VAE's: bottom-up and more

Contents

1 Introduction

When estimating true posterior $P(z|x)$ is difficult, we can instead learn an approximate posterior $Q(z|x)$ and enforce $P(z|x)$ to be close to $Q(z|x)$, using say, KL divergence. For simplicity we assume $Q(z|x) \sim \mathcal{N}(\mu(x), \Sigma(z))$ and learn it's parameters.

$$
\begin{aligned}
KL(Q(z|x) \| P(z|x)) &= \mathbb{E}_z \left[\log \frac{Q(z|x)}{P(z|x)} \right] \\
&= \mathbb{E}_z[\log Q(z|x)] - \mathbb{E}_z[\log P(z|x)] \\
&= \mathbb{E}_z[\log Q(z|x)] - \mathbb{E}_z \left[\log \frac{P(x|z)P(z)}{P(x)} \right] \\
&= \mathbb{E}_z[\log Q(z|x)] - \mathbb{E}_z \left[\log P(x|z) \right] - \mathbb{E}_z \left[\log P(z) \right] + \log P(x)
\end{aligned}
$$

VAE objective to maximize to learn latents

$$
\mathbb{E}_z[\log P(x|z)] - KL(Q(z|x) \| P(z))
$$

Assuming $P(z) \sim \mathcal{N}(0, 1)$,

$$
\begin{aligned}
KL(Q(z|x) \| P(z)) =& \mathbb{E}_z[-\frac{1}{2} \left((z - \mu(x))^\top \Sigma(x)^{-1} (z - \mu(x)) + \|z\|^2 \right)] - \log |\Sigma(x)| \\
=& \frac{-1}{2} \left(\mathbb{E}_z[\mathrm{Tr} \left(\Sigma(x)^{-1} (z - \mu(x))(z - \mu(x))^\top \right)] + \mathbb{E}_z[\|z\|^2] \right) - \log |\Sigma(x)| \\
=& \frac{-1}{2} \left(d + \mathbb{E}_z[\|z\|^2] \right) - \log |\Sigma(x)| \text{ where } d \text{ is the dimension of } z \\
=& \frac{-1}{2} \left(d + \mathbb{E}_z[\mathrm{Tr}(zz^\top)] \right) - \log |\Sigma(x)| \\
=& \frac{-1}{2} \left(d + \mathrm{Tr}(\Sigma(x) + \mu(x)\mu(x)^\top) \right) - \log |\Sigma(x)| \\
=& \frac{-1}{2} \left(d + \mathrm{Tr}(\Sigma(x)) + \|\mu(x)\|^2 \right) - \log |\Sigma(x)|
\end{aligned}
$$

Assuming $\Sigma(x) = \sigma \mathbf{I}_d$ the objective simplifies to $\frac{-1}{2} \sum_{k=1}^{d} \left(1 + \sigma + \mu_k(x)^2 + 2 \log \sigma \right)$
For numerical stability, we can model $\bar{\sigma} = \log \sigma$.

2 More on the VAE objective to minimize

We want the learnt distribution over latents $Q(z) = \mathbb{E}_{x\sim P(x)}[Q(z|x)]$, to match a prior distribution $P(z)$.

The VAE objective to maximize for the dataset

$$\begin{aligned}
&-\lambda\mathbb{E}_{x\sim P(x)}[KL(Q(z|x)\|P(z))] + \mathbb{E}_{x\sim P(x)}[\mathbb{E}_{z\sim Q(z|x)}[\log P(x|z)]] \\
&= -\lambda\mathbb{E}_{x\sim P(x)}\left[\mathbb{E}_{z\sim Q(z|x)}\left[\log\frac{Q(z)}{P(z)}\right]\right] + \mathbb{E}_{x\sim P(x)}[\mathbb{E}_{z\sim Q(z|x)}[\log P(x|z)]] \\
&= -\lambda\int_z\int_x \log\frac{Q(z)}{P(z)}Q(z|x)P(x)\mathrm{dxdz} + \mathbb{E}_{x\sim P(x)}[\mathbb{E}_{z\sim Q(z|x)}[\log P(x|z)]] \\
&= -\lambda\int_z \log\frac{Q(z)}{P(z)}Q(z)\mathrm{dz} + \mathbb{E}_{x\sim P(x)}[\mathbb{E}_{z\sim Q(z|x)}[\log P(x|z)]] \\
&= -\lambda KL(Q(z)\|P(z)) + \mathbb{E}_{x\sim P(x)}[\mathbb{E}_{z\sim Q(z|x)}[\log P(x|z)]]
\end{aligned}$$

2.1 [WIP]Challenges and Alternatives

2.1.1 Semantically not optimal

Minimizing $\mathbb{E}_{x\sim P(x)}[KL(Q(z|x)\|P(z))]$ intuitively means mapping $Q(z|x)$ to $P(z)$ for each x. Alternative: **MMD-VAE**

2.1.2 KL Vanishing Away Problem