1 - p_1}, \dots, \frac{p_{|\mathcal{X}|}}{1 - p_1}$$

conditional probability of  $X \neq 1$ .

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

# Decomposability of Entropy

## Generalization

In general, we have that for any  $m$  between 1 and  $|\mathcal{X}| - 1$ :

<https://eduassistpro.github.io>

Add WeChat [edu\\_assist\\_pro](#)

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

1 Decomposability of Entropy

# Assignment Project Exam Help

2 Relative Entropy / KL Divergence

3 Mutual Information

- De
- Joint and Conditional Mutual Information

<https://eduassistpro.github.io>

Add WeChat edu\_assist\_pr

4 Wrapping up

# 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 t

What happens

- e.g. b

n

If the true distribution is  $p$ , but we assume it is  $q$  to use

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

where  $D_{\text{KL}}(p||q)$  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:

$$D_{\text{KL}}(p \parallel q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

- Note:

- ▶ Both  $p(X)$  and  $q(X)$  are defined over  $\mathcal{X}$

- Conventions on log likelihood ratio:

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

# Relative Entropy

## Properties

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

- $D_{KL}$

- Not s

- Not satisfy triangle inequality:  $D_{KL}(p \parallel q)$

- ▶ Not a true distance since is not symmetric and triangle inequality

- ▶ Hence, “KL divergence” rather than “KL distance”

# Relative Entropy

Uniform  $q$

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

Assignment Project Exam Help

Relative entropy between  $p$  and  $q$ :

<https://eduassistpro.github.io>

Add WeChat edu\_assist\_pro

$$\begin{aligned} & \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \\ &= -H(X) + \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)} \\ &= -H(X) + \log |\mathcal{X}|. \end{aligned}$$

Matches intuition as penalty on number of bits for encoding

# Relative Entropy

Example (from Cover & Thomas, 2006)

## Assignment Project Exam Help

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

<https://eduassistpro.github.io>

Add WeChat [edu\\_assist\\_pro](#)

Compute  $D_{\text{KL}}(p||q)$  and  $D_{\text{KL}}(q||p)$



# Relative Entropy

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

## Assignment Project Exam Help

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

<https://eduassistpro.github.io>

Add WeChat: edu\_assist\_pro

$$= \frac{1}{4} \log \frac{\frac{1}{4}}{\frac{1}{2}} + \frac{3}{4} \log \frac{\frac{3}{4}}{\frac{1}{2}} = -1 + \frac{3}{4} \log 3 \approx 0.1887 \text{ bits}$$

1 Decomposability of Entropy

# Assignment Project Exam Help

2 Relative Entropy / KL Divergence

3 Mutu

- De

- Joint and Conditional Mutual Information

<https://eduassistpro.github.io>

Add WeChat edu\_assist\_pr

4 Wrapping up

# 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 and the product of the marginals:

<https://eduassistpro.github.io>

Add WeChat edu\_assist\_pro

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

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:

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

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:

Assignment Project Exam Help

- Mut

<https://eduassistpro.github.io>

- Self-information:

Add WeChat edu\_assist\_pro

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

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

# Assignment Project Exam Help

<https://eduassistpro.github.io>

Add WeChat edu\_assist\_pr

(From Mackay, p140; see his exercise 8.8)

# Mutual Information

Example 1 (from Mackay, 2003)

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

$$p(X = 0) = p \quad p(X = 1) = 1 - p$$

<https://eduassistpro.github.io>

(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)

# Assignment Project Exam Help

(a) As  $X \rightarrow Y$  and  $q = 1/2$  the noise will flip the outcome of  $X$  with probability  $q$ .  
pro re:

<https://eduassistpro.github.io>

Hence:

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

Add WeChat [edu\\_assist\\_pro](#)

Indeed for  $q = 1/2$  we see that  $Z \perp X$



## Mutual Information

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

(b)

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

So <https://eduassistpro.github.io/> (flip)

$= (1-p)q + p(1-q)$   
Add WeChat edu\_assist\_pro

and:

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

1 Decomposability of Entropy

# Assignment Project Exam Help

2 Relative Entropy / KL Divergence

3 Mutu

- De

- Joint and Conditional Mutual Information

<https://eduassistpro.github.io>

Add WeChat edu\_assist\_pr

4 Wrapping up

## Joint Mutual Information

Recall that for random variables  $X, Y$ ,

**Assignment Project Exam Help**

- Reduction in uncertainty in  $X$  due to knowledge of  $Y$

More generally

**<https://eduassistpro.github.io>**

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

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

**Add WeChat edu\_assist\_pro**

Symmetry also generalises:

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

## Conditional Mutual Information

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

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

Assignment Project Exam Help

Averaging over  $Z$  we obtain:

The condi

<https://eduassistpro.github.io>

$$= \mathbb{E}_{p(X,Y,Z)} \log \frac{p(X, Y, Z)}{p(X)p(Y)p(Z)}$$

Add WeChat edu\_assist\_pro

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.

1 Decomposability of Entropy

# Assignment Project Exam Help

2 Relative Entropy / KL Divergence

3 Mutual Information

- De
- Joint and Conditional Mutual Information

<https://eduassistpro.github.io>

Add WeChat edu\_assist\_pr

4 Wrapping up

# Assignment Project Exam Help

- Decomposability of entropy

- Rel

<https://eduassistpro.github.io>

- Mutual information

Add WeChat edu\_assist\_pr

- **Reading:** Mackay §2.5, Ch 8; Cover & Thomas §2.3 to §2.5

Next time

# Assignment Project Exam Help

Mutual information chain rule

Jensen'

"Informa

Data processing inequality

<https://eduassistpro.github.io>

Add WeChat edu\_assist\_pr