Lecture 3: Information Theory Continues

Thin Nguyen
Oregon State University
Shannon’s measure of information is the number of bits to represent the amount of uncertainty (randomness) in a data source, and is defined as entropy.

\[ H = -\sum_{i=1}^{n} p_i \log(p_i) \]

Where there are \( n \) symbols \( 1, 2, \ldots, n \), each with probability of occurrence of \( p_i \).
Entropy: Three properties

1. It can be shown that $0 \cdot H \cdot \log N$.

2. Maximum entropy ($H = \log N$) is reached when all symbols are equiprobable, i.e., $p_i = 1/N$.

3. The difference $\log N - H$ is called the redundancy of the source.
Joint Information

- X and Y are random variables.
- X and Y can have n and m possibilities, respectively. Then, the joint information is defined as:

\[
H(X, Y) = -\sum_{i=1}^{n} \sum_{j=1}^{m} r(x_i, y_j) \log(r(x_i, y_j))
\]

- \( r(x,y) \) is the joint probability of x and y.
- Why this definition?
Conditional Information

- X and Y are random variables.

- X and Y can have n and m possibilities, respectively. Then, the conditional information is defined as:

\[ H(Y | X) = - \sum_{i=1}^{n} \sum_{j=1}^{m} r(x_i, y_j) \log(q(y_j | x_i)) \]

- \( q(y|x) \) is the conditional probability.

- Why this definition?
Conditional Information

Properties of Conditional Information:

1. \[ H(Y \mid X) \geq 0 \]
2. \[ H(Y \mid X) \leq H(Y) \] with equality if \( X \) and \( Y \) are independent.
3. \[ H(X, Y) = H(X) + H(Y \mid X) = H(Y) + H(X \mid Y) \]
Mutual Information

- X and Y are random variables.
- X and Y can have n and m possibilities, respectively. Then, the mutual information is defined as:

\[
I(X,Y) = H(Y) - H(Y \mid X) = \sum_{i=1}^{n} \sum_{j=1}^{m} r(x_i, y_j) \log \left( \frac{r(x_i, y_j)}{q(y_j) p(x_i)} \right)
\]

- Why this definition?
Mutual Information

- Properties of Mutual Information:

\[ I(X, Y) = I(Y, X) \]
Relationship among entropy, conditional, and mutual information

H(X) H(Y)
H(X|Y) I(X,Y) H(Y|X)

X and Y are dependent variables.

H(X) H(Y)

X and Y are independent variables.
Example:

- A vase contains 5 black balls and 10 white balls. Experiment x involves the random drawing of a ball, without being replaced in the vase. Experiment Y involves random drawing of the second ball.

- 5 black balls
- 10 white balls
Example: Entropy

- How much uncertainty (information) does experiment $X$ contain?

  $P(\text{black}_X) = 1/3, P(\text{white}_X) = 2/3$

  $$H(X) = -(1/3)\log(1/3) - (2/3)\log(2/3) = 0.92 \text{ bit}$$
Example:

- How much uncertainty (information) in experiment Y given that the ball in experiment X is white?

<table>
<thead>
<tr>
<th>Probability</th>
<th>Calculation</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(\text{black}_Y</td>
<td>\text{white}_X) = 5/14</td>
<td></td>
</tr>
<tr>
<td>P(\text{white}_Y</td>
<td>\text{white}_X) = 9/14</td>
<td></td>
</tr>
<tr>
<td>H(\text{Y}</td>
<td>\text{white}_X) = -(5/14)\log(5/14) - (9/14)\log(9/14)</td>
<td>= .94</td>
</tr>
</tbody>
</table>
Example:

- How much uncertainty (information) in experiment Y given that the ball in experiment X is black?

Drawing

\[ P(\text{black}_Y|\text{black}_X) = \frac{4}{14} = \frac{2}{7} \]
\[ P(\text{white}_Y|\text{black}_X) = \frac{10}{14} = \frac{5}{7} \]
\[ H(Y|\text{black}_X) = -(\frac{2}{7})\log(\frac{2}{7}) - (\frac{5}{7})\log(\frac{5}{7}) = 0.86 \text{ bit} \]
Example:

- How much uncertainty does experiment Y contain?

\[
H(Y) = P(\text{black}_X) \cdot H(Y|\text{black}_X) + P(\text{white}_X) \cdot H(Y|\text{white}_X)
\]

\[
= (1/3)(0.86) + (2/3)(0.94) = 0.91 \text{ bit}
\]
Formal Derivation of Entropy

- Why do we have

\[ H = -\sum_{i=1}^{n} p_i \log(p_i) \]
Axiomatic Foundations

Assuming that information measure should satisfy the three following requirements (Chaundy and McLeod (1960)):

1. If all outcomes are split up into groups, then all the values of $H$ for the various groups, multiplied by the statistical weights, should lead to the overall $H$.

2. $H$ should be continuous in $p_i$.

3. If all $p_i$’s are equal, i.e. for all $i$, $(p_i = 1/n)$, then $H$ will increase monotonically as a function of $n$. That means the uncertainty will increase for an increasing number of equal probabilities.
Derivation of Entropy

- Theorem: The only function that satisfy the three requirements above is

\[ H = -K \sum_{i=1}^{n} p_i \log(p_i) \]

- Proof:
Summary

- History of information theory.

- Information theoretical entities
  - Information, self-information, entropy, conditional information, joint information, mutual information.

- Derivation of $H = - \sum p_i \log p_i$