

## Information Theory More

Linda Doyle  
CTVR

### point of learning

- we began the last day to look at two fundamental questions
  - the first - **what is the limit of compression?** - minimum to which something can be compressed
  - the second - **what is the maximum capacity of a channel?** - maximum information that can be sent over a channel
- to answer these we had to begin to find a measure or way of describing information and uncertainty
- this is where **entropy** came to hand
- we still have more basic building block work to do to answer these questions
- today we do more of that .....

### recall - entropy

The entropy of a random variable  $X$  with a probability mass function  $p(x)$  is defined by

$$H(X) = - \sum_x p(x) \log_2 p(x). \quad (1.1)$$

We also write  $H(p)$  for the above quantity.

### the memory less source

- Even though I did not say it explicitly during the last lecture we were looking at the entropy of a **discrete memoryless source**.
- The random variable  $X$  was produced by a **discrete source** - the message was generated symbol by symbol based on some set of probabilities.
- And entropy,  $H(X)$  is the number of bits on average required to describe the random variable,  $X$ .

### the source - thing that produces the random variable $X$

- A physical system, or a mathematical model of a system which produces such a sequence of symbols governed by a set of probabilities, is known as a **stochastic process**.
- We consider a discrete source, therefore, to be represented by a stochastic process

let's look at an experiment Shannon did to further explore concepts of interest

assume a 27 letter alphabet - 26 letters and a space

create a model for a discrete source that outputs the English language

### 1. Zero-order approximation

Suppose now all symbols of the alphabet are equally probable the following could be obtained

XFOML RXKHRJFFJUJ ZLPWCFWKCYJ  
FFJEYVKCQSGHYD QPAAMKBZAACIBZLHJQD

this does not resemble English!

### 2. First-order approximation

Suppose now we use symbols that are independent but with frequencies of English text we then get,

OCRO HLI RGWR NMIELWIS EU LL NBNSEBYA TH  
EEI ALHENHTTPA OOBTTVA  
NAH BRL

this is really not that much better!

let's think differently

- **digram**- two successive letters (especially two letters used to represent a single sound: 'sh' in 'shoe')
- there are lots of these in the English language
- maybe we could take these into account

er in ti on te al an at ic en is re ra le ri ro st ne ar

how do we take digrams into account?

to do this we can treat digrams as single symbols and give them probabilities according to their frequency of occurrence

### 3. Second-order approximation

Suppose we now take digrams into account in the model - we now get the following

ON IE ANTSOUTINYS ARE T INCTORE ST  
BE S DEAMY ACHIN D ILONASIVE  
TUCCOWEAT TEASONARE FUSO TIZIN ANDY  
TOBE SEACE CTISBE

so maybe a little progress??

### 4. Third-order Approximation

here we take trigrams into account

IN NO IST LAT WHEY CRATICT FROURE  
BIRS GROCID PONDENOME OF DEMONSTURES  
OF THE REPTAGIN IS REGOACTIONA OF  
CRE.

think again!

we could go on taking bigger and bigger combinations of letters into account but perhaps working on a word level might be helpful!

## 5. First-order approximation

We now go back to a first order approximation but on a word level - words are chosen independently based on their probability of occurrence.

REPRESENTING AND SPEEDILY IS AN GOOD  
APT OR COME CAN DIFFERENT NATURAL  
HERE HE THE A IN CAME THE TOOF TO  
EXPERT GRAY COME TO FURNISHES THE  
LINE MESSAGE HAD BE THESE.

## 6. Second-order Approximation

The word transition probabilities are correct but no further structure is included

THE HEAD AND IN FRONTAL ATTACK ON AN  
ENGLISH WRITER THAT THE CHARACTER OF  
THIS POINT IS THEREFORE ANOTHER  
METHOD FOR THE LETTERS THAT THE TIME  
OF WHO EVER TOLD THE PROBLEM FOR AN  
UNEXPECTED.

## How Shannon did all this ...

*"To construct (3) for example, one opens a book at random and selects a letter at random on the page. This letter is recorded. The book is then opened to another page and one reads until this letter is encountered. The succeeding letter is then recorded. Turning to another page this second letter is searched for and the succeeding letter recorded, etc. A similar process was used for (4), (5) and (6). It would be interesting if further approximations could be constructed, but the labor involved becomes enormous at the next stage."*

so ...

It appears then that a sufficiently complex stochastic process will give a satisfactory representation of a discrete source

the source - thing the produces  
the random variable  $X$

- We can think of a **discrete source** as generating the message, symbol by symbol.
- It chooses successive symbols according to certain probabilities
- The choice depends, in general, on preceding choices as well as the particular symbols in question

- therefore the source is now no longer one that can be said to be memoryless - it has memory - what happened last matters [i.e. certain letters are more likely to follow certain letters and likewise for words]
- typically in the work we will do later we are interested in these kinds of discrete sources as well as they map very well to reality
- in order to deal with such sources we need to introduce some new concepts around entropy

entropy      conditional entropy      mutual information  
 joint entropy      relative entropy

## joint entropy

### joint entropy

**Definition 1.3** (joint entropy) The *joint entropy*  $H(X, Y)$  of a pair of discrete random variables  $X$  and  $Y$  with joint distribution  $p(x, y)$  is defined as

$$H(X, Y) = - \sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 p(x, y).$$

how likely it is that two  
(or more) events happen at the same time

## conditional entropy

### intuitive approach - recall from last lecture

- Given a text composed from an alphabet of 32 letters (each letter equally probable)
- Person A chooses a letter  $X$  (randomly)
- Person B wants to know this letter
- B may ask only binary questions
- Question: how many binary questions must B ask in order to learn which letter  $X$  was chosen by A
- Answer: **entropy  $H(X)$**
- Here:  $H(X) = 5$  bit

### now conditional entropy

- The sky is blu\_
- How many binary questions must we ask to get the next letter??
- 5?
- No!
- Why?
- What's wrong?
- The context tells us "something" about the missing letter Y
- $H(Y|X)$  is this conditional entropy - in other there is less uncertainty now we know X

$$\begin{aligned}
 H(Y|X) &= \sum_{x \in \mathcal{X}} p(x) H(Y|X = x) \\
 &= - \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log_2 p(y|x) \\
 &= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log_2 p(y|x).
 \end{aligned}$$

the entropy of Y given X

- $H(Y|X) \leq H(Y)$
- In worst case - if we ignore X - we need H(Y) binary questions to determine Y
- Knowledge of Y cannot increase the number of binary questions
- Knowledge can never harm!

