Information Theory
Lecture 1: Measuring amount of information. Entropy

September 4, 2022
Discrete ensembles

Notations:

- $X = \{x\}$ is a discrete set
- $\{p(x)\}$ s.t. $p(x) \geq 0$, $\sum_x p(x) = 1$ is a probability distribution, $X = \{x, p(x)\}$ is discrete ensemble
- $\Omega = \{A\}$ is algebra of subsets (events), where $A \subseteq \{x\}$
- Probability measure on $\Omega$ is defined as
  \[ P(A) = \sum_{x \in A} p(x), A \in \Omega \]
- Conditional probability:
  \[ P(A|B) = \frac{P(AB)}{P(B)}, \text{if } P(B) \neq 0 \text{ and } 0 \text{ otherwise} \]
- In general,
  \[ P(A_1...A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1A_2)...P(A_n|A_1...A_{n-1}). \]
Discrete ensembles

• $A, B \in \Omega$ are independent, if:

$$P(AB) = P(A)P(B).$$

• $A_1, \ldots, A_n \in \Omega$ are mutually independent, if:

$$P(A_1 \ldots A_n) = P(A_1)P(A_2)\ldots P(A_n).$$

• Equivalently, $A, B \in \Omega$ are independent iff

$$P(A|B) = P(A); \ P(B|A) = P(B).$$

• Probability of union of events:

$$P\left(\bigcup_{m=1}^{M} A_m\right) \leq \sum_{m=1}^{M} P(A_m).$$
• Law of total probability:
  Given $M$ mutually exclusive events $H_i$, $i = 1, ..., M$ whose probabilities sum to 1, probability of an arbitrary event $A$ is

$$P(A) = \sum_{m=1}^{M} P(A|H_m)P(H_m),$$

• Bayes’ theorem

$$P(H_j|A) = \frac{P(A|H_j)P(H_j)}{\sum_{m=1}^{M} P(A|H_m)P(H_m)}$$
Discrete ensembles

- Product of $X = \{x, p_X(x)\}$ and $Y = \{y, p_Y(y)\}$, is determined by a joint probability distribution $\{p_{XY}(x, y)\}$ on Cartesian product $\{(x, y) : x \in X, y \in Y\}$. Product ensemble $XY = \{(x, y), p_{XY}(x, y)\}$.

- Conditional probability distribution

$$p(x | y) = \begin{cases} \frac{p(x, y)}{p(y)}, & \text{if } p(y) \neq 0, \\ 0, & \text{otherwise}, \end{cases} \quad x \in X.$$ 

- Ensembles $X$ and $Y$ are independent, if

$$p(x, y) = p(x)p(y), \quad x \in X, \quad y \in Y.$$
Numerical characteristics of random variables

Mathematical expectation of $x \in X$ is defined as

$$E\{x\} = \sum_{x \in X} xp(x),$$

variance of $x \in X$ is defined as

$$D\{x\} = E\{(x - E\{x\})^2\}.$$

Covariance moment of two random variables $x \in X$ and $y \in Y$ is defined as

$$\lambda(x, y) = E\{(x - E\{x\})(y - E\{y\})\}.$$
Amount of information = cost (time or space) required for transmitting (storing) the data.

Examples:

<table>
<thead>
<tr>
<th>Data</th>
<th>Representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numbers (e.g. measurements)</td>
<td>Decimal numbers, length depends on range</td>
</tr>
<tr>
<td>Text</td>
<td>8-bit (1-byte) characters</td>
</tr>
<tr>
<td>Digital speech</td>
<td>13-bits samples</td>
</tr>
<tr>
<td>CD-audio</td>
<td>16-bits samples</td>
</tr>
<tr>
<td>Image (e.g. bmp)</td>
<td>3 bytes per pixel</td>
</tr>
</tbody>
</table>
Let $X = \{x, p(x)\}$ be a discrete ensemble, $\mu(x)$ is information measure (amount of information) in $x$.

- **Non-negative**: $\mu(x) \geq 0$
- $\mu(x)$ should be a function of $p(x)$.
- **Monotonic**: if $x, y \in X$, $p(x) \geq p(y)$, then $\mu(x) \leq \mu(y)$
- **Additive**: if $x$ and $y$ are independent, then $\mu(x, y) = \mu(x) + \mu(y)$
- $\mu(p(x)^k) = k \mu(p(x))$

The only function which satisfies all axioms is

$$I(x) = -\log p(x), \ x \in X$$
Definition:

\[ H(X) = \mathbb{E}[I(x)] \]  
\[ H(X) = \sum_x p(x) I(x) \]  
\[ H(X) = \sum_x -p(x) \log p(x) \]
Entropy. Examples

A: $X = \{a, b, c\}; \ p(a) = p(b) = p(c) = 1/3, \ 
I(a) = I(b) = I(c) = H(X) = \log 3 = 1.59 \ 
bits,$

B: $X = \{a, b, c\}; \ p(a) = p(b) = 1/4, \ p(c) = 1/2 \ 
I(a) = I(b) = 2, \ I(c) = 1 \ 
H = 1.5 \ bits$

C: $X = \{0, 1\}; \ p(0) = 0.9, \ p(1) = 0.1 \ 
I(0) = 0.152, \ I(1) = 3.322; \ 
H = 0.469 \ bits$

D: $X = \{0, 1\}, \ p(0) = p(1) = 1/2 \ 
I(0) = I(1) = H(X) = 1 \ bit$
1 \( H(X) \geq 0 \).

2 \( H(X) \leq \log |X| \). With equality iff all elements of \( X \) are equiprobable.

3 If \( X = \{x, p(x)\} \) and \( Y = \{y = f(x), p(y)\} \), then \( H(Y) \leq H(X) \) with equality iff \( f \) is one-to-one mapping.

4 If \( X \) and \( Y \) are independent, then

\[
H(XY) = H(X) + H(Y).
\]
5 $H(X)$ is a convex function of the probability distribution on elements of $X$.

6 Let $X = \{x, p(x)\}$ and $A \subseteq X$. Consider $X' = \{x, p'(x)\}$. Let $p'(x)$ be:

$$p'(x) = \begin{cases} \frac{P(A)}{|A|}, & x \in A, \\ p(x), & x \notin A. \end{cases}$$

Then $H(X') \geq H(X)$.

7 If for ensembles $X$ and $Y$ probability distributions differ only in the order of elements then $H(X) = H(Y)$. 
Proof of Property (2)

\[
H(X) - \log |X| \quad \overset{(a)}{=} \quad - \sum_{x \in X} p(x) \log p(x) - \sum_{x \in X} p(x) \log |X| = \\
\overset{(b)}{=} \quad \sum_{x \in X} p(x) \log \frac{1}{p(x)|X|} \leq \\
\overset{(c)}{=} \quad \log e \left[ \sum_{x \in X} p(x) \left( \frac{1}{p(x)|X|} - 1 \right) \right] = \\
= \quad \log e \left( \sum_{x \in X} \frac{1}{|X|} - \sum_{x \in X} p(x) \right) = 0 .
\]
Proof of (c)

- \( \ln x \leq x - 1 \iff \log x \leq (x - 1) \log e. \)

**Figure:** Graphical interpretation of \( \ln(x) \leq x - 1 \)