Open Study Answer #135

E. Castedo Ellerman

Proposed answer to the following questions:

- What is a copredictor?

  Related questions:

  - What is fraction of predictive information?

WORK IN PROGRESS

Copredictor Definition

Given any discrete random variables X and Y, define a *copredictor* of Y with X as any random variable W such that

$$\begin{aligned} &X \text{ and } W \text{ are independent and} \\ &Y = f(X, W) \text{ for some function } f \end{aligned}$$

Copredictor Existence

Given any discrete random variables X and Y, there exists a copredictor W. The range of W is $(\operatorname{rng} X) \mapsto (\operatorname{rng} Y)$, that is all functions from the finite range of X to the range of Y. Consider any x, y, w from $\operatorname{rng} X$, $\operatorname{rng} Y$, and $(\operatorname{rng} X) \mapsto (\operatorname{rng} Y)$ respectively. Denote

$$S \quad = \quad \{\omega : W(\omega) = w\} \cap \{\omega : X(\omega) = x\} \cap \{\omega : Y(\omega) = y\}$$

Define W such that

$$\mathrm{P}(S) \quad = \quad \mathrm{P}(X = x) \prod_{x' \in \operatorname{rng} X} \mathrm{P}(Y = w(x') | X = x')$$

when $w(x) = y$ otherwise $\mathrm{P}(S) = 0$.

Proof TO DO

A Functional Representation Lemma can be found on page 626 of [1] and is a more powerful result beyond just existence of a copredictor. A further Strong Functional Representation Lemma can be found in [2].

Optimal Copredictor

Given any copredictor W of Y with X, it follows that

$$\begin{aligned} \mathrm{H}(Y) &= I(X,W;Y) + H(Y|X,W) \\ &= I(X,W;Y) + 0 \\ &= I(X;Y) + I(W;Y) - I(X;W;Y) \\ &= I(X;Y) + I(W;Y) + I(X;W|Y) \end{aligned}$$

since $I(X;W;Y) + I(X;W|Y) = I(X;Y) = 0$.

We seek a copredictor that maximizes how much predictive information is provided 'exclusively' by independent variables X and W and not their interaction $I(X;W|Y)$.

To this end we want the minimum $I(X;W|Y)$ across all possible copredictors W of Y with X. This minimum is defined as the *excess functional information* in [2]. In this document, any copredictor achieving this minimum is considered an *optimal copredictor*.

The sum of mutual information and excess functional information is the conditional entropy of Y given an optimal copredictor W with X:

$$\begin{aligned} H(Y|W) &= I(X;Y|Y) + H(Y|X,W) \\ &= I(X;Y) - I(X;W;Y) + 0 \\ &= I(X;Y) + I(X;W|Y) \end{aligned}$$

Thus minimizing $I(X;W|Y)$ is equivalent to minimizing $H(Y|W)$. The optimal copredictor is also a copredictor that achieves a minimum $H(Y|W)$.

Linear Subspace of Copredictors

Consider any finite discrete random variables X and Y. Let F denote $(\operatorname{rng} X) \mapsto (\operatorname{rng} Y)$, the set of all functions from the range of X to the range of Y.

The following linear subspace of copredictors captures all of the possible values of $H(Y|W)$ and $I(X;W|Y)$. Consider all copredictors W of Y with X where the range of W is F, and for each $w \in F$, $x \in \operatorname{rng} X$, $y \in \operatorname{rng} Y$ with

$$S = \{\omega : W(\omega) = w\} \cap \{\omega : X(\omega) = x\} \cap \{\omega : Y(\omega) = y\}$$

the following holds

$$\mathrm{P}(S) = \begin{cases} \mathrm{P}(Y = y)\,\mathrm{P}(X = x) & \text{if } w(x) = y \\ 0 & \text{otherwise} \end{cases}$$

Each possible copredictor W defines a point in the subspace $F \mapsto \mathbb{R}$ where $q_w := P(W = w)$ is the coordinate along the dimension $w \in F$.

