Signal & Data Analysis in Neuroscience 2016

Information Theory

Izhar Bar-Gad
Room: 408 Phone: 7141 Email: izhar.bar-gad@biu.ac.il

Outline

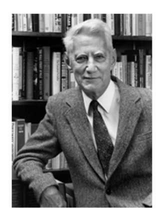
- Entropy
- Mutual information
- Continuous variables

Suggested reading:

- *Elements of Information Theory*, T. Cover & J. Thomas, Ch. 2.
- *Information Theory, Inference, and Learning Algorithms*, David J.C. MacKay, Ch. 2 (Online version is available on the course web site).

Introduction

- Information theory is a branch of mathematics founded by **Claude Shannon** in the 1940s.
- Information theory sets up **quantitative measures** of information and of the capacity of various systems to transmit, store, and otherwise process information.
- Usage: communication, compression, cryptography, computer science, biology, psychology, neuroscience, etc.



Entropy

- The **entropy** of a system is the amount of ***uncertainty*** about the state of that system.
- The entropy is measured by the number of bits required to fully describe the state of the system.
- Other symbols may easily be transformed to bits e.g. English letters may be represented by 5 bits.
- Could also be thought of as the number of yes/no questions required to establish full understanding.

This type of entropy is also termed Shanon's entropy or Information entropy to distinguish it from the entropy used in Thermodynamics

IBG

Simple example: coin flipping I

- A coin flip results in either heads or tails. We can mark the outcomes using 1 bit:

  Head = 0 **T**ail = 1

- Following this encoding scheme, the following sequences of coin flips are equivalent:

  H,H,T,H,T ←→ 00101

- Exactly 1 bit is required to represent each toss.

IBG

Simple example: coin flipping II

- Assuming that we flip two coins simultaneously, we can encode the outcomes as:

Coin A	H	H	T	T
Coin B	H	T	H	T
Encoding	00	01	10	11

- Following this encoding scheme the following sequences of coin flips are equivalent:

00101110 ←→

Trial	1	2	3	4
Coin A	H	T	T	T
Coin B	H	H	T	H

- Exactly 2 bits are required to represent each toss.

IBG

Simple example: coin flipping III

- What happens if we don't care about the order? We only care if we got both heads, both tails, or a mixed pair.
- The probability of each of these outcomes:
  - both heads - 25%
  - both tails - 25%
  - mixed - 50%
- We will use the following encoding scheme:
  - mixed - 0
  - both heads - 10
  - both tails - 11

Simple example: coin flipping IV

- Following this encoding scheme the following sequences of coin flips may be encoded as:

100110 ←

Trial	1	2	3	4
Coin A	H	T	T	T
Coin B	H	H	T	H

- The average number of bits we use:

Both heads: 0.25 x 2 bits = 0.5 bits
Both tails: 0.25 x 2 bits = 0.5 bits
Mixes: 0.5 x 1 bit = 0.5 bits
1.5 bits

Entropy & Information

- The **entropy** of a system is the ***uncertainty*** about the state of that system. It is the expected number of bits required to fully describe the state of the system.
  - In the final two-coin-flip example, we had a 1.5 bit uncertainty about the outcome.
- **Information** is, quite simply, the amount our uncertainty is reduced given ***new knowledge***.
  - In the two-coin-flip example, if we got new knowledge that the two coins flipped were the same, we will gain 0.5 bits of information (as there is only 1 bit of uncertainty left).

Entropy

- Entropy is the expected length in bits of a binary message conveying information
- Other common descriptions of the term: code complexity, uncertainty, missing/required information, expected surprise, information content, etc.
- Historically, entropy was defined in classic thermodynamics as the “amount of un-usable heat in system” and in statistical thermo-dynamics as the “measure of the disorder in the system”, the two were proven to be equivalent.

Shannon Information

- Smallest unit of information is the “bit”
- 1 bit = the amount of information needed to choose between two equally-likely outcomes (e.g. flip a coin)
- Properties:
  - Information for independent events adds
  - Information is zero if we already know the outcome

Shannon Information: Surprise I

The surprise of a single event is high for unexpected (low probability) events and low for expected events.

