Mathematical Foundations

Foundations of Statistical Natural Language Processing, chapter2

Presented by Jen-Wei Kuo（郭人瑋）
CSIE, NTNU
rogerkuo@csie.ntnu.edu.tw

Reference

- A First Course in Probability -Sheldon Ross
- Probability and Random Processes for Electrical Engineering -Algerto Leon-Garcia

Outline

- Elementary Probability Theory
  - Probability spaces
  - Conditional probability and independence
  - Bayes' theorem
  - Random variables
  - Expectation and variance
  - Joint and conditional distributions
  - Gaussian distributions
- Essential Information Theory
  - Entropy
  - Joint entropy and conditional entropy
  - Mutual information
  - Relative entropy or Kullback-Leibler divergence

Essential Information Theory

Entropy

- Entropy measures the amount of information in a random variable. It is normally measured in bits.

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

- We define

$$0 \log_2 0 = 0$$

Essential Information Theory

Entropy

- Example:

  Suppose you are reporting the result of rolling an 8-sided die. Then the entropy is:

$$H(X) = -\sum_{i=1}^{8} p(i)\log p(i) = -\sum_{i=1}^{8} \frac{1}{8}\log\frac{1}{8}$$

$$= -\log\frac{1}{8} = \log 8 = 3bits$$

Essential Information Theory

Entropy

- Entropy代表要傳遞這件事的平均資訊量，當我們建立系統時，希望Entropy愈低愈好。
- 傳遞機率時，由於機率不會超過1，故我們只需傳遞分母的值即可。

Essential Information Theory

Entropy

- Properties of Entropy:

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

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

$$= E\left(\log \frac{1}{p(x)} \right)$$

Essential Information Theory

Joint Entropy and Conditional Entropy

- Joint Entropy:

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

- Conditional Entropy:

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

Essential Information Theory

Joint Entropy and Conditional Entropy

- Proof of Conditional Entropy:

$$H(Y \mid X) = \sum_{x \in X} p(x) H(Y \mid X = x)$$

$$= \sum_{x \in X} p(x) \left[-\sum_{y \in Y} p(y \mid x) \log p(y \mid x) \right]$$

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

Essential Information Theory

Joint Entropy and Conditional Entropy

- Chain rule for Entropy:

$$H(X,Y) = H(X) + H(Y \mid X)$$

- Proof:

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

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

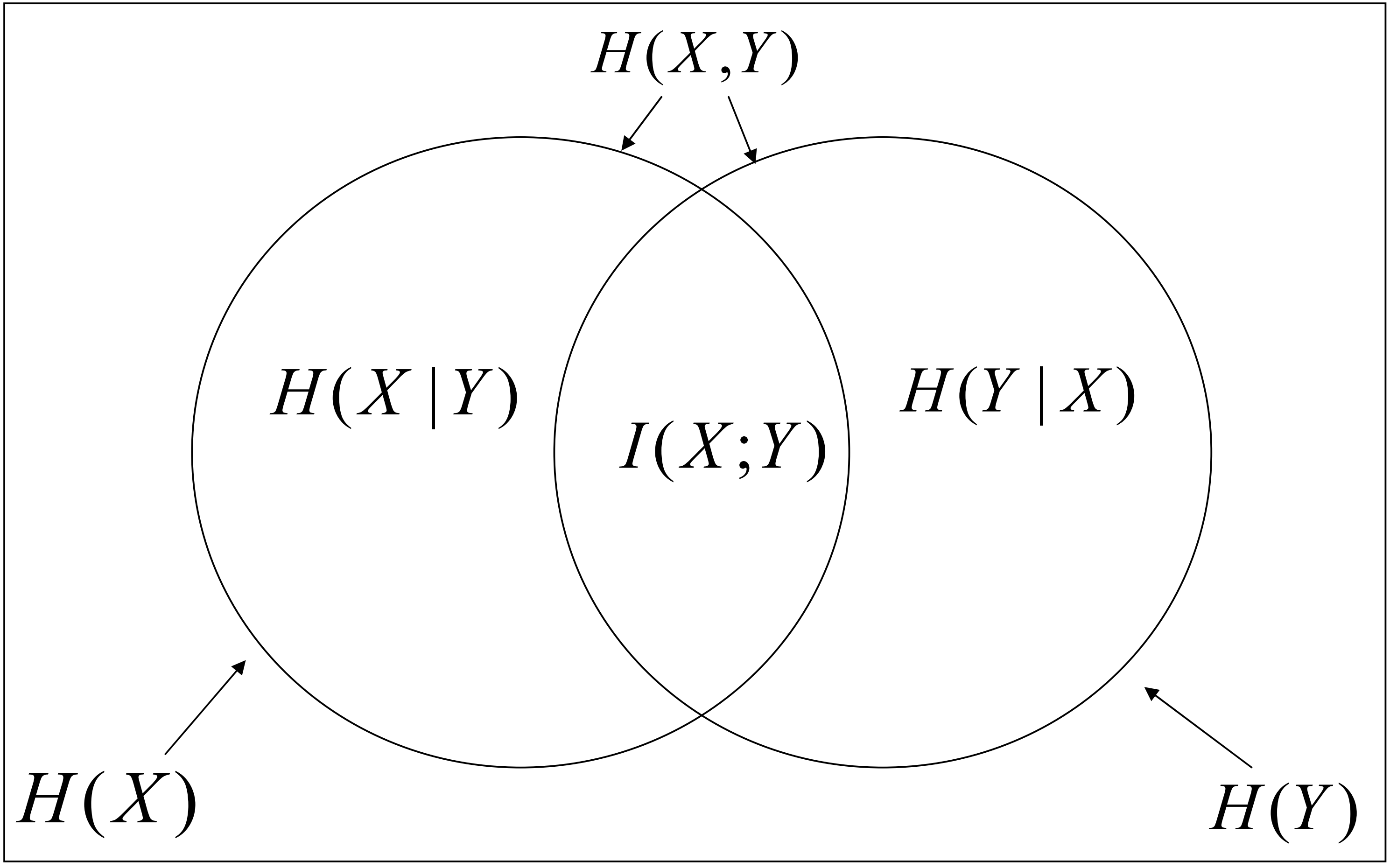
$$= -\sum_{x \in X} \sum_{y \in Y} p(x,y)\big(\log p(y \mid x) + \log p(x)\big)$$

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

$$= H(Y \mid X) + H(X)$$

Essential Information Theory

Mutual Information



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

Essential Information Theory
Mutual Information

- This difference is called the ***mutual information*** between X and Y.

- The amount of information one random variable contains about another.

- It is 0 only when two variables are independent.
  也就是說，兩個獨立事件的mutual Information爲0。

Essential Information Theory

Mutual Information

- How to simply calculate Mutual Information？

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

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

$$= \sum_x p(x)\log\frac{1}{p(x)} + \sum_y p(y)\log\frac{1}{p(y)} + \sum_{x,y} p(x,y)\log p(x,y)$$

$$= \sum_{x,y} p(x,y)\log\frac{1}{p(x)} + \sum_{x,y} p(x,y)\log\frac{1}{p(y)} + \sum_{x,y} p(x,y)\log p(x,y)$$

$$= \sum_{x,y} p(x,y)\left[\log\frac{1}{p(x)} + \log\frac{1}{p(y)} - \log\frac{1}{p(x,y)}\right]$$

$$= \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

Essential Information Theory

Mutual Information

- Define the ***pointwise mutual information*** between two particular points.

$$I(x,y) = \log \frac{p(x,y)}{p(x)p(y)}$$

This has sometimes been used as a measure of association between elements.

Essential Information Theory

Relative Entropy or Kullback-Leibler divergence

- For two probability mass functions, *p(x)* , *q(x)* their relative entropy is given by:

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

define $\quad 0 \log \frac{0}{q} = 0 \quad \textit{and} \quad p \log \frac{p}{0} = \infty$

Essential Information Theory

Relative Entropy or Kullback-Leibler divergence

- 意義：It is the average number of bits that are wasted by encoding events from a distribution p with a code based on a not-quite-right distribution q.

