Information Theory II: mutual information and efficient coding

NEU 314, Fall 2021
Lecture 18

Jonathan Pillow
Entropy

\[ H(x) = - \sum_x p(x) \log p(x) \quad \text{in “bits”} \]

\[ = \mathbb{E}[- \log p(x)] \]

other ways of writing it:

\[ H(x) = \sum_{i=1}^{N} p_i (\log \left( \frac{1}{p_i} \right)) \]

- average number of “yes/no” questions needed to identify x
- average “surprise” from encountering a sample from p(x)
Conditional Entropy

\[ H(x|y) = -\sum_y p(y) \sum_x p(x|y) \log p(x|y) \]

- Averaged over \( p(y) \)
- Entropy of \( x \) given some fixed value of \( y \)
Conditional Entropy

\[ H(x|y) = -\sum_y p(y) \sum_x p(x|y) \log p(x|y) \]

averaged over \( p(y) \)

entropy of \( x \) given some fixed value of \( y \)

\[ = -\sum_{x,y} p(x, y) \log p(x|y) \]

“On average, how uncertain are you about \( x \) if you know \( y \)?”

“On average, how many questions do you need to identify \( x \) when you know \( y \)?”
Exercise

Compute the conditional entropy:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>(0)</td>
<td>(0)</td>
<td>(0)</td>
<td>(1/2)</td>
<td>(1/4)</td>
<td>(0)</td>
<td>(0)</td>
<td>(0)</td>
</tr>
<tr>
<td></td>
<td>(0)</td>
<td>(0)</td>
<td>(0)</td>
<td>(0)</td>
<td>(0)</td>
<td>(0)</td>
<td>(0)</td>
<td>(1)</td>
</tr>
<tr>
<td>Y</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p(Y)</td>
<td>(2/3)</td>
<td>(1/3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[H(p(X|Y=0)) = 3/2\]

\[H(p(X|Y=1)) = 0\]

\[H(X \, | \, Y) = \frac{2}{3} \left(\frac{3}{2}\right) + \frac{1}{3} \left(0\right) = 1 \text{ bit}\]

“On average, you need 1 question to guess X when you know Y”
Mutual Information

\[ I(x, y) = H(x) - H(x|y) \]

Total entropy in X minus conditional entropy of X given Y

\[ = H(y) - H(y|x) \]

Total entropy in Y minus conditional entropy of Y given X

\[ = H(x) + H(y) - H(x, y) \]

Sum of entropies minus joint entropy

“How much does X tell me about Y (or vice versa)?”

“How much is your uncertainty about X reduced from knowing Y?”

“What is the difference between (# of questions needed to guess X) and (# questions needed to guess X when you’re given Y)”
Venn diagram of entropy and information
Kullback-Leibler Divergence

for two distributions $P(x)$ and $Q(x)$

$$D_{KL}(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

$$= \sum x \ P(x) \log P(x) - \sum x \ P(x) \log Q(x)$$

$$= \sum x \ P(x)(-\log Q(x)) - \sum x \ P(x)(-\log P(x))$$

- quantifies the number of extra bits required to code samples from $P(x)$ if you use a codebook (“question asking strategy”) based on $Q(x)$

Properties:

- $D_{KL}(P||Q) \geq 0, \ \forall P, Q$
- $D_{KL}(P||Q) = 0, \ \text{iff} \ P = Q$
- $\text{KL is not in general symmetric: } D_{KL}(P||Q) \neq D_{KL}(Q||P)$
Illustrating non-symmetry of KL divergence

\[ D_{KL}(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)} \]

1st probability distribution \( P_1(X) \)

\[
\begin{array}{cccccccc}
1/8 & 1/8 & 1/8 & 1/8 & 1/8 & 1/8 & 1/8 & 1/8 \\
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
\end{array}
\]

2nd probability distribution: \( P_2(X) \)

\[
\begin{array}{cccccccc}
0 & 0 & 0 & 1/8 & 1/8 & 1/8 & 1/8 & 1/2 \\
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
\end{array}
\]

Exercise: 1) What is KL(\( P_2 || P_1 \))?