This is the linear subspace of possible q defined by the linear constraints:

$$P(Y = y|X = x) \quad = \quad \sum_{\substack{w \in F \\ w(x)=y}} q_w$$

for all $y \in \operatorname{rng} Y$, and $x \in \operatorname{rng} X$ and

$$q_w \quad \geq \quad 0$$

for all $w \in F$.

We show that $H(Y|W)$ is a linear function in terms of q_w across $w \in F$. Because of this, minimizing $H(Y|W)$ is minimizing a linear objective function in the subspace.

For any $w \in F$ and $y \in \operatorname{rng} Y$,

$$\begin{aligned}
\mathrm{P}(Y = y|W = w) &= \frac{\mathrm{P}(Y = y, W = w)}{\mathrm{P}(W = w)} \\
&= \sum_{x \in \operatorname{rng} X} \frac{\mathrm{P}(Y = y, X = x, W = w)}{\mathrm{P}(W = w)} \\
&= \sum_{\substack{x \in \operatorname{rng} X \\ w(x)=y}} \frac{\mathrm{P}(X = x)\,\mathrm{P}(W = w)}{\mathrm{P}(W = w)} \\
&= \sum_{\substack{x \in \operatorname{rng} X \\ w(x)=y}} \mathrm{P}(X = x) \\
&= \mathrm{P}(X \in w^{-1}(y))
\end{aligned}$$

For any $w \in F$ let

$$m(w) \quad = \quad \sum_{y \in \operatorname{rng} Y} \mathrm{P}(X \in w^{-1}(y)) \log_2 \frac{1}{\mathrm{P}(X \in w^{-1}(y))}$$

Note that $m(w)$ does not depend on any probabilities $\mathrm{P}(W = w)$ of a specific copredictor W. The function m only depends on functions $w \in F$. Combining the previous results shows that $H(Y|W)$ is a linear objective function.

$$\begin{aligned}
H(Y|W) &= \sum_{w \in F} P(W = w) \sum_{y \in \operatorname{rng} Y} P(Y = y|W = w) \log_2 \frac{1}{P(Y = y|W = w)} \\
&= \sum_{w \in F} q_w\, m(w)
\end{aligned}$$

Any copredictor Z can be simplified such that $H(Y|Z) = H(Y|W)$ for some W in the above linear subspace. For every value z in the range of Z, it must correspond to some function $w \in F = (\operatorname{rng} X \mapsto \operatorname{rng} Y)$ since $H(Y|X,W) = 0$. Let W be a random variable with range F such that $P(W = w)$ equals the sum of probabilities $\mathrm{P}(Z = z)$ for which z corresponds to w.

Linear Optimization

Minimization of the previously identified linear objective function within the linear subspace of copredictors is a standard form linear programming problem [3]. It is a primal problem for which a dual problem exists. Although the primal problem is an optimization on $|Y|^{|X|}$ variables, the dual problem is an optimzation on only $|Y| \cdot |X|$ variables.

In matrix notation, the dual linear program is to maximize $b^T r$ under constraint $A^T r \leq m$. The resulting maximum $b^T r$ equals the minimum possible $H(Y|W)$.

The vector b consists of the values $P(Y = y|X = x)$ across values $x \in \operatorname{rng} X$ and $y \in \operatorname{rng} Y$. The vector m consists of the values $m(w)$ across $w \in F$. The matrix expression $A^T r \leq m$ represents the linear constaints

$$\sum_{x \in \operatorname{rng} X} r_{(x,w(x))} \leq m(w)$$

across all $w \in F$. The maximized $b^T r$ represents

$$\sum_{\substack{x \in \operatorname{rng} X \\ y \in \operatorname{rng} Y}} \mathrm{P}(Y = y|X = x)\, r_{(x,y)}$$

and will equal the minimum possible $H(Y|W)$ across all copredictors W.

References

1. El Gamal AA, Kim Y-H (2011) Network information theory. Cambridge University Press, Cambridge ; New York

2. Li CT, Gamal AE (2017) Strong functional representation lemma and applications to coding theorems. In: 2017 IEEE International Symposium on Information Theory (ISIT). IEEE, Aachen, Germany, pp 589–593

3. Cornuejols G, Tütüncü R (2007) Optimization methods in finance. Cambridge University Press, Cambridge, UK ; New York