Approach based on pairwise conditional probabilities, M signals and $N+1$ nodes

Drazen Prelec, March 23, 2019

1 Notation: Images, pixels, responses

The setup is the same as the note of 3/16/19, with an 'image' variable S, taking on values $s = (s_1, .., s_M)$, $s_i \in \{0, 1\}$, and $N+1$ binary node response variables, Y^j, $j = 0, .., N$. Each pixel $i = 1, .., M$, can be On ($s_i = 1$) or Off ($s_i = 0$). The set of images is I.

As before, the 'strategy' for node Y^k is a vector of real-valued weights $y^j = (y_1^j, .., y_M^j)$ which determines response probabilities for each image,

$$\Pr(Y^j = 1|S = s) = \frac{1}{1 + \exp(-y_0^j - \sum_{i=1}^{M} y_i^j s_i)} \tag{1}$$

and $\Pr(Y^j = 2|S = s) = 1 - \Pr(Y^j = 1|S = s)$.

There are four changes in notation relative to previous versions:

- there is an additional 'baseline' parameter y_0^j in equation (1); without it, we would necessarily have $\Pr(Y^j = 1|S = (0, , .0)) = 0.5$, which may not always be desirable.
- averages are replaced by sums in the λ, κ parts of the utility function (2) below; this just rescales the relations between μ, λ, κ.
- the complete expected utility function is expressed as $EV(Y^k)$ instead of $U(Y^k)$, to avoid confusion with common notation for entropy, U = 'uncertainty'
- $\Pr(Y^k = i)$ is abbreviated as $p(y_i^k)$, etc..

Remark. If we constrain images so that exactly one pixel i is On, $s_i = 1$, and all others Off, $s_k = 0$, $k \neq i$, then this is equivalent to the earlier model. Using the old $x_1(s_i)$ notation, and letting $x_0 = 0$ in (1):

$$x_1(s_i) = \frac{1}{1 + \exp(-\sum_{i=1}^{M} x_i s_i)} = \frac{1}{1 + \exp(-x_i)}$$

so the 'weight' x_i in the new notation is:

$$x_i = \log \frac{x_1(s_i)}{1 - x_1(s_i)}$$

One would expect that the new 'image' model will perform just well as the old 'stimulus' model on these elementary patterns, unless there are differences in the updating process.

2 Entropies

The utility functions and mutual information are linear combinations of the following entropies,

$$H(S) = -\sum_{s\in I} p(s)\log p(s)$$

$$H(S,Y^k) = -\sum_{s\in I}\sum_{y^k} p(s,y^k)\log p(s,y^k) \qquad y^k = 1,2; k = 0,..,N$$

$$H(Y^k) = -\sum_{y^k} p(y^k)\log p(y^k) \qquad y^k = 1,2; k = 0,..,N$$

$$H(Y^j,Y^k) = -\sum_{y^j}\sum_{y^k} p(y^j,y^k)\log p(y^j,y^k) \qquad y^k = 1,2; y^j = 1,2; k,j = 0,..,N; j \neq k$$

$$H(Y^0,..,Y^N) = -\sum_{y^0}..\sum_{y^N} p(y^0,..,y^N)\log p(y^0,..,y^N) \qquad y^j = 1,2; j = 0,..,N$$

Remark. It is probably most efficient to compute, at each update step for node Y^k, the entropies $H(Y^k)$, $H(S,Y^k)$, $H(Y^j,Y^k)$. The final term $H(Y^0,..,Y^N)$ is needed for mutual information $I(S:(Y^0,..,Y^N))$, but not for the utilities. It could therefore be computed off-line, after the learning process ends. $H(S)$ is of course a constant.

3 Utilities

As before, Y^k adjusts strategies at each update step to maximize an expected score (utility) function, treating the strategies of nodes Y^j , $j \neq k$, as parameters:

$$\begin{aligned} EV(Y^k) =&\mu\sum_{s\in I}\sum_{y^k} p(s,y^k)\log p(y^k|s) \\ &+\lambda\sum_{s\in I}\sum_{j\neq k}\sum_{y^k}\sum_{y^j} p(y^j,y^k,s)\log \Pr(y^j|y^k) \\ &-\kappa\sum_{s\in I}\sum_{j\neq k}\sum_{y^k}\sum_{y^j} p(y^j,y^k,s)\log \Pr(y^k|y^j) \end{aligned}$$

$EV(Y^k)$ is also a linear function of the entropies (S disappears from the λ and κ weighted terms):

$$\begin{aligned} EV(Y^k) =&\mu(H(S) - H(S,Y^k)) \\ &+\lambda\sum_{j\neq k}(H(Y^k) - H(Y^j,Y^k)) \\ &-\kappa\sum_{j\neq k}(H(Y^j) - H(Y^j,Y^k)) \end{aligned} \tag{2}$$