- Some authors use the name "KL distance", but note that relative entropy isn't a metric (it doesn't satisfy the triangle inequality)

Essential Information Theory

Relative Entropy or Kullback-Leibler divergence

Properties of KL-divergence:

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

$$= D(p(x,y) \,||\, p(x)p(y))$$

Define the Conditional Relative Entropy:

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

Essential Information Theory

Relative Entropy or Kullback-Leibler divergence

Properties of KL-divergence:

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

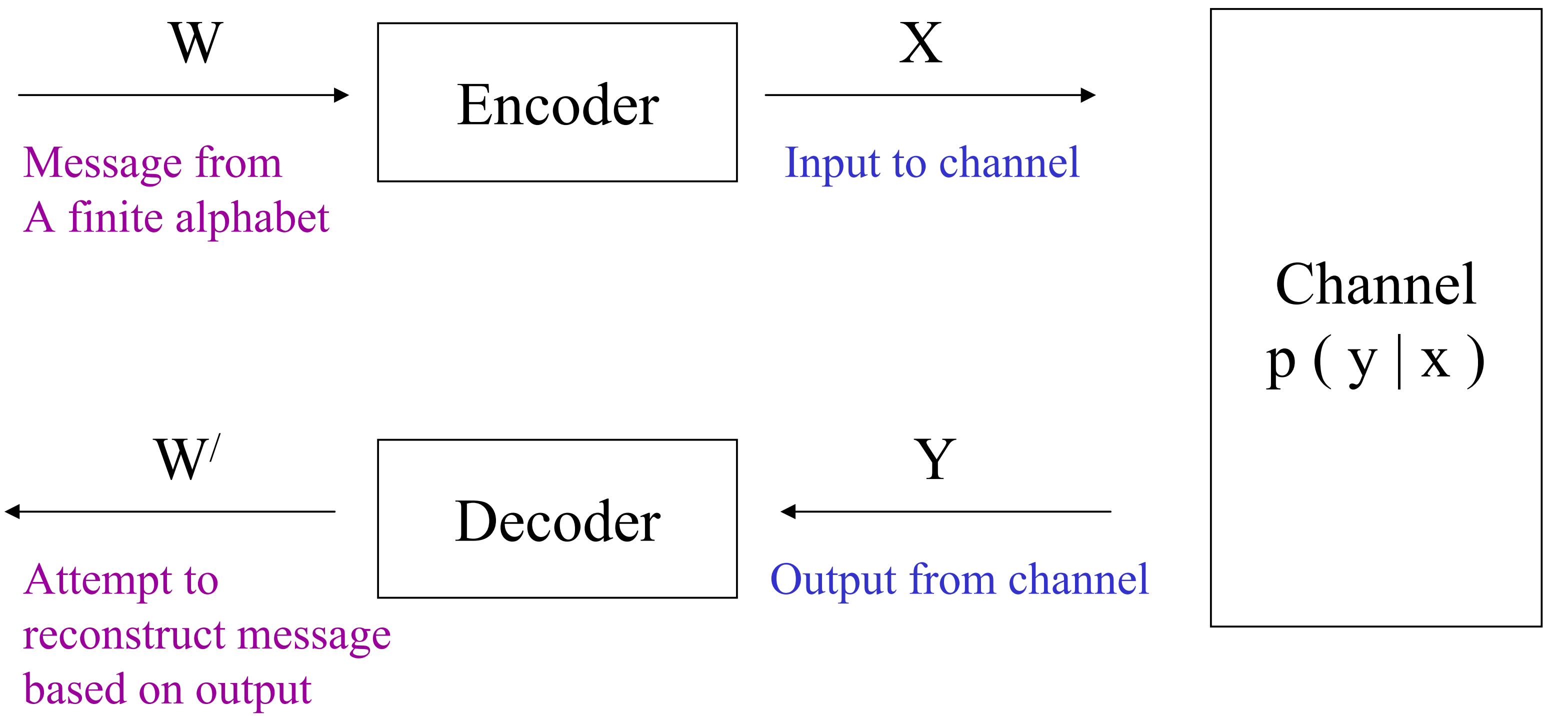
$$= D(p(x,y) \,\|\, p(x)p(y))$$

Define the Conditional Relative Entropy:

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

Essential Information Theory

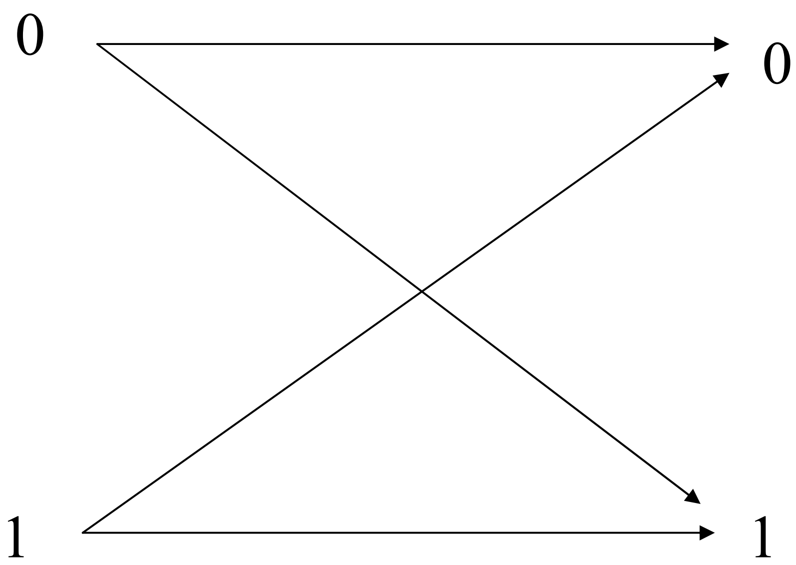
The noisy channel model



The noisy channel model

Essential Information Theory

The noisy channel model



A binary symmetric channel

Essential Information Theory

The noisy channel model

Capacity:

The channel capacity describes the rate at which one can transmit information through the channel with an arbitrarily low probability of being unable to recover the input from the output.

$$C = \max_{p(X)} I(X;Y) = \max_{p(X)} H(Y) - H(Y \mid X) = H(Y) - H(p) = 1 - H(p)$$

$$0 < C \leq 1$$

$$if \quad p = 0 \quad or \quad p = 1 \ \Rightarrow C = 1$$

$$if \quad p = \frac{1}{2} \ \Rightarrow C = 0$$

Essential Information Theory

The noisy channel model

Application: (In speech recognition)

Input: *word sequences*
Output: *observed speech signal*
P(input): *probability of word sequences*
P(output|input): *acoustic model (channel prob.)*

Bayes' theorem

$$\hat{I} = \arg\max_i p(i \mid o) = \arg\max_i \frac{p(i)\,p(o \mid i)}{p(o)} = \arg\max_i \boxed{p(i)}\boxed{p(o \mid i)}$$

Essential Information Theory

Cross entropy

Cross entropy:

The cross entropy between a random variable X with true probability distribution p(X) and another pmf q (normally a model of p) is given by:

$$\begin{aligned}
H(X,q) &= H(X) + D(p \| q) \\
&= \sum_{x \in X} p(x) \log \frac{1}{p(x)} + \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} \\
&= \sum_{x \in X} p(x) \left[\log \frac{1}{p(x)} + \log \frac{p(x)}{q(x)} \right] \\
&= \sum_{x \in X} p(x) \left[\log \frac{1}{q(x)} \right] \\
&= -\sum_{x \in X} p(x) \log q(x)
\end{aligned}$$

Essential Information Theory

Cross entropy

Cross entropy of a language :

suppose

Language L = (X_i) ~ p(x) according to a model m by

$$H(L,m) = -\lim_{n\to\infty}\frac{1}{n}\sum_{x_{1n}} p(x_{1n})\log m(x_{1n})$$

We cannot calculate this quantity without knowing p. But if we make certain assumptions that the language is 'nice,' then the cross entropy for the language can be calculated as:

$$H(L,m) = -\lim_{n\to\infty}\frac{1}{n}\log m(x_{1n})$$

Essential Information Theory

Cross entropy

Cross entropy of a language :

We do not actually attempt to calculate the limit, but approximate it by calculating for a sufficiently large n:

$$H(L,m) \approx -\frac{1}{n}\log m(x_{1n})$$

This measure is just the figure for our average surprise. Our goal will be to try to minimize this number. Because H(X) is fixed, this is equivalent to minimizing the relative entropy, which is a measure of how much our probability distribution departs from actual language use.

Essential Information Theory

Perplexity

In the speech recognition community, people tend to refer to perplexity rather than cross entropy. The relationship between the two is simple:

$$Perplexity(x_{1n}, m) = 2^{H(x_{1n}, m)}$$

$$= 2^{-\frac{1}{n}\log m(x_{1n})}$$

$$= m(x_{1n})^{-\frac{1}{n}}$$

Why we use perplexity not cross entropy?

Because it is much easier to impress funding bodies by saying that "we've managed to reduce perplexity from 950 to only 540" than by saying that "we've reduced cross entropy from 9.9 to 9.1 bits."