Information Theory

Mark van Rossum
School of Informatics, University of Edinburgh
January 15, 2018

Why information theory

Understanding the neural code.
- Encoding and decoding. We imposed coding schemes, such as 2nd-order kernel, or NLP. We possibly lost information in doing so.
- Instead, use information:
  - Don’t need to impose encoding or decoding scheme (non-parametric).
    In particular important for 1) spike timing codes, 2) higher areas.
  - Estimate how much information is coded in certain signal.

Caveats:
- No easy decoding scheme for organism (upper bound only)
- Requires more data and biases are tricky

Overview
- Entropy, Mutual Information
- Entropy Maximization for a Single Neuron
- Maximizing Mutual Information
- Estimating information
- Reading: Dayan and Abbott ch 4, Rieke

Definition

The entropy of a quantity is defined as
\[ H(X) = -\sum_x P(x) \log_2 P(x) \]

This is not ‘derived’, nor fully unique, but it fulfills these criteria:
- Continuous
- If \( p_i = \frac{1}{n} \), it increases monotonically with \( n \). \( H = \log_2 n \).
- Parallel independent channels add.

"Unit": bits
Entropy can be thought of as physical entropy, “richness” of distribution [?, ?, ?]
Entropy

Discrete variable

\[ H(R) = - \sum_r p(r) \log_2 p(r) \]

Continuous variable at resolution \( \Delta r \)

\[ H(R) = - \sum_r p(r) \Delta r \log_2 (p(r) \Delta r) = - \sum_r p(r) \Delta r \log_2 p(r) - \log_2 \Delta r \]

letting \( \Delta r \to 0 \) we have

\[ \lim_{\Delta r \to 0} [H + \log_2 \Delta r] = - \int p(r) \log_2 p(r) dr \]

(also called differential entropy)

Joint, Conditional entropy

Joint entropy:

\[ H(S, R) = - \sum_{r,s} p(s, r) \log_2 p(s, r) \]

Conditional entropy:

\[ H(S|R) = \sum_r p(r) H(S|R = r) = - \sum_r p(r) \sum_s p(s|r) \log_2 p(s|r) = H(S, R) - H(R) \]

If \( S, R \) are independent

\[ H(S, R) = H(S) + H(R) \]

Mutual information

Mutual information:

\[ I_m(R; S) = \sum_{r,s} p(r, s) \log_2 \frac{p(r, s)}{p(r)p(s)} \]

= \( H(R) - H(R|S) = H(S) - H(S|R) \)

- Measures reduction in uncertainty of \( R \) by knowing \( S \) (or vice versa)
- \( I_m(R; S) \geq 0 \)
- The continuous version is the difference of two entropies, the \( \Delta r \) divergence cancels

The joint histogram determines mutual information.

Given \( P(r, s) \Rightarrow I_m \).
Mutual Information: Examples

<table>
<thead>
<tr>
<th>Y₁</th>
<th>non smoker</th>
<th>smoker</th>
</tr>
</thead>
<tbody>
<tr>
<td>lung cancer</td>
<td>1/3</td>
<td>0</td>
</tr>
<tr>
<td>no lung cancer</td>
<td>0</td>
<td>2/3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Y₁</th>
<th>non smoker</th>
<th>smoker</th>
</tr>
</thead>
<tbody>
<tr>
<td>lung cancer</td>
<td>1/9</td>
<td>2/9</td>
</tr>
<tr>
<td>no lung cancer</td>
<td>0</td>
<td>2/3</td>
</tr>
</tbody>
</table>

Only for the left joint probability $I_m > 0$ (how much?). On the right, knowledge about $Y_1$ does not inform us about $Y_2$.

Kullback-Leibler divergence

- KL-divergence measures distance between two probability distributions
- $D_{KL}(P || Q) = \int P(x) \log_2 \frac{P(x)}{Q(x)} dx$, or $D_{KL}(P || Q) \equiv \sum_i P_i \log_2 \frac{P_i}{Q_i}$
- Not symmetric, but can be symmetrized
- $I_m(R; S) = D_{KL}(p(r, s) || p(r)p(s))$.
- Often used as probabilistic cost function: $D_{KL}(data || model)$.
- Other probability distances exist (e.g. earth-movers distance)

\[
I(Y_1; Y_2) = \int \int P(y_1, y_2) \log_2 \frac{P(y_1, y_2)}{P(y_1)P(y_2)} \, dy_1 \, dy_2 = -\frac{1}{2} \log_2(1 - \rho^2)
\]

$\rho$ is (Pearson-r) correlation coefficient.
Populations of Neurons

Given
\[ H(R) = - \int p(r) \log_2 p(r) dr - N \log_2 \Delta r \]
and
\[ H(R_i) = - \int p(r_i) \log_2 p(r_i) dr - \log_2 \Delta r \]
We have
\[ H(R) \leq \sum_i H(R_i) \]
(proof, consider KL divergence)

Mutual information in populations of Neurons

Redundancy can be defined as (compare to above)
\[ R = \sum_{i=1}^{n_r} I(r_i; s) - I(r; s) \]
Some codes have \( R > 0 \) (redundant code), others \( R < 0 \) (synergistic)
Example of synergistic code: \( P(r_1, r_2, s) \) with
\[ P(0,0,1) = P(0,1,0) = P(1,0,0) = P(1,1,1) = \frac{1}{4} \]

