Information Relevance

The original formulation of information theory by Shannon purposely ignored the problem of information relevance, i.e., judging the value—of information to the recipient. Only concern: transmission efficiency.

Problem of information relevance is a priori not well-posed in the absence of additional knowledge about information structure/usage.

Example: speech → complex sound wave

Idea: the relevance of information conveyed about \( X \) in the representation \( X \) can be measured via the introduction of an additional relevance variable.

Tradeoff: there is a natural tradeoff between the representation size and the expected distortion of the meaningful content. Akin to rate-distortion theory.

Choice of relevant variable ↔ Feature selection
Rate distortion theory

\[ X \rightarrow \widetilde{X} : \text{encoding channel} \]

stochastic mapping \( p(\widetilde{X}|X) = \) soft partitioning/quantization

\[ \Omega \quad \Omega \]

\( n^{H}(X|X) = V = \text{average volume of input space that get mapped to the same code.} \)

\[ \# \text{ distinguishable states} = 2^{n^{H}(X)} \quad \frac{1}{2^{n^{H}(X)}} 2^{nI(X;X)} \]

Quality of the quantization is measured by the rate of bits needed to specify input without confusion, which is bounded below by \( I(X;\widetilde{X}) \).

Trade-off between rate and quality of the reconstruction as measured via a distortion function;

\[ d : X \times \widetilde{X} \rightarrow \mathbb{R}^+ : \text{ small when } X \text{ can be faithfully reconstructed from } \widetilde{X} \]

\[ \mathbb{E}_{p(x,\widetilde{x})}[d(x,\widetilde{x})] : \text{ expected distortion given a code} \]

\( \mathbb{E} \) the larger the rate, the smaller the achievable distortion

Convex optimization setting:

\[ R(D) = \min_{p(\widetilde{x}|x)} I(X;\widetilde{X}) \quad \text{s.t. } \mathbb{E}[d] \leq D \]

under constraint of expected distortion
**Variational problem and iterative algorithm**

\[ F[p(x|x)] = I(x; \tilde{x}) + \beta \mathbb{E}[d(x, \tilde{x})] + \lambda(x) \sum \mathbb{P}(\tilde{x} | x) \]

Lagrangian

Lagrangian multiplier

normalization

\[ F[p(\tilde{x}|x)] = \sum_{x' \in x} p(x)p(\tilde{x}|x) \log p(\tilde{x}|x) - \sum_{x} p(x) \log p(x) + \beta \sum_{x} \sum \mathbb{P}(\tilde{x} | x) d(x, \tilde{x}) + \lambda(x) \sum \mathbb{P}(\tilde{x} | x) \]

non-local term

\[ \frac{\delta F}{\delta p(\tilde{x}|x)} = p(x)(1 + \log p(\tilde{x}|x)) - \sum_{x'} \frac{Sp(\tilde{x})}{Sp(\tilde{x}'|x)} (1 + \log p(\tilde{x}')) + \beta p(x) d(x, \tilde{x}) + \lambda(x) \]

\[ = \delta_{\tilde{x} x} p(x|x) \]

assumed non-zero

\[ \frac{\delta F}{\delta p(\tilde{x}|x)} = 0 \Rightarrow p(\tilde{x}|x) = \frac{p(x)}{Z(x, \beta)} e^{-\beta d(x, \tilde{x})} \]

\[ Z(x, \beta) \leftarrow \text{normalization constant} \]

"Free energy"

Moreover, at optimum:

\[ SF = SI + \beta \delta \mathbb{E}[d] = 0 \implies \beta = -\frac{\delta F}{\delta D} \]

**Self consistent equations**

\[ (**\star) \quad p(\tilde{x}|x) = \sum_{x} p(\tilde{x} | x) p(x) \]

\[ (**\star\star) \quad p(\tilde{x}|x) = \frac{p(x) e^{-\beta d(x, \tilde{x})}}{Z(x, \beta)} \]

\[ R(D) \]

- \( \beta \): slope of rate distortion curve

achievable rate given distortion

\[ D \]
Iterative Blahut algorithm: \( P_{n+1}(x) = \sum_x P_n(x'|x) p(x) \)

\( P_{n+1}(x|x) = P_{n+1}(x) e^{-\beta d(x, \tilde{x})} \)

\( Z(x, p) \)

Justification:

\[ R(D) = \min_{p(x|z)} \max_{p(z|x)} I(x, y) = \min_{p(z|x)} \min_{p(x|z)} \max_{p(z|x)} F(p(x|z), p(y)) \]

\[ \text{s.t.} \quad \mathbb{E}[d] \leq D \]

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

\[ \text{minimization of } F \text{ can be performed independently on the convex sets of probabilities for } p(x|z) \text{ and } p(y) \]
Information bottleneck

relevant feature \[\downarrow\]
input representation \[\downarrow\]

\[Y \leftrightarrow X \leftrightarrow \tilde{X}\]
\[\tilde{X} \mid X\text{ and } X \mid \tilde{X}\text{ are independent}\]
\[\Rightarrow\text{ Markov information channel}\]

A double arrow notation to emphasize lack of directionality

\[Y \to X \to \tilde{X} = Y \leftarrow X \leftarrow \tilde{X}\]

* Relevant representations are those for which \(I(X,Y)\) is high. Remember that \(I(X,Y) \leq I(X,Y)\) by the data processing inequality.