2) What is KL(\( P_1 || P_2 \))?
Mutual Information identities

\[ I(x, y) = H(x) - H(x|y) \]

\[ = - \sum_{x} p(x) \log p(x) + \sum_{x,y} p(x, y) \log p(x|y) \]

\[ = - \sum_{x,y} p(x, y) \log p(x) + \sum_{x,y} p(x, y) \log p(x|y) \]

\[ = \sum_{x,y} p(x, y) \log \frac{p(x|y)}{p(x)} \]

\[ = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} = D_{KL}(p(x, y)\|p(x)p(y)) \]

KL divergence between joint distribution and product of marginals

\[ = - \sum_{x,y} p(x, y) \log p(y) + \sum_{x,y} p(x, y) \log p(y|x) \]

\[ = - \sum_{y} p(y) \log p(y) + \sum_{x,y} p(x, y) \log p(y|x) \]

\[ = H(y) - H(y|x) \]
Data Processing Inequality

Suppose $S \rightarrow R_1 \rightarrow R_2$ form a Markov chain, that is

$$P(R_1, R_2|S) = P(R_2|R_1)P(R_1|S)$$

Then necessarily:

$$I(S, R_2) \leq I(S, R_1)$$

• in other words, we can only lose information during processing
Summary with formulas:

“surprise” function: \(- \log[p(x)]\)

Entropy: “avg # Y/N Q’s” = \(- \sum_x P(x) \log P(x)\) (in bits)

Conditional Entropy: \(H(x|y) = - \sum P(x, y) \log P(x|y)\)

Mutual information: \(I(x, y) = H(x) - H(x|y) = H(y) - H(y|x)\)
\(= KL[p(x, y)\|p(x)p(y)]\)

KL divergence \(D_{KL}(P\|Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}\)
Barlow’s
“Efficient Coding Hypothesis”
Efficient Coding Hypothesis:

- goal of nervous system: maximize information about environment (one of the core “big ideas” in theoretical neuroscience)

**mutual information:**

\[ I(x, y) = H(y) - H(y|x) \]

- avg # yes/no questions you can answer about x given y (“bits”)
Barlow’s original version:

\[ I(x, y) = H(y) - H(y|x) \]

if responses are noiseless

**mutual information:**

- response entropy
- “noise” entropy

Barlow 1961
Atick & Redlich 1990
Barlow’s original version:

mutual information:

\[ I(x, y) = H(y) - H(y|x) \]

- response entropy
- “noise” entropy

noiseless system

\[ \Rightarrow \text{brain should maximize response entropy} \]
- use full dynamic range
- decorrelate (“reduce redundancy”)

- mega impact: huge number of theory and experimental papers focused on decorrelation / information-maximizing codes in the brain
basic intuition

natural image

nearby pixels exhibit strong dependencies

pixels

desired encoding

neural representation

pixel \(i\)

pixel \(i+1\)

neural response \(i\)

neural response \(i+1\)
Application Example: single neuron encoding stimuli from a distribution $P(x)$

- **stimulus prior**
  
  $x \sim P(x)$

- **noiseless, discrete encoding**
  
  $y = f(x), \quad y \in \{y_1, y_2, \ldots, y_n\}$

Q: what solution for infomax?
Application Example: single neuron encoding stimuli from a distribution $P(x)$

stimulus prior \[ x \sim P(x) \]

noiseless, discrete encoding \[ y = f(x), \quad y \in \{y_1, y_2, \ldots, y_n\} \]

Q: what solution for infomax?
A: histogram-equalization

\[ I(X, Y) = H(Y) - H(Y|X) \]
Laughlin 1981: blowfly light response

- first major validation of Barlow’s theory
luminance-dependent receptive fields

(a) $S/N=10$

High SNR
(“whitening” / decorrelating)

(b) $S/N=2$

Middle SNR
(partial decorrelating)

(c) $S/N=0.1$

Low SNR
(averaging / correlating)
summary: info theory

- entropy
- conditional entropy
- mutual information
- data processing inequality
- efficient coding hypothesis (Barlow)
  - neurons should “maximize their dynamic range”
  - multiple neurons: responses should decorrelate
  - Atick & Redlich: extended to noisy responses