Variational Deterministic Information Bottleneck

DJ Strouse & David Schwab

January 5, 2018

Abstract

Alemi et al. [2016] recently introduced a variational version of the information bottleneck (IB) [Tishby et al., 1999]. Here we introduce a variational version of the deterministic information bottleneck (DIB) [Strouse and Schwab, 2017].

1 Facts

The IB cost function that Alemi et al. [2016] developed a variational upper bound on was $L_{\mathrm{IB}}\left[q(z \mid x)\right] = I(X;Z) - \beta I(Z;Y)$. Here, we want to develop an upper bound on $L_{\mathrm{DIB}}[q(z \mid x)] = H(Z) - \beta I(Z;Y)$. Since the second terms are identical, we can borrow the (lower) bound from Alemi et al. [2016]. So we need only seek an upper bound on $H(Z)$. Just like in Alemi et al. [2016], we introduce the variational approximation $r(Z)$ to the marginal $q(Z)$, and use that $\mathrm{KL}[q(Z) \mid r(Z)] \geq 0 \implies \int dz\, q(z) \log q(z) \geq \int dz\, q(z) \log r(z)$ to write down:

$$H(Z) \leq -\int dx\, dz\, p(x)\, q(z \mid x) \log r(z)\,. \tag{1}$$

This should be compared with eqn 14 of Alemi et al. [2016]. The difference amounts to replacing $\mathrm{KL}[q(Z \mid X) \mid r(Z)]$ with the cross-entropy $H(q(Z \mid X)\,, r(Z))$.

2 Interpretation

What change in model behavior should we expect shifting from the KL to cross-entropy? To understand, note that the KL is exactly the cross-entropy minus the encoder entropy, that is:

$$\mathrm{KL}[q(Z \mid X = x) \mid r(Z)] = H(q(Z \mid X = x)\,, r(Z)) - H(q(Z \mid X = x))\,. \tag{2}$$

In the case that $r(Z)$ is fixed, the KL cost encourages the encoding to match the entire distribution, while the cross-entropy preferentially matches the mode and highest probability components of $r(Z)$. For continuous Z, this would likely result in a pathological collapse to the encoder using an infinitestimally small range of Z around the mode of $r(Z)$. However, for discrete Z, this would, as with the original DIB [Strouse and Schwab, 2017], have the potentially desirable effect of using only as many latents as needed to solve the task at hand, and ignoring the rest. This could, for example, be used to perform clustering, where the algorithm would be encouraged to only use as few clusters as it could get away with.

3 Geometric clustering

In geometric clustering, X becomes the data point index i, Y becomes the data point identity $\mathbf{x}$, and T becomes the cluster c. The VIB objective minimized in this case (of discrete latents) is:

$$
\begin{aligned}
L_{\text{VIB}} &= \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{c} -q(c \mid i)\log q(x_i \mid c) + \beta \text{KL}[q(c \mid i) \mid r(c)]\right] && (3)\\
&= \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{c} -q(c \mid i)\log q(x_i \mid c) + \beta \sum_{c} q(c \mid i)\log \frac{q(c \mid i)}{r(c)}\right] && (4)\\
&= \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{c} -q(c \mid i)\log q(x_i \mid c) - \beta \sum_{c} q(c \mid i)\log r(c) - \beta H(q(c \mid i))\right], && (5)
\end{aligned}
$$

whereas the DVIB objective would be:

$$
\begin{aligned}
L_{\text{DVIB}} &= \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{c} -q(c \mid i)\log q(x_i \mid c) + \beta H(q(c \mid i)\,, r(c))\right] && (6)\\
&= \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{c} -q(c \mid i)\log q(x_i \mid c) - \beta \sum_{c} q(c \mid i)\log r(c)\right]. && (7)
\end{aligned}
$$

Both encourage assigning data points to clusters which assign high probability to the observed data point location through the first term, but they differ in their "priors" over clusters, with VIB using a KL towards a variational marginal approximation, and DVIB using a cross-entropy.

References

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. *CoRR*, abs/1612.00410, 2016. URL http://arxiv.org/abs/1612.00410.

DJ Strouse and David J Schwab. The Deterministic Information Bottleneck. *Neural Computation*, 2017.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *Proceedings of The 37th Allerton Conference on Communication, Control, and Computing*, pages 368–377, 1999.