Mansoura University
Faculty of Computers and Information
Department of Information System

[IS311T] Information Theory

Grade: 3rd grade (IS)

Lecture 03: Entropy, relative entropy, and mutual information

Dr. Islam Reda

Outline

- Entropy
- Joint entropy
- Conditional entropy
- Chain rule for entropy
- Relative entropy
- Mutual information

Information

- Information theory is concerned with representing data in a compact fashion (data compression or source coding), as well as with transmitting and storing it in a way that is robust to errors (error correction or channel coding).
- Information theory is concerned with quantifying information for communication (subfield of mathematics).
- The intuition behind quantifying information is the idea of measuring how much surprise there is in an event.
  - ➢ Low probability event (rare): High information (surprising).
  - ➢ High probability event (common): Low information (unsurprising).
- Example:
  - "The sun rose this morning" → uninformative
  - "There was a solar eclipse this morning" → very informative

2.1 Entropy

- Entropy is a measure of the uncertainty of a random variable.
- Let X be a discrete random variable with alphabet $\mathcal{X}$ and probability mass function $p(x) = \Pr\{X = x\}, x \in \mathcal{X}$.
- The entropy of a discrete random variable X with a probability mass function $p(x)$ is defined by:

$$H(X) = \sum_x p(x) \log(\frac{1}{p(x)}) = -\sum_x p(x) \log p(x).$$

- We use logarithms to base 2. The entropy will then be measured in bits.
- If the base of the logarithm is e, the entropy is measured in nats.

2.1 Entropy

- The entropy is a measure of the average uncertainty in the random variable.
- It is the number of bits on average required to describe the random variable.
- Entropy is a function of the distribution of X. It does not depend on the actual values taken by the random variable X, but only on the probabilities.
- The entropy of X can also be interpreted as the expected value of the random variable $\log\frac{1}{p(X)}$.

$$H(X) = E(\log\frac{1}{p(X)})$$

2.1 Entropy

- **Example 1.1.1:** Consider a random variable that has a uniform distribution over 32 outcomes. The entropy of this random variable is:

$$H(X) = \sum_{i=1}^{32} p(i) \log \frac{1}{p(i)} = \sum_{i=1}^{32} \frac{1}{32} \log 32 = \log 32 = 5\ bits.$$

- **Example 1.1.2:** Suppose that we have a horse race with eight horses taking part. Assume that the probabilities of winning for the eight horses are$(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64})$. The entropy of the horse race is:

$$H(x) = \frac{1}{2} \log 2 + \frac{1}{4} \log 4 + \frac{1}{8} \log 8 + \frac{1}{16} \log 16$$

$$+4 \times \frac{1}{64} \log 64 = 2\ bits$$

2.1 Entropy

- $H(X) \geq 0$.
- Proof: $0 \leq p(x) \leq 1$ *implies that* $\log \frac{1}{p(x)} \geq 0$
- **Example 2.1.1:** Let

$$X = \begin{cases} 1 & \text{with probability } p, \\ 0 & \text{with probability } 1-p. \end{cases}$$

Then

$$H(X) = -p \log p - (1-p)\log(1-p) \overset{\text{def}}{=\!=} H(p).$$

- $H(X) = 1\, bit$ when $p = \frac{1}{2}$. The graph of the function $H(p)$ is shown in Figure 2.1. This figure represents the **binary entropy function**.

2.1 Entropy

- $H(p)$ is a concave function.
- $H(x) = 0$ when $p = 0 \ or \ 1$. When $p = 0 \ or \ 1$, the variable is not random and there is no uncertainty.
- Similarly, the uncertainty is maximum when $p = \frac{1}{2}$.

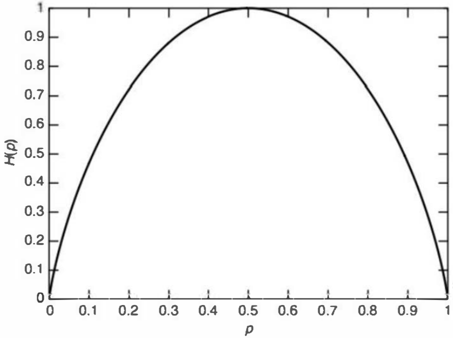

FIGURE 2.1. $H(p)$ vs. p.

2.1 Entropy

$$X = \begin{cases} a & \text{with probability } \frac{1}{2}, \\ b & \text{with probability } \frac{1}{4}, \\ c & \text{with probability } \frac{1}{8}, \\ d & \text{with probability } \frac{1}{8}. \end{cases}$$

The entropy of X is

$$H(X) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{8}\log\frac{1}{8} - \frac{1}{8}\log\frac{1}{8} = \frac{7}{4} \text{ bits.}$$

- The entropy of a random variable is a lower bound on the average number of bits required to represent the random variable.
- In the previous example, if X has a uniform distribution, the entropy is maximized.