### the chain rule

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

joint entropy of random variables X and Y =  
entropy of X + conditional entropy of Y given X

### proof of chain rule

$$H(X, Y) = - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y) \quad (2.15)$$

$$= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x) p(y|x) \quad (2.16)$$

$$= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x) - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x) \quad (2.17)$$

$$= - \sum_{x \in \mathcal{X}} p(x) \log p(x) - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x) \quad (2.18)$$

$$= H(X) + H(Y|X). \quad (2.19)$$

### example from Cover

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

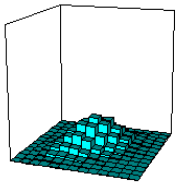
$Y \backslash X$	1	2	3	4
1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$
2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$
3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
4	$\frac{1}{4}$	0	0	0

The marginal distribution of X is  $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8})$  and the marginal distribution of Y is  $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ , and hence  $H(X) = \frac{7}{4}$  bits and  $H(Y) = 2$  bits. Also,

aside

### marginal distribution

- Given two jointly distributed random variables  $X$  and  $Y$ , the marginal distribution of  $X$  is simply the probability distribution of  $X$  ignoring information about  $Y$ ,
- typically calculated by summing or integrating the joint probability distribution over  $Y$ .



- Suppose a joint distribution of two random variables.
- If you integrate out one variable, you can obtain a **marginal distribution** of the other variable.

<http://homepage2.nifty.com/hashimoto-t/misc/margin-e.html>

$$H(X|Y) = \sum_{i=1}^4 p(Y=i)H(X|Y=i) \quad (2.22)$$

$$= \frac{1}{4}H\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right) + \frac{1}{4}H\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{8}, \frac{1}{8}\right) + \frac{1}{4}H\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) + \frac{1}{4}H(1, 0, 0, 0) \quad (2.23)$$

$$= \frac{1}{4} \times \frac{7}{4} + \frac{1}{4} \times \frac{7}{4} + \frac{1}{4} \times 2 + \frac{1}{4} \times 0 \quad (2.24)$$

$$= \frac{11}{8} \text{ bits.} \quad (2.25)$$

Similarly,  $H(Y|X) = \frac{13}{8}$  bits and  $H(X, Y) = \frac{27}{8}$  bits.

# mutual information

### mutual information

This can be defined as follows:

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

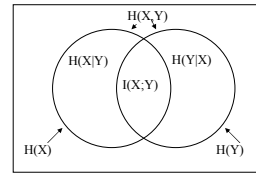
Intuitively, the mutual information " $I(X; Y)$ " measures **the information about  $X$  that is shared by  $Y$**

## mutual inofrmation

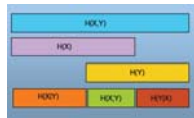
In classical information theory, the mutual information of two random variables is a quantity that measures the mutual dependence of the two variables.

A discrete memoryless source will produced independent variables and hence there is zero dependence between two successive variables.

## Mutual information and entropy



- $I(X;Y)$  is 0 iff two variables are independent
- For two dependent variables, mutual information grows not only with the degree of dependence, but also according to the entropy of the variables

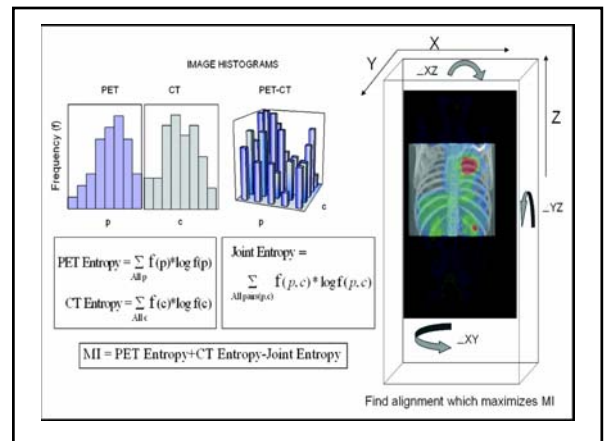


## observation

- From the diagram on the previous page it is possible to see that joint entropy is smallest and, consequently, mutual information is largest when variables are closely aligned - i.e. when the two circles of the Venn diagram overlap

## the example

- A **PET scan** measures important body functions, such as blood flow, oxygen use, and sugar (glucose) metabolism, to help doctors evaluate how well organs and tissues are functioning.
- **CT imaging** uses special x-ray equipment, and in some cases a contrast material, to produce multiple images or pictures of the inside of the body.



so what was the point?

- we are not just interested in discrete memoryless sources - we are interested in sources in which the current symbol has some dependence on what went previously
- hence we need to be able to talk about the information or uncertainty associated with a random variable that has some dependence on a previous random variable

• But now that we know how to talk about information, how do we put that in to use???

• And we still have not answered those two questions.

• Tune in next day!

short film about information theory

Ray and Charles Eames - a most amazing film made in 1953 - it is unbelievably relevant