Information Bottleneck

(Mostly in Deep Learning)

Ivan Skorokhodov

September 16, 2019

Plan

1. Origins and intuitive understanding
2. IB is not a hidden variable model (no generative assumption)
3. It usually assumes that joint is known
4. Connection to minimal sufficient statistics
5. Connection to VAE/variational inference in general
6. Trivial solution for $\beta \leq 1$.
7. Problems in deterministic scenario
8. Implicit optimization in DL: SZT experiments and Saxe's critics
9. Explicit optimization in DL: TODO
10. Why IB? — Generalization bounds
11. Why IB? — Bias-variance tradeoff
12. MI estimators

Origins and intuitive understanding

- TODO: do we always can compute z from a new x, i.e. W is always invertible?
- PCA solves:

$$\min_{W,Z} \|WZ - X\|_F$$

  i.e. we try to find linear mapping W and latent codes Z such that X is “reconstructed well” from Z.
- IB solves

$$\min_{Z} I(X : Z) - \beta I(Z : Y)$$

  i.e. we try to find such Z that Y is “reconstructed well” and *we do not care about* X.
- Q1: Why should we care about minimizing $I(X : Z)$ if we only care about reconstructing Y?
- A1: $I(X : Z) = H(Z) - H(Z|X)$, so $H(Z)$ is minimized (a good property)
- Q2: But $H(Z|X)$ is maximized: why?
- A2: The reason is subtle, we can do well without that

$$
\begin{aligned}
\mathrm{D}_{\mathrm{KL}}[p(w) \parallel \tilde{q}(w)] &= \int \log \frac{p(w)}{\tilde{q}(w)} p(w) dw \\
&= \int \log \frac{p(w)}{\tilde{p}(w)} p(w) dw + \int \log \frac{\tilde{p}(w)}{\tilde{q}(w)} p(w) dw \\
&= \mathrm{D}_{\mathrm{KL}}[p(w) \parallel \tilde{p}(w)] + \int \sum_{i=1}^{d} \log \frac{\tilde{p}(w_i)}{\tilde{q}(w_i)} p(w) dw \\
&= \mathrm{D}_{\mathrm{KL}}[p(w) \parallel \tilde{p}(w)] + \sum_{i=1}^{d} \mathrm{D}_{\mathrm{KL}}[\tilde{p}(w_i) \parallel \tilde{q}(w_i)] \\
&= \mathrm{D}_{\mathrm{KL}}[p(w) \parallel \tilde{p}(w)] + \mathrm{D}_{\mathrm{KL}}[\tilde{p}(w) \parallel \tilde{q}(w)]
\end{aligned}
$$