* There is a tradeoff between preserving relevant information and compressing the representation. The goal is to pass as much information about \(Y\) via \(X\), but through the "bottleneck" formed by compact representations \(\tilde{X}\).

* Variational Formulation: \(L[p(\tilde{X}|x)] = I(X, \tilde{X}) - \beta I(X, Y)\)

\(\beta = 0\): most sketchy representation

\(\beta \to \infty\): arbitrary detailed quantization

\(\Rightarrow\text{ no explicit distortion function}\)

\(L\): nonlinear convex optimization problem
Self consistent equations

\[ L[p(x|x)] = \sum_{x,x'} p(x) p(x|x') \log p(x|x) - \sum_x p(x) \log p(x) \leq \text{non-local} \]

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

\[ - \sum_x \lambda(x) \sum_{x'} p(x|x') \leq \text{normalization constraints} \]

\[ \frac{\delta L}{\delta p(x|x)} = \frac{p(x)}{p(x|x)} \left( 1 + \log p(x|x) \right) - \frac{\delta p(x)}{\delta p(x|x)} \left( 1 + \log p(x) \right) \]

\[ - \beta \sum_y p(y) \frac{\delta p(x|y)}{\delta p(x|x)} \left( 1 + \log p(x|y) \right) + \beta \frac{\delta p(x)}{\delta p(x|x)} \left( 1 + \log p(x) \right) \]

\[ - \lambda(x) \]

Markov chain \( y \leftarrow x \leftarrow \tilde{x} : \)

\[ p(\tilde{x}) = \sum_{x} p(\tilde{x}|x) p(x) \]

\[ p(x|y) = \sum_{x'} p(\tilde{x}|x') p(x|y) \]

\[ \frac{\delta p(x)}{\delta p(x|x)} = p(x) \quad \text{and} \quad \frac{\delta p(x|y)}{\delta p(x|x)} = p(x|y) \]

\[ \frac{\delta L}{\delta p(x|x)} = p(x) \left\{ 1 + \log \frac{p(x|x)}{p(x)} - \lambda(x) \right\} \]

\[ \uparrow \quad \lambda(x) = \frac{\lambda(x)}{p(x)} \]

\[ = p(x) \left\{ \log \frac{p(x|x)}{p(x)} - \beta \sum_y p(y|x) \log \frac{p(y|x)}{p(y)} - \lambda(x) \right\} \]

\[ = p(x) \left\{ \log \frac{p(x|x)}{p(x)} + \beta \sum_y p(y|x) \log \frac{p(y|x)}{p(y|x)} \right\} \]

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

\[ \text{new multipliers:} \quad \lambda(x) \leftarrow - \left( \frac{\lambda(x)}{p(x)} + \beta \sum_y p(y|x) \log \frac{p(y|x)}{p(y)} \right) \]
\[
\frac{\delta L}{\delta p(\tilde{x}|x)} = 0 \Rightarrow p(\tilde{x}|x) = \frac{p(\tilde{x})}{Z(x/\beta)} e^{-\beta d(x, \tilde{x})}
\]

The Kullback-Leibler divergence \(d(x, \tilde{x}) = D_{KL}(p(y|x) || p(y|\tilde{x}))\) emerges as the relevant distortion function.

At optimum: self-consistent equations

\[
\begin{align*}
(*) & \quad p(\tilde{x}) = \sum_x p(\tilde{x}|x) p(x) \\
(\ast) & \quad p(y|x) = \sum p(y|x) p(x|\tilde{x}) \\
(\ast \ast) & \quad p(\tilde{x}|x) = \frac{p(\tilde{x})}{Z(x/\beta)} e^{-\beta d(x, \tilde{x})}
\end{align*}
\]

Markov chain: \(Y \leftrightarrow X \leftrightarrow \tilde{X}\)
Iterative algorithm

1. $p_n(\tilde{x}|x) = \frac{p_n(\tilde{x})}{Z_n(x, p)} e^{-\beta D_{KL}(p(y|x)\|p_n(y|x))}$ ← (***)

2. $p_{n+1}(\tilde{x}) = \sum_x p_n(\tilde{x}|x) p(x)$ ← (*)

2'. $p_{n+1}(y|\tilde{x}) = \sum_x p(y|x) p_n(x|\tilde{x})$ ← (***)

Interpretation

\[
\min_{p(\tilde{x}|x)} \min_{p(x)} \min_{p(y|x)} \min_{p(y|\tilde{x})} F[p(\tilde{x}|x), p(x), p(y|x)]
\]

s.t. $I(x, y) > D$

\[
F[p(\tilde{x}|x), p(x), p(y|x)] = I(x, \tilde{x}) + \beta E_{p(x, \tilde{x})} [D_{KL}(p(y|x)\|p(y|\tilde{x}))]
\]

+ $K[x, y] \leftarrow$ constant

regular rate-distortion function minimized when (*) and (***) hold at fixed $p(\tilde{x}|x)$

Functional of $p(y|x)$ minimized when (**) holds at fixed $p(x, \tilde{x})$