COMP2610 / COMP6261 - Information Theory

Lecture 7: Relative Entropy and Mutual Information

Australian National University

13 August 2018

Last time

- Information content and entropy: definition and computation

- Ent

- Entropy and minimum expected number of bi

- Joint and conditional entropies, chain rule

Information Content: Review

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

Let $p(x)$

The (Sha

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

As $p(x) \to 0$, $h(x) \to +\infty$ (rare outcomes ar

Entropy: Review

The entropy is the average information content of all outcomes:

$$\frac{1}{\quad}$$

Entropy i m:

$$0 \leq H(X) \leq \text{lo}$$

Entropy is related to minimal number of bits needed t variable

This time

- The decomposability property of entropy
- Rel
- Mutual information

Outline

1. Decomposability of Entropy

2. Relat

3. Mutu
   - Definition
   - Joint and Conditional Mutual Information

4. Wrapping up

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 $X \neq$ … r $X = 3$

The prob…

$$p(X = 1) = $$
$$p(X = 2) = $$
$$p(X = 3) = \frac{1}{4}$$

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 imagi

1. Firs
   -
   -

2. If $X \neq 1$ we learn the value of the second coin fli
   - Also binary variable with $\mathbf{p}^{(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 $\mathbf{p} = (p_1, \ldots, p_{|\mathcal{X}|})$:

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

$H(p_1, 1$ $X =$ "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 $\neq 1$.

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

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

$$H(\mathbf{p}) = H\left(\sum_{i=1}^{m} p_i, \sum_{i=m+1}^{|\mathcal{X}|} p_i\right)$$

$$+ \left(\sum_{i=1}^{m} p_i\right) H\left(\frac{p_1}{\sum_{i=1}^{m} p_i}, \ldots, \frac{p_m}{\sum_{i=1}^{m} p_i}\right)$$

$$+ \left(\sum_{i=m+1}^{|\mathcal{X}|} p_i\right) H\left(\frac{p_{m+1}}{\sum_{i=m+1}^{|\mathcal{X}|} p_i}, \ldots, \frac{p_{|\mathcal{X}|}}{\sum_{i=m+1}^{|\mathcal{X}|} p_i}\right)$$

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 t

What hap

- e.g. b n

If the true distribution is p, but we assume it is 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:

$1 \quad 1$

$x\in\mathcal{X}$

- Note:

  - Both $p(X)$ and $q(X)$ are defined over t
- 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

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

- Not s

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

  - Not a true distance since is not symmetric an triangle inequality
  - Hence, “KL divergence” rather than “KL distance”

Relative Entropy

Uniform q

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

Relative entropy between p and q:

$$x \in \mathcal{X}$$

$$= -H(X) + \sum_{x \in \mathcal{X}}$$

$$= -H(X) + \log |\mathcal{X}|.$$

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:

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

Relative Entropy

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

$$D_{KL}(p \; q) = \theta_p \log \frac{\theta_p}{} + (1 \quad \theta_p) \log \frac{1 - \theta_p}{}$$

bits

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

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

1 Decomposability of Entropy

Assignment Project Exam Help

2 Relative Entropy / KL Divergence

3 Mutu https://eduassistpro.github.i

- De
- Joint and Conditional Mutual Information

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 distributi

$$= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y$$

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:

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

$$p(x\ y)$$

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

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

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:

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

- Mut

- Self-information:

$$I(X;X) =$$

- 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 $X, Y \in \{0, 1\}$, $X \perp Y$ and:

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

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

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

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

Hence:

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

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

Mutual Information

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

(b)

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

Si

(flip)

$$
= (1-p)q + p(1-
$$

$$
= q + p - 2pq
$$

and:

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

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 gen

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

- Reduction in uncertainty in $X_1,\ldots,X_n$ du ... $_m$

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 $Z = z_k$:

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

Averaging over Z we obtain:

The condi

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

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
- Rel
- Mutual information
- Reading: Mackay §2.5, Ch 8; Cover & Thomas §2.3 to §2.5

Next time

Mutual information chain rule

Jensen’

“Informa

Data processing inequality