Inverting the Information Bottleneck

ATH and DS

October 31, 2013

The Information Bottleneck

In their paper, *The Information Bottleneck Method*, Tishby, Pereira, and Bialek outline a principled method for choosing a distortion function given another relevant variable. The goal is very straightforward—to map signals $x \in X$ to a set of codewords $\tilde{x} \in \tilde{X}$, such that we retain a much information as possible about another signal $y \in Y$. The term 'bottleneck' implies that $|X| > |\tilde{X}|$.

In order to find this mapping, here is a set of conditionals—$p(\tilde{x}|x)$, we must minimize the functional:

$$\begin{aligned}\mathcal{L} &= I(X,\tilde{X}) - \beta I(\tilde{X},Y) - \sum_{x,\tilde{x}} \lambda(x) p(\tilde{x}|x) \\ &= \sum_{x,\tilde{x}} p(\tilde{x}|x)\, p(x) \log\left[\frac{p(\tilde{x}|x)}{p(\tilde{x})}\right] - \beta \sum_{\tilde{x},y} p(\tilde{x}|y)\, p(y) \log\left[\frac{p(\tilde{x}|y)}{p(\tilde{x})}\right] - \sum_{x,\tilde{x}} \lambda(x) p(\tilde{x}|x) \end{aligned} \tag{1}$$

with respect to the partitioning $\{p(\tilde{x}|x)\}$. The third term is simply a normalization constraint at each x. Taking derivatives with respect to each conditional for a given x and $\tilde{x}$, we get:

$$\begin{aligned}\frac{\delta \mathcal{L}}{\delta p(\tilde{x} = \tilde{x}^* | x = x^*)} &= 0 \\ &= p(x^*) \left(\log\left[\frac{p(\tilde{x}^*|x^*)}{p(\tilde{x}^*)}\right] - \beta \sum_y p(y|x^*) \log\left[\frac{p(y|\tilde{x}^*)}{p(y)}\right] - \frac{\lambda(x^*)}{p(x^*)} \right) \\ &= p(x^*) \left(\log\left[\frac{p(\tilde{x}^*|x^*)}{p(\tilde{x}^*)}\right] - \beta \sum_y p(y|x^*) \log\left[\frac{p(y|x^*)}{p(y|\tilde{x}^*)}\right] - \tilde{\lambda}(x^*) \right) \end{aligned} \tag{2}$$

Where we've just done some rearranging so that $\tilde{\lambda}(x^*)$ contains all the tens independent of $\tilde{x}$.

$$\tilde{\lambda}(x^*) = \frac{\lambda(x^*)}{p(x^*)} - \beta \sum_y p(y|x^*) \log\left[\frac{p(y|x^*)}{p(y)}\right] \tag{3}$$

Solving for $p(\tilde{x}|x)$ we see that our distortion measure has become the KL-divergence between the mapping of $Y \to X$ and the mapping of $Y \to \tilde{X}$:

$$\begin{aligned} p(\tilde{x}|x) &= \frac{1}{\mathcal{Z}(x,\beta)} p(\tilde{x}) \exp\left(\sum_y p(y|x) \log\left[\frac{p(y|x)}{p(y|\tilde{x})}\right] \right) \\ &= \frac{1}{\mathcal{Z}(x,\beta)} p(\tilde{x}) \exp\left(D_{KL}\left[p(y|x) \,||\, p(y|\tilde{x}) \right] \right) \end{aligned} \tag{4}$$

Where $\mathcal{Z}(x, \beta)$ is the usual normalization.

$$\mathcal{Z}(x, \beta) = \sum_{\tilde{x}} p(\tilde{x}) \exp\left(D_{KL}\left[p(y|x)\,||p(y|\tilde{x})\right]\right) \tag{5}$$

From here, Tisby et al. go on to explain an iterative algorithm for finding $\{p(\tilde{x}|x)\}$, $\{p(\tilde{x})\}$, and $\{p(y|\tilde{x})\}$ at every value of β. While this may end up being important, we will forgo a review at this time.

Finding Relevance

Working through the information bottleneck posses a natural follow up question—if I have some mapping $\{p(\tilde{x}|x)\}$ that arose as the result of relevant quantization, can I invert the process and find $\{p(y|\tilde{x})\}$ and/or $\{p(y|x)\}$? As stated the problem is underdetermined, but we can begin by asking what else do I need to know about $p(y)$ in order to reach a solution[1].

Let us begin from the derivative of the functional in eq. 2.

$$\frac{\delta \mathcal{L}}{\delta p(\tilde{x} = \tilde{x}^*|x = x^*)} = 0 = p(x^*)\left(\log\left[\frac{p(\tilde{x}^*|x^*)}{p(\tilde{x}^*)}\right] - \beta \sum_y p(y|x^*) \log\left[\frac{p(y|\tilde{x}^*)}{p(y)}\right] - \frac{\lambda(x^*)}{p(x^*)}\right) \tag{6}$$

Now assuming that we know $p(\tilde{x}^*|x^*)$, $p(x^*)$, $p(\tilde{x}^*)$, and $p(x^*)$, we can simplify the above expression.

$$\begin{aligned}
\sum_y p(y|x^*) \log\left[\frac{p(y|\tilde{x}^*)}{p(y)}\right] &= \frac{1}{\beta}\left[\log\left[\frac{p(\tilde{x}^*|x^*)}{p(\tilde{x}^*)}\right] - \frac{\lambda(x^*)}{p(x^*)}\right] \\
&= \frac{1}{\beta}\left[\eta(x^*, \tilde{x}^*) - \Lambda(x^*)\right] \\
&= \frac{1}{\beta}\Phi(x^*, \tilde{x}^*) \\
\sum_y p(y|x^*) \log\left[\frac{\sum_{x'} p(y|x')\, p(x'|\tilde{x}^*)}{p(y)}\right] &= \frac{1}{\beta}\Phi(x^*, \tilde{x}^*) \quad \forall\ x^*, \tilde{x}^*
\end{aligned} \tag{7}$$

We have some additional constraints corresponding to normalization and the fact that $\tilde{X}$ cannot encode anything about Y that is not encoded by X. At this point, it seems the first question one should ask is—given the correct β and the correct cardinality of Y, can one find a set $\{p(y|x)\}$ that satisfy eq. 7. We have NM equations where $N = |X|$ and $M = |\tilde{X}|$ and for each one $\Phi(x^*, \tilde{x}^*)$ is a known scalar. My hope is that this actually highly degenerate and there are many such sets. From there we can choose a set with desirable external properties.

I am going to set about trying to work through a small example at this point, but any help or input would be much appreciated, especially with regards to how to solve the system of equations given by eq. 7.

---

[1] We will leave for a moment the question of whether or not such a solution is useful.