Past-future and observable information bottlenecks

November 3, 2020

1 Past-future information bottleneck

For a Markov process $\{x_t\}$ in the phase space Ω, the principle of past-future information bottleneck (PIB), also called predictive information bottleneck, defines a bottleneck variable $r_t = r(x_t)$ as the solution to

$$\max_r \mathcal{I}(r_t||x_{t+\tau}) - \gamma \cdot \mathcal{I}(r_t||x_t),$$

where $\mathcal{I}(\cdot||\cdot)$ denotes the mutual information (MI) between two random variables, and the hyperparameter $\gamma \geq 0$ can be selected due to the tradeoff between the complexity and prediction [1]. For simplicity of analysis, we assume in this article that $\dim(r_t) < \dim(x_t)$ and let $\gamma = 0$ as in [2].

According to [3],

$$\mathcal{I}(r_t||x_{t+\tau}) \leq \mathcal{I}(x_t||x_{t+\tau})$$

and the equality holds iff

$$p(x_{t+\tau}|r_t) = p(x_{t+\tau}|x_t),$$

i.e., $r_t = r(x_t)$ is a low-dimensional sufficient statistics for predicting $x_{t+\tau}$.

2 Donsker-Varadhan representation of PIB

Goal *For a given long trajectory $\{x_1, x_t, \ldots, x_T\}$, find a mapping r so that the MI $\mathcal{I}(r_t||x_{t+\tau})$ is (approximately) minimized.*

In [2], the MI is approximated by assuming that $p(x_{t+\tau}|r_t)$ is a multivariate normal distribution, which is obviously over-simplified. Based on the Donsker-Varadhan representation of MI [4], we have

$$\mathcal{I}(r_t||x_{t+\tau}) = \max_f \mathbb{E}_J \left[f(r_t, x_{t+\tau})\right] - \log \mathbb{E}_I \left[\exp f(r_t, x_{t+\tau})\right],$$

where $\mathbb{E}_J$ denotes the expectation over the joint distribution of $(r_t, x_{t+\tau})$, and $\mathbb{E}_I$ denotes the expectation with r_t and $x_{t+\tau}$ being independently sampled.

By modeling r, f by neural networks, we can then obtain the ideal bottleneck variable by solving

$$\max_{r,f} \mathbb{E}_J \left[f(r_t, x_{t+\tau}) \right] - \log \mathbb{E}_I \left[\exp f(r_t, x_{t+\tau}) \right].$$

3 Observable information bottleneck

In many practical applications, we are only interested in an (deterministic or random) observable y_t of the state instead of the whole state x_t.

Example 1. When investigating the reaction between two conformational states A, B, y_t can be defined as $(1_{x_t \in A}, 1_{x_t \in B})$ or the committor function.

Example 2. For a binding process of two proteins, y_t can be considered as the distance between the proteins.

For such cases, we propose a new bottleneck variable r_t as a solution to

$$\max_r \mathcal{L}(r) = \mathcal{I}(r_t || r_{t+\tau}, y_{t+\tau}) - \mathcal{I}(x_t || r_{t+\tau}, y_{t+\tau}).$$

It can be shown that $\mathcal{L}(r) \leq 0$, and we can obtain the following theorem, which implies that r_t is a sufficient statistics for predicting future observables and the its dynamics is Markovian. We call the solution r_t as the observable information bottleneck (OIB) variable in this article.

Theorem 3. *If* $\mathcal{L}(r) = 0$,

1. $p(r_{t+\tau}, y_{t+\tau} | r_t) = p(r_{t+\tau}, y_{t+\tau} | x_t)$,
2. $\{r_t\}$ *is a Markov process with*

$$p(r_{t+\tau} | r_t, r_{t-\tau}, \ldots) = p(r_{t+\tau} | r_t),$$

3. $\{r_t, y_t\}$ *is a Markov process with*

$$p((r_{t+\tau}, y_{t+\tau}) | (r_t, y_t), (r_{t-\tau}, y_{t-\tau}), \cdots) = p((r_{t+\tau}, y_{t+\tau}) | (r_t, y_t))$$

4. $p(y_{t+\tau}, r_{t+\tau} | r_t, y_t) = p(y_{t+\tau}, r_{t+\tau} | r_t)$.

Proof. 1.
$\mathcal{I}(r_{t+\tau}, y_{t+\tau} || r_t) \leq \mathcal{I}(r_{t+\tau}, y_{t+\tau} || x_t)$
$\Rightarrow r_t = \arg\max_s \mathcal{I}(r_{t+\tau}, y_{t+\tau} || s(x_t))$
By [3], we have
$p\{r_{t+\tau}, y_{t+\tau} | r_t\} = p\{r_{t+\tau}, y_{t+\tau} | x_t\}$. □

Proof. 2.

For an arbitrary $k, (r_t, r_{t-\tau}, \ldots, r_{t-k\tau}) \rightarrow x_t \rightarrow r_{t+\tau}$ is a Markov chain, which implies that

$$\mathcal{I}\left(r_{t+\tau} \| r_t, r_{t-\tau}, \ldots, r_{t-k\tau}\right) \leq \mathcal{I}\left(r_{t+\tau} \| x_t\right) = \mathcal{I}\left(r_{t+\tau} \| r_t\right)$$

Therefore, r_t is a sufficient statistics of $(r_t, r_{t-\tau}, \ldots, r_{t-k\tau})$ for predicting $r_{t+\tau}$ and

$$p\left(r_{t+\tau} \mid r_t, r_{t-\tau}, \ldots, r_{t-k\tau}\right) = p\left(r_{t+\tau} \mid r_t\right)$$

□

Proof. 3.