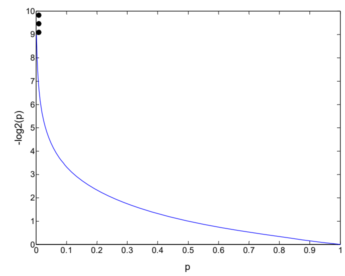
$$p(r_1) = 1 \quad \Rightarrow \quad h(p(r_1)) = 0$$
$$p(r_2) \to 0 \quad \Rightarrow \quad h(p(r_2)) \to \infty$$

Independent events: $p(r_1, r_2) = p(r_1)\, p(r_2)$

Implies: $h(p(r_1, r_2)) = h(p(r_1)) + h(p(r_2))$

$$\boxed{h(p(r)) = -\log_2(p(r))}$$

Shannon Information: Surprise II



IBG

Logarithms – useful formulas

$$\log_a X \cdot Y = \log_a X + \log_a Y$$

$$\log_a \frac{X}{Y} = \log_a X - \log_a Y$$

$$\log_a X^Y = Y \log_a X$$

$$\log_a X = \frac{\log_b X}{\log_b a}$$

$$\frac{d \log_a X}{dX} = \frac{\log_a e}{X}$$

IBG

Entropy - definition

- Entropy is the mean value of the information over all possible observations

$$H(X) = E_p[-\log_2 p(x)]$$

- In the discrete case:

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

IBG

Example: a two outcome event I

- The entropy of the result of a fair coin toss:

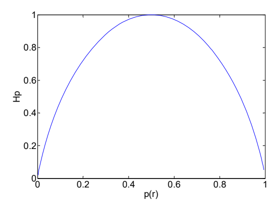
$$H = -[0.5 \cdot \log_2(0.5) + (1-0.5) \cdot \log_2(1-0.5)]$$
$$= -[-0.5 - 0.5] = 1$$

- The entropy of an unfair (99% head) coin toss:

$$H = -[0.99 \cdot \log_2(0.99) + (1-0.99) \cdot \log_2(1-0.99)]$$
$$= -[-0.0144 - 0.0644] = 0.08$$

Example: a two outcome event II

- In the general case:



$$H = -[p \cdot \log_2(p) + (1-p) \cdot \log_2(1-p)]$$

Entropy properties

- Entropy is always positive
- Entropy is maximum if p(r) is constant
  - Least certain of the result
- Entropy is minimum if p(r) is a delta function
- The higher the entropy, the more you learn (on average) by observing values of the random variable
- The higher the entropy, the less you can predict the values of the random variable

Calculating Entropy: The simple case

- If all n possible outcomes of situation X are equally probable, then our uncertainty about which one will occur can be calculated by:

$$H(X) = \log_2(n) \text{ bits}$$

- Out of gold eight coins, one of which is a fake, while you know the other seven are real. You know the fake one has a different weight than the rest. How many weightings on a balance scale will it take to determine the fake? What if you only had seven coins with one fake? What if you had nine coins with one fake?

IBG

Encoding based on entropy I

- Suppose we have 4 symbols: A C G T with
- The symbol probabilities are:
  $P_a=0.5$ $P_C=0.25$ $P_g=P_t=0.125$
- Leading to surprises:
  $h_a=1$bit $h_c=2$bit $h_g=h_T=3$ bit
- Thus the mean uncertainty of a symbol is:
  H=1*0.5+2*0.25+0.125*3+0.125*3=1.75 bit

IBG

Encoding based on entropy II

- One option for encoding uses 2 bits for each symbol: A=00 C=01 G=10 T=11
- In the other option the number of binary digits equals the surprise: A=1 C=01 G=000 T=001
- So the string ***ACATGAAC*** which has frequencies the same as the probabilities defined above, is coded as:

Method 1	0001001110000001	16 (2 bits per symbol)
Method 2	10110010001101	14 (1.75 bits per symbol)

IBG

Encoding based on entropy III

- In this specific case, can we find a better (shorter) encoding ?
- In the general case, how can we formulate the optimal encoding ?
- These questions are handled under the data compression topic…

Elements of Information Theory, T. Cover & J. Thomas, Chapter 5.

