Information Theory

A mathematical theory of communication, Claude Shannon 1948

- Entropy
- Conditional Entropy
- Mutual Information
- KL Divergence
- Data Processing Inequality
- Efficient Coding Hypothesis (Barlow 1961)
- Maxent distributions
Entropy

# yes/no questions needed, on average, to determine the value of a random variable
**motivating example #1:** I’m thinking of a number between 1 and 8. How many Y/N questions do you need to guess it?

1 2 3 4 5 6 7 8

**Strategy 1:**

Q1: is it 1?
Q2: is it 2?
...  
Q2: is it 8?

- worst case: 8 questions
- average case (assuming uniform): 4 questions

Can we do better???
motivating example #1: I’m thinking of a # between 1 and 8. How many Y/N questions do you need to guess it?

1  2  3  4  5  6  7  8

Strategy 2: Q1: is it >4?
**motivating example #1:** I’m thinking of a # between 1 and 8. How many Y/N questions do you need to guess it?

<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>Q1: is it $&gt;4$?</td>
<td>N</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Strategy 2: 1,2,3,4 | 5,6,7,8
motivating example #1: I’m thinking of a # between 1 and 8. How many Y/N questions do you need to guess it?

1  2  3  4  5  6  7  8

Strategy 2:  Q1: is it >4?  N  Y

1,2,3,4  5,6,7,8

Q2: is it >2?  N  Y  N  Y  is it >6?

1,2  3,4  5,6  7,8
motivating example #1: I’m thinking of a # between 1 and 8. How many Y/N questions do you need to guess it?

Strategy 2:

Q1: is it >4? N Y

1,2,3,4 5,6,7,8

Q2: is it >2? N Y  N Y is it >6?

1,2 3,4 5,6 7,8

Q3: is it 2? is it 4? is it 6? is it 8?
motivating example #1: I’m thinking of a # between 1 and 8. How many Y/N questions do you need to guess it?

1  2  3  4  5  6  7  8

Strategy 2: Q1: is it >4?  N  Y
        1,2,3,4           5,6,7,8

Q2: is it >2?  N  Y  N  Y  is it >6?
        1,2  3,4  5,6  7,8

Q3: is it 2?  is it 4?  is it 6?  is it 8?
        N  Y  N  Y  N  Y  N  Y

1  2  3  4  5  6  7  8

• average: 3 questions
motivating example #1: I’m thinking of a # between 1 and 8. How many Y/N questions do you need to guess it?

1  2  3  4  5  6  7  8

General result:

\[ 2^N \text{ numbers} \implies N \text{ binary questions} \]

Therefore: # questions = \( \log_2(K) \)

# possibilities (unknowns)
motivating example #2: how many Y/N questions needed?

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

- 3 questions would still suffice
- But can we do better?
motivating example #2: how many Y/N questions needed?

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

Strategy: Q1: is it 8?

N

Y

4, 5, 6, 7

8 (done!)
motivating example #2: how many Y/N questions needed?

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

Strategy:

Q1: is it 8?

- N
- Y

Q2: is it >5?

- N
- Y

4,5,6,7

8 (done!)
motivating example #2: how many Y/N questions needed?

P(X) | 0  | 0  | 0  | 1/8 | 1/8 | 1/8 | 1/8 | 1/8 | 1/2  
-----|----|----|----|-----|-----|-----|-----|-----|------
     | 1  | 2  | 3  | 4   | 5   | 6   | 7   | 8   |      

Strategy:

Q1: is it 8? N Y

Q2: is it >5? N Y

Q3: is it 5? Y

4,5,6,7

is it 7? Y

4 5 6 7
motivating example #2: how many Y/N questions needed?

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

Strategy:

Q1: is it 8?

Q2: is it >5?

Q3: is it 5?

what is the average # of questions?

\[
\text{1/2 (1 question) + 1/2 (3 questions)} = \frac{1}{2} + \frac{3}{2} = 2 \text{ questions on average!}
\]
motivating example #2: how many Y/N questions needed?

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

General results:

• optimal strategy: divide probability in half w/ each question
• need N questions to identify options with probability $1/2^N$
• thus: $\log_2(K)$ questions for options with probability $1/K$
• or: $-\log_2(p)$ questions for options with probability $p$

(using: $\log(K) = -\log(1/K) = -\log(p)$)

code length
P(X) | 0  | 0  | 0  | 1/8 | 1/8 | 1/8 | 1/8 | 1/2  \\
     | 1  | 2  | 3  | 4   | 5   | 6   | 7   | 8    \\

Strategy:

Q1: is it 8?  N  Y

Q2: is it >5? N  Y

Q3: is it 5?  4,5,6,7  8

- or: \(-\log_2(p)\) questions for options with probability \(p\)

**code:**
8: Y
7: NYY
6: NYN
5: NNY
4: NNN

**code length**
Entropy

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

- number of “yes/no” questions needed to identify x (on average)

for distribution on K bins,
- maximum entropy = \(\log K\) (achieved by uniform dist)
- minimum entropy = 0 (achieved by all probability in 1 bin)
log-likelihood and entropy

model: \( P(x|\theta) \)

entropy \( H \): \(-\mathbb{E}[\log P(x|\theta)]\)

How would we compute a Monte Carlo estimate of this?

draw samples: \( x_i \sim P(x|\theta) \) for \( i = 1, \ldots, N \)

compute average: \( \hat{H} = -\frac{1}{N} \sum_{i=1}^{N} \log P(x_i|\theta) \)

• Avg neg log likelihood = Monte Carlo estimate for entropy!
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 \( y \)

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

\[ = H(x) \quad \text{if} \quad P(x,y) = P(x)P(y) \]

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

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

- total entropy in X minus conditional entropy of X given Y

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

- 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?”
Venn diagram of entropy and information

\[ H(x) \quad H(y) \]

\[ H(x|y) \quad I(x, y) \quad H(y|x) \]

\[ H(x, y) \]
Kullback-Leibler Divergence

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

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

• 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:

• $D_{KL}(P||Q) \geq 0, \forall P, Q$

• $D_{KL}(P||Q) = 0$, iff $P = Q$

• KL is not in general symmetric: $D_{KL}(P||Q) \neq D_{KL}(Q||P)$
Illustrating non-symmetry of KL divergence

1\textsuperscript{st} probability distribution \quad 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}

2\textsuperscript{nd} probability distribution: \quad 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
Data Processing Inequality

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

$$P(R_1, R_2 | \mathcal{S}) = P(R_2 | R_1) P(R_1 | \mathcal{S})$$

Then necessarily:

$$I(\mathcal{S}, R_2) \leq I(\mathcal{S}, R_1)$$

• 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)

Differential Entropy: \(- \int P(x) \log P(x) dx\)

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

ML estimator: minimizes \(D_{KL}[p_{true}(x)\|p(x|\theta)]\)

Monte Carlo Integration: \(\int p(x) f(x) dx \approx \frac{1}{N} \sum f(x_i) \quad x_i \sim p(x)\)