Statistics for Computing

MA4413 Lecture 11B

Kevin O'Brien

kevin.obrien@ul.ie

Dept. of Mathematics & Statistics,
University *of* Limerick

Autumn 2013

Entropies: Example (a)

- The input source to a noisy communication channel is a random variable X over the four symbols $\{a, b, c, d\}$.
- The output from this channel is a random variable Y over these same four symbols.

Entropies: Example (b)

The joint distribution of these two random variables is as follows:

	x=a	x=b	x=c	x=d
y=a	1/8	1/16	1/16	1/4
y=b	1/16	1/8	1/16	0
y=c	1/32	1/32	1/16	0
y=d	1/32	1/32	1/16	0

Entropies: Example (c)

- Write down the marginal distribution for X and compute the marginal entropy $H(X)$.
- Write down the marginal distribution for Y and compute the marginal entropy $H(Y)$.
- What is the joint entropy $H(X,Y)$ of the two random variables?
- What is the conditional entropy $H(Y|X)$?
- What is the conditional entropy $H(X|Y)$?
- What is the mutual information $I(X;Y)$ between the two random variables?

Entropies: Example (d)

The marginal distribution of these two random variables is as follows:

	x=a	x=b	x=c	x=d	P(Y)
y=a	1/8	1/16	1/16	1/4	0.50
y=b	1/16	1/8	1/16	0	0.25
y=c	1/32	1/32	1/16	0	0.125
y=d	1/32	1/32	1/16	0	0.125
P(X)	0.25	0.25	0.25	0.25	

Entropies: Example (e)

- H(X) , the entropy of X, is computed as

$$H(X) = -\sum P(x_i)\log_2 P(x_i)$$

- $H(X) = (-0.25 \times -2) + (-0.25 \times -2) + (-0.25 \times -2) + (-0.25 \times -2)$
- $H(X) = 2$b

- H(X) , the entropy of Y, is computed as

$$H(Y) = -\sum P(y_j)\log_2 P(y_j)$$

- $H(Y) = (-0.5 \times -1) + (-0.25 \times -2) + (-0.125 \times -3) + (-0.125 \times -3)$
- $H(Y) = 1.75$b

Entropies: Example (f)

- To compute the joint entropy $H(X,Y)$, we will use $H(X,Y) = -\sum P(x_i, y_y) \log_2 P(x_i, y_j)$
- This means we should compute the entropy component for each cell of the table, and sum up all the resultant terms.
- To save time, we will aggregate similar results,
  - there are 4 cells where the probability is $1/32$,
  - 6 cells with probability $1/16$,
  - 2 cells with probability $1/8$
  - and 1 cell with probability $1/4$.
- Solving

$$H(X,Y) = [4 \times -\frac{1}{32}\log_2\frac{1}{32}] + [6 \times -\frac{1}{16}\log_2\frac{1}{16}] + \ldots + [1 \times -\frac{1}{4}\log_2\frac{1}{4}]$$

Entropies: Example (g)

- Simplifying

$$H(X,Y) = [4 \times -\frac{1}{32}\log_2\frac{1}{32}] + [6 \times -\frac{1}{16}\log_2\frac{1}{16}] + \ldots + [1 \times -\frac{1}{4}\log_2\frac{1}{4}]$$

- Simplifying

$$H(X,Y) = [-\frac{4}{32} \times -5] + [-\frac{6}{16} \times -4] + [-\frac{2}{8} \times -3] + [-\frac{1}{4} \times -2]$$

- $H(X,Y) = 27/8$ b.

Entropies: Example (h)

From last lecture, two useful relationships among the types of entropies are

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

Re-arranging these formulae

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

Entropies: Example (i)

Re-arranging these formulae

- $H(X|Y) = H(X,Y) - H(Y) = 27/8 - 14/8 = 13/8$ b.
- $H(Y|X) = H(X,Y) - H(X) = 27/8 - 16/8 = 11/8$ b.
- Remark $1.75 = 14/8$ and $2 = 16/8$.
- Also: we will derive $H(Y|X$ and $H(X|Y)$ from first principles in a tutorial.

Entropies: Example (j)

There are three alternative ways to obtain the answer:

- $I(X;Y) = H(Y) - H(Y|X) = 7/4 - 11/8 = 3/8$ b.
- $I(X;Y) = H(X) - H(X|Y) = 2 - 13/8 = 3/8$ b.
- $I(X;Y) = H(X) + H(Y) - H(X,Y) =$
  $2 + 7/4 - 27/8 = (16 + 14 - 27)/8 = 3/8$b.

Channel Capacity

- In information theory, channel capacity is the most conservative upper bound on the amount of information that can be reliably transmitted over a communications channel.
- It is given by the maximum of the mutual information between the input and output of the channel (maximum in respect to input probabilities).

Channel Capacity

A. Channel Capacity per Symbol C:
The channel capacity per symbol of a DMC is defined as

$$C_s = \max_{(P(x_i))} I(X;Y) \text{ b/symbol}$$

where the maximization is over all possible input probability distributions $P(x_i)$ on X. Note that the channel capacity C_s is a function of only the channel transition probabilities that define the channel.

B. Channel Capacity per Second :

If r symbols are being transmitted per second, then the maximum rate of transmission of information per second is rC_s.