IBG

Outline

- Entropy
- Mutual information
- Continuous variables

IBG

Joint entropy

- The joint entropy may be considered a single vector valued random variable:

$$H(X,Y) = E_{p(x,y)}[-\log_2 p(x,y)]$$

- In the discrete case:

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

IBG

Conditional entropy

Same formulation, but using the conditional density:

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

IBG

The conditional entropy chain rule

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

Proof:

$$
\begin{aligned}
H(Y|X) &= -E_{p(x,y)} \log\ p(y|x) \\
&= -E_{p(x,y)} \log \left(\frac{p(y,x)}{p(x)} \right) \\
&= -E_{p(x,y)} (\log p(y,x) - \log p(x)) \\
&= -E_{p(x,y)} \log p(y,x) + E_{p(x)} \log p(x) \\
&= H(Y,X) - H(X).
\end{aligned}
$$

Thus:

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

IBG

Mutual information I

- The entropy tells us how much we can learn (therefore how much we don't know)
- The mutual information between *r* and *s* is:
  - How much do we learn about *r* by observing *s?*
  - How much more do we know about *r* after observing *s*?
  - How much easier is it to predict *r* after observing *s*?
- Therefore: How much has the entropy of *r* decreased after observing *s*?

IBG

Mutual information II

- Mutual information = How is the entropy of *r* decreased by knowing *s*?

$$H_{noise} = H(R \mid S) = -\sum_r \sum_s p(r,s) \cdot \log(p(r \mid s))$$

$$\begin{aligned} I(R;S) &= H(R) - H(R \mid S) \\ &= -\sum_r p(r) \cdot \log(p(r)) + \sum_r \sum_s p(r,s) \cdot \log(p(r \mid s)) \\ &= -\sum_r \sum_s p(r,s) \cdot \log(p(r)) + \sum_r \sum_s p(r,s) \cdot \log(p(r \mid s)) \\ &= \sum_r \sum_s p(r,s) \cdot [-\log(p(r)) + \log(p(r \mid s))] \\ &= \sum_r \sum_s p(r,s) \cdot \log(\frac{p(r \mid s)}{p(r)}) = \sum_r \sum_s p(r,s) \cdot \log(\frac{p(r,s)}{p(r) \cdot p(s)}) \end{aligned}$$

The doctor example I

- We're back to the doctor who need to distinguish between:
  - The flu $p(x_1)$=0.9
  - Severe infection $p(x_2)$ =0.1
- He has two tests:

Blood test Y	Flu	Infection
Positive	0.2	0.7
Negative	0.8	0.3

Urine test Z	Flu	Infection
Positive	0.1	0.5
Negative	0.9	0.5

- **Which test gives more information about the state of the patient?**

The doctor example II

$$I_m = \sum_{s,r} P[s]\, P[r \mid s] \log_2 \left(\frac{P[r \mid s]}{P[r]} \right)$$

$P(y_+)$=0.9*0.2+0.1*0.7=0.25 $P(y_-)$=0.75
$P(z_+)$=0.9*0.1+0.1*0.5=0.14 $P(z_-)$=0.86

H(X)=-(0.9*$\log_2$(0.9)+0.1*$\log_2$(0.1))=0.436

I(Y;X)=0.9*0.2*$\log_2$(0.2/0.25)+0.9*0.8*$\log_2$(0.8/0.75)+ 0.1*0.7*$\log_2$(0.7/0.25)+0.1*0.3*$\log_2$(0.3/0.75)=0.0734

I(Z;X)=0.9*0.1*log2(0.1/0.14)+0.9*0.9*$\log_2$(0.9/0.86)+ 0.1*0.5*$\log_2$(0.5/0.14)+0.1*0.5*$\log_2$(0.5/0.86)=0.0621

Thus, the blood test is more informative…

Properties of mutual information I

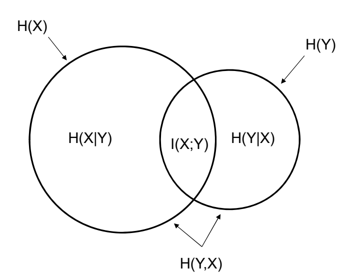
- Zero if r and s are independent
  $p(r,s) = p(r)p(s) \Rightarrow I(R,S) = 0$
- Cannot be more than the entropy
  $I(R,S) \le H(R) \quad I(R,S) \le H(S)$
- Cannot be increased by math alone
  $I(f(R),S) \le I(R,S)$

This is critical: holds true FOR ANY f(), so no transmission line, neural network, or laboratory computation (no matter how clever) can ever squeeze out more information.