An alternative (equivalent) formulation highlights the mutual information terms:

$$
\begin{aligned}
EV(Y^k) =& \mu I(S : Y^k) \\
& -(\kappa - \lambda) \sum_{j \neq k} I(Y^j : Y^k) \\
& +(N\kappa - \mu) H(Y^k) \\
& -\lambda \sum_{j \neq k} H(Y^j)
\end{aligned} \tag{3}
$$

Remark. If stimulus s is presented, the expected utility for $Y^k = i$ is:

$$
\begin{aligned}
EV(Y^k = i|s) =& \mu \log p(y_i^k|s) \\
& +\lambda \sum_{j \neq k} \sum_{y^j} p(y^j|s) \log p(y^j|y_i^k) \\
& -\kappa \sum_{j \neq k} \sum_{y^j} p(y^j|s) \log p(y_i^k|y^j)
\end{aligned}
$$

The difference in expected utilities, which is a plausible candidate for reinforcement signal, is:

$$
\begin{aligned}
EV(Y^k = 1|s) - EV(Y^k = 2|s) =& \mu \log \frac{p(y_1^k|s)}{p(y_2^k|s)} \\
& +\lambda \sum_{j \neq k} \sum_{y^j} p(y^j|s) \log \frac{p(y^j|y_1^k)}{p(y^j|y_2^k)} \\
& -\kappa \sum_{j \neq k} \sum_{y^j} p(y^j|s) \log \frac{p(y_1^k|y^j)}{p(y_2^k|y^j)}
\end{aligned}
$$

From (1) we have $\Pr(Y^j = 1|S = s)/\Pr(Y^j = 2|S = s) = (\exp(y_0^j - \sum_{i=1}^{M} y_i^j s_i))^{-1}$, hence the first term above is:

$$
\log \frac{p(y_1^k|s)}{p(y_2^k|s)} = y_0^j + \sum_{i=1}^{M} y_i^j s_i
$$

The other two terms have more complicated expressions however.

4 Mutual information

The mutual information between S and the ensemble is a simple function of stimulus entropy, pairwise node entropies and the additional term $H(Y^0, .., Y^N)$:

$$
I(S : (Y^0, .., Y^N)) = (N+1)H(S) - \sum_{j=0}^{N} H(S, Y^j) + H(Y^0, .., Y^N) \tag{3}
$$

The above equation exploits conditional independence

$$p(s,y^0,..,y^N) = \sum_{s\in S} p(s)\prod_{i=0}^{N} p(y^i|s)$$

or $H(S,Y^0,..,Y^N) = \sum_{j=0}^{N} H(S,Y^j) - NH(S)$. Although $H(Y^0,..,Y^N)$ needs to be evaluated for 2^{N+1} distinct combinations, this is also somewhat simplified by conditional independence.

5 Objective for each node

Ideally, one would like 'local' maximization of $EV(Y^k)$ by each node to also improve the 'global' encoding quality, as measured by $I(S:(Y^0,..,Y^N))$. This condition does not hold exactly, but an interesting relationship between mutual information $I(S:(Y^0,..,Y^N))$ and the utilities $EV(Y^k)$ can nevertheless be derived.

We start with an alternative expression for:

$$I(S:(Y^0,..,Y^N)) = \sum_{j=0}^{N} I(S:Y^j) - \sum_{j=0}^{N} H(Y^j) + H(Y^0,..,Y^N)$$

and solve for $I(S:Y^k)$:

$$I(S:Y^k) = I(S:(Y^0,..,Y^N)) - \sum_{j\neq k} I(S:Y^j) + \sum_{j=0}^{N} H(Y^j) - H(Y^0,..,Y^N)$$

After substituting $I(S:Y^k)$ into (3) and some manipulations, we obtain:

$$\begin{aligned} EV(Y^k) =& \mu I(S:(Y^0,..,Y^N)) \\ & -\mu H(Y^0,..,Y^N) \\ & +(\mu-\lambda)\sum_{j\neq k} H(Y^j|Y^k) \\ & +\kappa\sum_{j\neq k} H(Y^k|Y^j) \end{aligned}$$

For example, if $\lambda=\mu$, the conditional entropies $H(Y^j|Y^k)$ disappear, and the objective for Y^k will be to maximize information about S plus internal response constraint (that is, negative ensemble entropy, $-H(Y^0,..,Y^N)$) while alsostriving for pairwise independence (that is, increasing $H(Y^k|Y^j)$). Holding $I(S:(Y^0,..,Y^N)$ constant, the utility function puts a premium on higher-order, i.e., above pairwise statistical relationships in the ensemble response vector. Individual optimization by each node should produce complex response patterns.