This is the channel capacity per second and is denoted by C (b/sec).

$$C = rC_s \text{ b/sec}$$

Capacities of special channels

Lossless Channel

- For a lossless channel, the mutual information (information transfer) is equal to the input (source) entropy), and no source information is lost in transmission.
- It can be shown that $H(X|Y) = 0$ (If y_i is the output, there is certainty about the input). Also $I(X;Y) = H(X)$.
- Consequently, the channel capacity per symbol is

$$C_s = \max_{P(x_i)} H(X) = \log_2 m$$

  where m is the number of symbols in X.
- For example, if there are $m = 4$ input channels, then $C = \log_2 4 = 2$ b/symbol

Capacities of special channels

Deterministic Channel:

- The mutual information (information transfer) is equal to the output entropy.
- It can be shown that $H(Y|X) = 0$ (If x_i is the input, there is certainty about the output). Also $I(X;Y) = H(Y)$.
- The channel capacity per symbol is

$$C_s = \max_{P(x_i)} H(Y) = \log_2 n$$

where n is the number of symbols in Y.

Capacities of special channels

Noiseless Channel:

- Since a noiseless channel is both lossless and deterministic , we can say that $I(X;Y) = H(X) = H(Y)$. The mutual information (information transfer) is equal to the output entropy).
- The channel capacity per symbol is

$$C_s = \log_2 m = \log_2 n$$

Capacities of special channels

Binary Symmetric Channel:

- It can be shown that, for a binary symmetric channel, the the channel capacity per symbol is

$$C_s = 1 + p\log_2 p + 1 - p\log_2(1-p)$$

Source Coding

- A conversion of the output of a DMS into a sequence of binary symbols (binary code word) is called ***source coding***.
- The device that performs this conversion is called the source encoder.
- An objective of source coding is to minimize the average bit rate required for representation of the source by reducing the redundancy of the information source.

Source Coding : Code Length and Code Efficiency

- Let X be a DMS with finite entropy $H(X)$ and an alphabet $\{x_1, \ldots, x_m\}$, each with corresponding probabilities of occurrence $P(x_i)$.
- Let the binary code word assigned to symbol x_i by the encoder have length n_i b.
- The length of a code word is the number of binary digits in the code word. The average code word length L, per source symbol is given by

$$E(L) = \sum_{i=1}^{m} P(x_i) n_i$$

Source Coding : Code efficiency and Code redundancy

- The parameter L (estimated by $E(L)$) represents the average number of bits per source symbol used in the source coding process.
- The code efficiency is defined as

$$\eta = \frac{L_{min}}{L}$$

  where L_{min} is the minimum possible value of L. When η approaches unity, the codes is said to be efficient.
- The code redundancy γ is defined as $\gamma = 1 - \eta$.

Source Coding Theorem

- The source coding theorem states that for a DMS X with entropy $H(X)$, the average code word length L per symbol is bounded as $L \geq H(X)$
- Furthermore L can be made as close to H(X) as required for some suitably chosen code.
- Thus, with $L_{min} \geq H(X)$, the code efficiency can be rewritten as

$$\eta = \frac{H(X)}{L}$$

- We will use this definition for efficiency. (Remark L is estimable by $E(L)$.)

Classification of Codes

In this section we look at to classify codes according to the following categories.

1. Fixed Length Codes
2. Variable Length Codes
3. Distinct Codes
4. Prefix-Free Codes
5. Uniquely decodable codes
6. Instantaneous Codes
7. Optimal Codes

Classification of Codes

Classification of codes is best illustrated by an example. Consider the table below where a source of size 4 has been encoded in binary codes with symbol 0 and 1.

X	Code 1	Code 2	Code 3	Code 4	Code 5	Code 6
x_1	00	00	0	0	0	1
x_2	01	01	1	10	01	01
x_3	00	10	00	110	011	001
x_4	11	11	11	111	0111	0001

1. Fixed-Length Codes: A fixed-length code is one whose code word length is fixed. Code 1 and code 2are fixed-length codes with length 2.
2. Variable-Length Codes: A variable-length code is one whose code word length is not fixed. All codes except codes 1 and 2 are variable-length codes.
3. Distinct Codes: A code is distinct if each code word is distinguishable from other code words. All codes except code 1 are distinct codes. Notice the codes for x_l and x_3.
4. Prefix-Free Codes: A code in which no code word can be formed by adding code symbols to another code word is called a prefix-free code. Thus, in a prefix-free code no code word is a prefix of another. Codes 2, 4, and 6 are prefix-free codes.

5. Uniquely Decodable Codes

- A distinct code is uniquely decodable if the original source sequence can be reconstructed perfectly from the encoded binary sequence.
- Note that code 3 is not a uniquely decodable code.
- For example, the binary sequence 1001 may correspond to the source sequences $x_2x_3x_2$ or $x_2x_1x_1x_2$.
- A sufficient condition to ensure that a code is uniquely decodable is that no code word is a prefix of another.
- Thus, the prefix-free codes 2, 4, and 6 are uniquely decodable codes. Note that the prefix -free condition is not a necessary condition for unique decodability.
- For example, code 5 does not satisfy the prefix-free condition, and yet it is uniquely decodable since the bit 0 indicates the beginning of each code word of the code.

6. Instantaneous Codes
   - A uniquely decodable code is called an instantaneous code if the end of any code word is recognizable without examining subsequent code symbols.
   - The instantaneous codes have the property previously mentioned that no code word is a prefix of another code word.
7. Optimal Codes
   - A code is said to be optimal if it is instantaneous and has minimum average length L for a given source with a given probability assignment for the source symbols.

Kraft inequality

- Let X be a DMS with alphabet ($x_i = \{1, 2, \ldots, m\}$). Assume that the length of the assigned binary code word corresponding to x, is n.
- A necessary and sufficient condition for the existence of an instantaneous binary code is

$$K = \sum_{i=1}^{m} 2^{-n_i} \leq 1$$

  which is known as the **Kraft inequality**.
- Note that the Kraft inequality assures us of the existence of an instantaneously decodable code with code word lengths that satisfy the inequality. But it does not show us how to obtain these code words, nor does it say that any code that satisfies the inequality is automatically uniquely decodable