Basic Concepts in Information Theory

Spring 2024

Instructor: Shandian Zhe
zhe@cs.utah.edu
School of Computing
Coding theory

- Let us start with discrete random variables
Coding theory

• How to represent the information contained in the random variables?

\[ h(x) \geq 0 \]

\[ h(x, y) = h(x) + h(y) \quad \text{if } x, y \text{ are independent} \]

\[ p(x, y) = p(x)p(y) \]

\[ h(x) = -\log(p(x)) \]
Entropy

- The average amount of information needs to transmit

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

<table>
<thead>
<tr>
<th>$x$</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
<th>g</th>
<th>h</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(x)$</td>
<td>$\frac{1}{2}$</td>
<td>$\frac{1}{4}$</td>
<td>$\frac{1}{8}$</td>
<td>$\frac{1}{16}$</td>
<td>$\frac{1}{64}$</td>
<td>$\frac{1}{64}$</td>
<td>$\frac{1}{64}$</td>
<td>$\frac{1}{64}$</td>
</tr>
</tbody>
</table>

\[
\begin{align*}
H[x] &= -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{4} \log_2 \frac{1}{4} - \frac{1}{8} \log_2 \frac{1}{8} - \frac{1}{16} \log_2 \frac{1}{16} - \frac{4}{64} \log_2 \frac{1}{64} \\
&= 2 \text{ bits}
\end{align*}
\]

Entropy is also the average code length
Entropy reflects uncertainty

Figure 1.30 Histograms of two probability distributions over 30 bins illustrating the higher value of the entropy \( H \) for the broader distribution. The largest entropy would arise from a uniform distribution that would give \( H = -\ln\left(\frac{1}{30}\right) = 3.40 \).

We can extend the definition of entropy to include distributions \( p(x) \) over continuous variables \( x \) as follows. First divide \( x \) into bins of width \( \Delta \). Then, assuming \( p(x) \) is continuous, the mean value theorem (Weisstein, 1999) tells us that, for each such bin, there must exist a value \( x_i \) such that

\[
\int_{i\Delta}^{(i+1)\Delta} p(x) \, dx = p(x_i) \Delta.
\]

We can now quantize the continuous variable \( x \) by assigning any value \( x \) to the value \( x_i \) whenever \( x \) falls in the \( i \)th bin. The probability of observing the value \( x_i \) is then \( p(x_i) \Delta \). This gives a discrete distribution for which the entropy takes the form

\[
H = -\sum_i p(x_i) \Delta \ln (p(x_i) \Delta) = -\sum_i p(x_i) \Delta \ln p(x_i) - \ln \Delta.
\]
Maximum entropy

- Consider a discrete R.V. with M possible status. We want to find the distribution has the maximum entropy: \( H[p] = - \sum_i p(x_i) \ln p(x_i) \).

\[
\tilde{H} = - \sum_i p(x_i) \ln p(x_i) + \lambda \left( \sum_i p(x_i) - 1 \right)
\]

\[
p(x_i) = \frac{1}{M} \quad \text{uniform distribution}
\]
Differential entropy

- Entropy is naturally defined on discrete random variables.
- But how about continuous variables?
Differential entropy

- Let us divide \( x \) into bins of \( \Delta \)

Mean-value theorem

\[
\int_{i\Delta}^{(i+1)\Delta} p(x) \, dx = p(x_i) \Delta
\]

Entropy on discretized probability

\[
H_{\Delta} = - \sum_i p(x_i) \Delta \ln (p(x_i) \Delta) = - \sum_i p(x_i) \Delta \ln p(x_i) - \ln \Delta
\]

\[
\sum_i p(x_i) \Delta = 1
\]
Differential entropy

\[ H_\Delta = - \sum_i p(x_i) \Delta \ln (p(x_i) \Delta) = - \sum_i p(x_i) \Delta \ln p(x_i) - \ln \Delta \]

\[ \lim_{\Delta \to 0} \left\{ \sum_i p(x_i) \Delta \ln p(x_i) \right\} = \int p(x) \ln p(x) \, dx \]

\[ H[x] = - \int p(x) \ln p(x) \, dx \]
Differential entropy

• The term that is thrown out reflects that to specify a continuous variable very precisely requires many many bits

• Note: differential entropy can be negative!
Differential entropy

• Given a continuous variable \( x \) with mean \( \mu \) and variance \( \sigma^2 \), which distribution has the largest entropy?

\[
\begin{align*}
\int_{-\infty}^{\infty} p(x) \, dx &= 1 \\
\int_{-\infty}^{\infty} x p(x) \, dx &= \mu \\
\int_{-\infty}^{\infty} (x - \mu)^2 p(x) \, dx &= \sigma^2.
\end{align*}
\]
Differential entropy

\[
\max - \int_{-\infty}^{\infty} p(x) \ln p(x) \, dx + \lambda_1 \left( \int_{-\infty}^{\infty} p(x) \, dx - 1 \right) \\
+ \lambda_2 \left( \int_{-\infty}^{\infty} x p(x) \, dx - \mu \right) + \lambda_3 \left( \int_{-\infty}^{\infty} (x - \mu)^2 p(x) \, dx - \sigma^2 \right)
\]

\[
p(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\}
\]

Gaussian distribution!
Conditional entropy

• Given \( \mathbf{x} \), how much information is left for \( \mathbf{y} \)

\[
H[\mathbf{y}|\mathbf{x}] = - \int \int p(\mathbf{y}, \mathbf{x}) \ln p(\mathbf{y}|\mathbf{x}) \, d\mathbf{y} \, d\mathbf{x}
\]

\[
H[\mathbf{x}, \mathbf{y}] = H[\mathbf{y}|\mathbf{x}] + H[\mathbf{x}]
\]

Prove it by yourself
Kullback-Leibler (KL) divergence

- Also called relative entropy

\[
\text{KL}(p\|q) = -\int p(x) \ln q(x) \, dx - \left( -\int p(x) \ln p(x) \, dx \right)
\]

\[
= -\int p(x) \ln \left( \frac{q(x)}{p(x)} \right) \, dx.
\]

If we use \( q \) to transmit information for \( p \), how much extra information do we need
Kullback-Leibler (KL) divergence

- KL divergence is widely used to measure the difference between two distributions

\[ KL(p \parallel q) \geq 0 \]

=0 iff \( p = q \)

- However, it is not symmetric!

\[ KL(p \parallel q) \neq KL(q \parallel p) \]

Prove it with convexity
And Jensen’s inequality
KL Divergence

- KL divergence plays the key role in approximate inference
- All the deterministic approximate methods aim to minimize the KL divergence between the true and approximate posteriors (or in the reversed direction)
- In general, we have alpha divergence
- We will discuss these in detail later
Mutual information

How many information do the two random variables share?

\[ I[x, y] \equiv \text{KL}(p(x, y) \| p(x)p(y)) \]

\[ = - \int \int p(x, y) \ln \left( \frac{p(x)p(y)}{p(x, y)} \right) \, dx \, dy \]


Prove it by yourself
What you need to know

• Definition of entropy
• How is differential entropy is derived
• Entropy is an indicator for uncertainty
• KL divergence and properties (especially asymmetric)
• Mutual information