Properties of mutual information II

- I(X;Y)=H(X)-H(X|Y)
- I(X;Y)=H(Y)-H(Y|X)
- I(X;Y)=H(X)+H(Y)-H(Y,X)
- I(X;Y)= I(Y;X)
- I(X;X)=H(X)

Entropy and Mutual information



Relative entropy ≡ Kullback Liebler (KL) divergence

The Kullback-Leibler (KL) divergence is a 'distance' measure between probability distributions.

$$D_{KL}(p,q) = \sum_r p(r) \log_2 \frac{p(r)}{q(r)}$$

$D_{KL}(p,q) \neq D_{KL}(q,p)$, and $D_{KL} \geq 0$

Thus,

$$I_m = D_{KL}(p(r,s), p(r)p(s))$$

• The excess message length needed to use p(x) - optimized code for messages based on q(x)

Relative entropy properties

$$I_m = \sum_{s,r} P[r,s] \log_2 \left(\frac{P[r,s]}{P[r]P[s]} \right)$$

$$D_{KL}(p,q) = \sum_r p(r) \log_2 \frac{p(r)}{q(r)}$$

$$D_{KL}(p(r,s), p(r)p(s)) = \sum_{r,s} p(r,s) \log_2 \frac{p(r,s)}{p(r)p(s)}$$

$$D_{KL}(p(r)p(s), p(r,s)) = \sum_{r,s} p(r)p(s) \log_2 \frac{p(r)p(s)}{p(r,s)}$$

$$\downarrow$$

$$I_m = D_{KL}(p(r,s), p(r)p(s))$$
$$I_m = D_{KL}(p(s,r), p(s)p(r))$$
$$I_m \neq D_{KL}(p(r)p(s), p(r,s))$$

Additional (in) equalities

- D(p||q)≥0 (information inequality)
  D(p||q)=0 iff p(x)=q(x) for every x
- I(X;Y)≥0 (Non negativity of mutual information)
  I(X;Y)=0 iff Y & X are independent
- H(X|Y)≤H(X) (Conditioning reduces entropy)
- $H(X_1, X_2, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i)$ (Independence bound)

Mostly proved by: If *f* is convex → *Ef(X)≥f(EX)* (Jensen inequality)

Outline

- Entropy
- Mutual information
- Continuous variables

Elements of Information Theory, T. Cover & J. Thomas, Chapter 9.

Continuous variables

- A real number has an infinite number of bits, therefore theoretically, **infinite information**.
- However, there is always noise (or quantization) which defines a number of discriminable levels

Entropy & Differential entropy

- Usage of probability density instead of probability

$$\begin{aligned} H &= -\sum p[r]\Delta r \log_2(p[r]\Delta r) \\ &= -\sum p[r]\Delta r \log_2 p[r] - \log_2 \Delta r \end{aligned}$$

- Note: for $\Delta r \rightarrow 0$ the log diverges…

$$h(r) = \lim_{\Delta r}\{H(r) + \log_2 \Delta r\} = -\int p(r)\log_2 p(r)dr$$

Differential entropy

$$h(x) = -\int p(x)\log_2 p(x)dx$$

- Example 1: **Uniform distribution** (interval [0,a])

$$h(x) = -\int_0^a \frac{1}{a}\log_2 \frac{1}{a}dx = -\log_2 \frac{1}{a} = \log_2 a$$

Note: for a<1 the differential entropy is negative

- Example 2: **Normal distribution** (μ=0,σ)

$$h(x) = \frac{-1}{\sigma\sqrt{2\pi}}\int e^{-x^2/2\sigma^2} \log\left(\frac{1}{\sigma\sqrt{2\pi}}e^{-x^2/2\sigma^2}\right)dx = \frac{1}{2}\log_2(2\pi e\sigma^2)$$

IBG

Entropy of a sampled continuous variable

- Following a *n* bit quantization of the variable (i.e. accuracy of 2^{-n})
  $H(X)=h(X)-\log(2^{-n})=h(X)+n$
- Example: a uniform distribution over the interval [0,1] with a resolution of ~0.001
  $H(X)=\log_2(1)+\log_2(1000)\sim 10$
- Example: a uniform distribution over the interval [0,¼] with a resolution of ~0.001
  $H(X)=\log_2(¼)+\log_2(1000)\sim 8$
  Since the first two bits are always 0.

IBG

Neurophysiological based information theoretic questions

- How much information do the neurons **convey**?
- How much information is conveyed through a **spike**?
- How much does spiking activity tell us about a stimulus?
- Is the neural representation **optimal**?
- Is the information encoded by a neuronal population **redundant**?
- Can **rate** by itself encode all the information?
- Is there and if so, what is the **theoretical limit** on the information in the nervous system?

These hard questions will be addressed only in the next lesson…

IBG