The variational autoencoder

Clayton W. Seitz

February 21, 2022

Outline

References

The logic of generative modeling

Say we have a set of variables $\mathrm{x} = (x_1, x_2, ..., x_n)$ which might have some statistical dependence

The variable x might be an amino acid sequence, gene expression data, microscopy image, etc.

- Often we are handed a batch of empirical samples $\{\mathrm{x}_i\}_{i=1}^N$
- We want to know the generating distribution $p(\mathrm{x})$

In supervised generative learning, we try to explicity learn the joint distribution $p(\mathrm{x}) = \prod_{i=1}^{N-1} p(x_i|x_{i+1:N})p(x_N)$, which is generally more difficult than discriminative learning.

Perks of generative modeling

- Fitting complete multivariate distributions $p(\mathrm{x})$ goes beyond correlation-based or clustering approaches
- Correlations cannot discover partial correlation in the context of other neighbors
- Fitting $p(\mathrm{x})$ permits sampling based inference

Why generative modeling is difficult

When describing a distribution over multiple variables, we may not know the proper normalization Z. That is,

$$p(\mathrm{x}) = \frac{1}{Z}\tilde{p}(\mathrm{x})$$

This very important situation arises in several contexts:

1. In Bayesian inference where $p(x_1|x_2) = p(x_2|x_1)p(x_1)/p(x_2)$ is intractable due to $Z = p(x_2) = \int p(x_2|x_1)p(x_1)dx_1$. This integral can be very difficult or impossible to compute.

2. In models from statistical physics, e.g. the Ising model, we only know $\tilde{p}(\mathrm{x}) = e^{-H(\mathrm{x})}$ where $H(\mathrm{x})$ is the Hamiltonian

Bayesian inference

The variable x has a latent representation or code z. We often say that z is the *causal source* of x. Ultimately, we would like to know the distribution $P_{\Phi}(\mathsf{x})$

$$P_{\phi}(\mathsf{x}) = \frac{P_{\phi}(\mathsf{x}|\mathsf{z})P_{\phi}(\mathsf{z})}{Q_{\psi}(\mathsf{z}|\mathsf{x})}$$

in order to find the model parameters that maximize the likelihood of the observed data:

$$\Phi^* = \underset{\Phi}{\mathrm{argmin}} - \log P_{\Phi}(\mathsf{x})$$

but we generally do not know $P_{\psi}(\mathsf{z}|\mathsf{x})$ due to the intractable integral $Z = \int P_{\phi}(\mathsf{x}|\mathsf{z})P_{\phi}(\mathsf{z})dz$ (see slide 5)

Computing the evidence

We can rewrite the evidence as

$$\begin{aligned} P_\phi(\mathsf{x}) &= \int P_\phi(\mathsf{z})P_\phi(\mathsf{x}|\mathsf{z})dz \\ &= \int P_\phi(\mathsf{z})P_\phi(\mathsf{x}|\mathsf{z})\frac{P_\phi(\mathsf{z}|\mathsf{x})}{P_\phi(\mathsf{z}|\mathsf{x})}dz \\ &= \mathbb{E}_{\mathsf{z}\sim P_\phi(\mathsf{z}|\mathsf{x})}\frac{P_\phi(\mathsf{z})P_\phi(\mathsf{x}|\mathsf{z})}{P_\phi(\mathsf{z}|\mathsf{x})} \end{aligned}$$

where $P_\phi(\mathsf{z}|\mathsf{x})$ is our model "encoder"

The evidence lower bound (ELBO)

$$
\begin{aligned}
\log P_\phi(\mathrm{x}) &= \log \int_z P(x,z)dx \\
&= \log \int_z P(x,z)\frac{Q(z|x)}{Q(z|x)}dz \\
&= \log \mathbb{E}_{\mathrm{z}\sim P_\phi(\mathrm{z}|\mathrm{x})} \frac{P(x|z)P(z)}{Q(z|x)} \\
&\geq \mathbb{E}_{\mathrm{z}\sim P_\phi(\mathrm{z}|\mathrm{x})} \log \frac{Q(x|z)}{P(z)} + \log P(x|z)
\end{aligned}
$$

$$
-\log P_\phi(\mathrm{x}) \leq \mathbb{E}_{\mathrm{z}\sim P_\phi(\mathrm{z}|\mathrm{x})} \log \frac{Q(x|z)}{P(z)} - \log P(x|z)
$$

The ELBO objective

$$
\begin{aligned}
\Phi^* &= \mathcal{L}(\Phi) \\
&= \underset{\Phi}{\operatorname{argmin}} \;\; \mathbb{E}_{\mathsf{x}\sim\mathrm{Pop},\, \mathsf{z}\sim P_\phi(\mathsf{z}|\mathsf{x})} \log \frac{Q_\Psi(\mathsf{z}|\mathsf{x})}{P(\mathsf{z})} - \log P(\mathsf{x}|\mathsf{z})
\end{aligned}
$$

The ELBO can be rewritten in terms of a KL-divergence and population entropy. We often think of Φ as "position" and the loss $\mathcal{L}$ as an "energy"

References I