Entropy Maximization for a Single Neuron

\[ I_m(R; S) = H(R) - H(R|S) \]

Let \( r = f(s) \) and \( s \sim p(s) \). Which \( f \) (assumed monotonic) maximizes \( H(R) \) using max firing rate constraint? Require:
\[ P(r) = \frac{1}{r_{\text{max}}} \]
\[ p(s) = p(r) \frac{dr}{ds} = \frac{1}{r_{\text{max}}} \frac{df}{ds} \]
Thus \( \frac{df}{ds} = r_{\text{max}}p(s) \) and
\[ f(s) = r_{\text{max}} \int_{s_{\min}}^{s} p(s') ds' \]
This strategy is known as **histogram equalization** in signal processing.
Fly retina
Evidence that the large monopolar cell in the fly visual system carries
out histogram equalization

Contrast response for fly large monopolar cell (points) matches
environment statistics (line) [?] (but changes in high noise conditions)

Information of time varying signals
Single analog channel with Gaussian signal \( s \) and Gaussian noise \( \eta \):
\[
r = s + \eta
\]
\[
I = \frac{1}{2} \log_2(1 + \frac{\sigma_s^2}{\sigma_\eta^2}) = \frac{1}{2} \log_2(1 + \text{SNR})
\]
For time dependent signals \( I = \frac{1}{2} T \int \frac{d\omega}{2\pi} \log_2(1 + \frac{s(\omega)}{\eta(\omega)}) \)
To maximize information, when variance of the signal is constrained,
use all frequency bands such that signal+noise = constant.
Whitening. Water filling analog:

Information of graded synapses

V1 contrast responses

Similar in V1, but On and Off channels [?]
Spiking neurons: maximal information

Spike train with $N = T/\delta t$ bins \(\delta t\) “time-resolution”.

$pN = N_1$ events, \#words = \(N_1/(N-N_1)\)!

Maximal entropy if all words are equally likely.

\[
H = -\sum p_i \log_2 p_i = \log_2 N! - \log_2 N_1! - \log_2 (N-N_1)!
\]

Use for large $x$ that $\log x! \approx x \log x - x$.

\[
H = -\frac{T}{\delta t} [p \log_2 p + (1-p) \log_2(1-p)]
\]

For low rates $p \ll 1$, setting $\lambda = (\delta t)p$:

\[
H = T \lambda \log_2 \left(\frac{e^{\lambda \delta t}}{\lambda \delta t}\right)
\]

Calculation incorrect when multiple spikes per bin. Instead, for large
bins maximal information for exponential distribution:

\[
P(n) = \frac{1}{Z} \exp[-n \log(1 + 1/\langle n \rangle)]
\]

\[
H = \log_2(1 + \langle n \rangle) + \langle n \rangle \log_2(1 + 1/\langle n \rangle) \approx \log_2(1 + \langle n \rangle) + 1
\]

Spiking neurons: rate code

Measure rate in window $T$, during which stimulus is constant.

Periodic neuron can maximally encode \(1 + (f_{\text{max}} - f_{\text{min}})T\) stimuli

\[
H \approx \log_2[1 + (f_{\text{max}} - f_{\text{min}})T]. \text{Note, only } \propto \log(T)
\]

Similar behaviour for Poisson: $H \propto \log(T)$
Spiking neurons: dynamic stimuli

Maximizing Information Transmission: single output

Single linear neuron with post-synaptic noise

\[ \mathbf{v} = \mathbf{w} \cdot \mathbf{u} + \eta \]

where \( \eta \) is an independent noise variable

\[ I_m(\mathbf{u}; \mathbf{v}) = H(\mathbf{v}) - H(\mathbf{v}|\mathbf{u}) \]

- Second term depends only on \( p(\eta) \)
- To maximize \( I_m \) need to maximize \( H(\mathbf{v}) \); sensible constraint is that \( ||\mathbf{w}||^2 = 1 \)
- If \( \mathbf{u} \sim N(\mathbf{0}, \mathbf{Q}) \) and \( \eta \sim N(0, \sigma^2_\eta) \) then \( \mathbf{v} \sim N(0, \mathbf{w}^T \mathbf{Q} \mathbf{w} + \sigma^2_\eta) \)

Infomax

Infomax: maximize information in multiple outputs wrt weights

\[ \mathbf{v} = \mathbf{w} \mathbf{u} + \eta \]

\[ H(\mathbf{v}) = \frac{1}{2} \log \det(\langle \mathbf{v} \mathbf{v}^T \rangle) \]

Example: 2 inputs and 2 outputs. Input is correlated. \( w_{k1}^2 + w_{k2}^2 = 1 \).

At low noise independent coding, at high noise joint coding.
Information estimation requires a lot of data. Most statistical quantities are unbiased (mean, var,...). But both entropy and noise entropy have bias.

Try to fit $1/N$ correction [?]

Common technique for $I_m$: shuffle correction [?]

See also: [?], [?]

Summary

- Information theory provides non parametric framework for coding
- Optimal coding schemes depend strongly on noise assumptions and optimization constraints
- In data analysis biases can be substantial