For an arbitrary k, $((r_t, y_t), (r_{t-\tau}, y_{t-\tau}), \ldots, (r_{t-k\tau}, y_{t-k\tau})) \rightarrow (x_t, y_t) \rightarrow (r_{t+\tau}, y_{t+\tau})$ is a Markov chain, which implies that

$$\begin{aligned}\mathcal{I}\left((r_{t+\tau}, y_{t+\tau}) \| ((r_t, y_t), (r_{t-\tau}, y_{t-\tau}), \ldots, (r_{t-k\tau}, y_{t-k\tau}))\right) &\leq \mathcal{I}\left((r_{t+\tau}, y_{t+\tau}) \| (x_t, y_t)\right) \\ &= \mathcal{I}\left((r_{t+\tau,}, y_{t+\tau}) \| (r_t, y_t)\right)\end{aligned}$$

Therefore, (r_t, y_t) is a sufficient statistics of $((r_t, y_t), (r_{t-\tau}, y_{t-\tau}), \ldots, (r_{t-k\tau}, y_{t-k\tau}))$ for predicting $(r_{t+\tau}, y_{t+\tau})$ and

$$p((r_{t+\tau}, y_{t+\tau}) | (r_t, y_t), (r_{t-\tau}, y_{t-\tau}), \cdots) = p((r_{t+\tau}, y_{t+\tau}) | (r_t, y_t))$$

□

Proof. 4.

$\mathcal{I}(y_{t+\tau}, r_{t+\tau} \| r_t) \leq \mathcal{I}(y_{t+\tau}, r_{t+\tau} \| r_t, y_t) \leq \mathcal{I}(y_{t+\tau}, r_{t+\tau} \| r_t, y_t, x_t)$

$(r_t, y_t) \rightarrow x_t \rightarrow (r_{t+\tau}, y_{t+\tau})$ is a Markov Chain.

$\Rightarrow \mathcal{I}(y_{t+\tau}, r_{t+\tau} \| r_t, y_t, x_t) = \mathcal{I}(y_{t+\tau}, r_{t+\tau} \| x_t)$

$\mathcal{I}(y_{t+\tau}, r_{t+\tau} \| r_t) = \mathcal{I}(y_{t+\tau}, r_{t+\tau} \| x_t) \leq \mathcal{I}(y_{t+\tau}, r_{t+\tau} \| x_t)$ □

$\Rightarrow \mathcal{I}(y_{t+\tau}, r_{t+\tau} \| r_t) = \mathcal{I}(y_{t+\tau}, r_{t+\tau} \| r_t, y_t)$

$\Rightarrow p(y_{t+\tau}, r_{t+\tau} \| r_t) = p(y_{t+\tau}, r_{t+\tau} \| r_t, y_t)$

We can obtain the following equivalent formulation of OIB based on the Donsker-Varadhan representation:

$$\begin{aligned}\max_{r,f} \min_{f_x} \quad & \mathbb{E}_J \left[f(r_t; r_{t+\tau}, y_{t+\tau}) - f_x(x_t; r_{t+\tau}, y_{t+\tau}) \right] \\ & - \log \mathbb{E}_I \left[\exp f(r_t; r_{t+\tau}, y_{t+\tau}) \right] \\ & + \log \mathbb{E}_I \left[\exp f_x(x_t; r_{t+\tau}, y_{t+\tau}) \right]\end{aligned}$$

Letting

$$\Delta f(x_t; r_{t+\tau}, y_{t+\tau}) = f_x(x_t; r_{t+\tau}, y_{t+\tau}) - f(r_t; r_{t+\tau}, y_{t+\tau}),$$

we can rewrite the formulation as

$$\begin{aligned}\max_{r,f}\min_{\Delta f}\mathcal{L}(r,f,\Delta f) &= \mathbb{E}_J\left[-\Delta f(x_t;r_{t+\tau},y_{t+\tau})\right] \\ &\quad -\log\mathbb{E}_I\left[\exp f(r_t;r_{t+\tau},y_{t+\tau})\right] \\ &\quad +\log\mathbb{E}_I\left[\exp\left(f(r_t;r_{t+\tau},y_{t+\tau})+\Delta f(x_t;r_{t+\tau},y_{t+\tau})\right)\right],\end{aligned}$$

and obtain the optimal $r, f, \Delta f$ by deep learning. Notice that $\mathcal{L}(r,f,\Delta f)=0$ if $\Delta f\equiv 0$, therefore $\mathcal{L}(r,f,\Delta f)\le 0$ for the optimal Δf.

References

[1] Tishby, N., Pereira, F. C. and Bialek, W. The information bottleneck method. Preprint at https://arxiv.org/abs/physics/0004057 (2000).

[2] Wang, Y., Ribeiro, J. M. L. and Tiwary, P. Past-future information bottleneck for sampling molecular reaction coordinate simultaneously with thermodynamics and kinetics, Nature Communications, 10:3573 (2019).

[3] Zhu, Z. Neural Sufficient Statistics Learning for Likelihood-free Inference. (Unpublished)

[4] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. arXiv preprint arXiv:1801.04062, 2018.