The entropy in that case will equal?

2.2 Joint entropy and conditional entropy

- Entropy is the uncertainty of a single random variable.
- We now extend the definition to a pair of random variables.
- The joint entropy $H(X,Y)$ of a pair of discrete random variables (X,Y) with a joint distribution $p(x,y)$ is defined as

$$H(X,Y) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log \left(\frac{1}{p(x,y)}\right)$$

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

which can also be expressed as

$$H(X,Y) = -E \log p(X,Y).$$

2.2 Joint entropy and conditional entropy

- **Conditional entropy** $H(X|Y)$ is the entropy of a random variable conditional on the knowledge of another random variable.
- Conditional entropy $H(X|Y)$ is defined as:

$$H(X|Y) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y)\log(\frac{1}{p(x|y)})$$

$$H(X|Y) = -E\log p(X|Y).$$

- Conditioning reduces entropy.
  - $H(X|Y) \leq H(X)$
  - When are the two quantities equal?

Chain rule for entropy

- The entropy of a pair of random variables = the entropy of one + the conditional entropy of the other.

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

- Proof

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

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

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

Chain rule for entropy

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

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

- $H(X,Y) \; = H(X) + H(Y \big| X) = H(Y) + H(X|Y)$
- $H(Y \big| X) = H(X,Y) - H(X)$
- $H(X \big| Y) = H(X,Y) - H(Y)$
- It follows that
- $H(X,Y|Z) = H(X|Z) + H(Y|X,Z)$.

2.2 Joint entropy and conditional entropy

- **Example 2.2.1** Let (X,Y) have the following joint distribution:

X

		1	**2**	**3**	**4**
Y	1	1/8	1/16	1/32	1/32
	2	1/16	1/8	1/32	1/32
	3	1/16	1/16	1/16	1/16
	4	1/4	0	0	0

Calculate $H(X), H(Y), H(X|Y), H(Y|X), H(X,Y)$

2.2 Joint entropy and conditional entropy

- **Solution**

X

Y

	1	**2**	**3**	**4**	**$p(Y)$**
1	1/8	1/16	1/32	1/32	1/4
2	1/16	1/8	1/32	1/32	1/4
3	1/16	1/16	1/16	1/16	1/4
4	1/4	0	0	0	1/4
$p(X)$	1/2	1/4	1/8	1/8	1

$$H(X) = \frac{1}{2}\log 2 + \frac{1}{4}\log 4 + \frac{1}{8}\log 8 + \frac{1}{8}\log 8 = \frac{7}{4} \text{ bits}$$

$$H(Y) = 4 \times (\frac{1}{4}\log 4) = 2 \text{ bits}$$

2.2 Joint entropy and conditional entropy

X

		1	2	3	4
Y	1	1/8	1/16	1/32	1/32
	2	1/16	1/8	1/32	1/32
	3	1/16	1/16	1/16	1/16
	4	1/4	0	0	0

$$H(X,Y) = \left[\frac{1}{8}\log 8 + \frac{1}{16}\log 16 + \frac{1}{32}\log 32 + \frac{1}{32}\log 32\right]$$

$$+ \left[\frac{1}{16}\log 16 + \frac{1}{8}\log 8 + \frac{1}{32}\log 32 + \frac{1}{32}\log 32\right]$$

$$+ 4 \times \left[\frac{1}{16}\log 16\right] + \left[\frac{1}{4}\log 4\right] = \frac{27}{8}\text{ bits}$$

$$H(X|Y) = H(X,Y) - H(Y) = \frac{27}{8} - 2 = \frac{11}{8}\,bits$$

$$H(Y|X) = H(X,Y) - H(X) = \frac{27}{8} - \frac{7}{4} = \frac{13}{8}\,bits$$

2.3 Relative entropy and mutual information

- Mutual information $I(X;Y)$ is the reduction in uncertainty due to another random variable.
- $I(X;Y)$ is a measure of the amount of information that one random variable contains about another random variable.
- $I(X;Y)$ is a measure of the dependence between the two random variables.
- It is symmetric in X and Y and always nonnegative.
- It is equal to zero if and only if X and Y are independent.

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

- Mutual information is a special case of a more general quantity called relative entropy.

2.3 relative entropy and mutual information

- Relative entropy or Kullback–Leibler distance $D(p||q)$ is a measure of the "distance" between two probability mass functions p and q.

$$D(p||q) = \sum_{x} p(x) \log \frac{p(x)}{q(x)}$$

- $Relative\ entropy \geq 0$ and is zero if and only if $p = q$.
- It is not a true distance between distributions since it is not symmetric and does not satisfy the triangle inequality.
- Mutual information $I(X;Y)$ is the relative entropy between the joint distribution and the product distribution $p(x)p(y)$.

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

THANK YOU