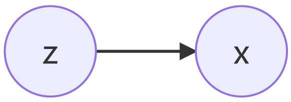
Author：Liu Jian

Time：2021-06-23

生成模型——变分自编码器 (VAE)

VAE 的构建

VAE 的概率图和 GMM 相同：



模型基于可微生成器网络给出分布参数，构建过程如下：

- 隐变量 $z \sim \mathcal{N}(z; 0, I)$，记分布为 $p_0(z)$；
- $x|z \sim \mathcal{N}(x; \mu_\theta(z), \Sigma_\theta(z))$，参数 $\mu_\theta(z), \Sigma_\theta(z)$ 经由一个神经网络输出，其网络结构给定，z 为网络输入，θ 为待求网络参数，记分布为 $p_\theta(x|z)$；

最终，构建的生成模型为：

$$p_\theta(x) = \int p_\theta(x, z)dz = \int p_\theta(x|z)p_0(z)dz$$

VAE 的训练

接下来就是模型的训练了，一般而言，有了 $p_\theta(x)$ 的形式，直接极大似然估计即可，可这里涉及到积掉隐变量 z，MLE 可能很难求解，因此，我们采用 EM 算法 + 学成近似推断求解。此外，为了便于描述和区别，我们假设从真实环境中只采集了一个样本，记为 $x^{(i)}$ ($x^{(i)} \sim p_{real}(x)$)，以此来描述学习过程。

首先假设分布 $q(z|x)$ 的形式，并基于可微生成器网络给出分布参数：

$$z|x \sim \mathcal{N}(z; \mu_\phi(x), \Sigma_\phi(x))$$

参数 $\mu_\phi(x), \Sigma_\phi(x)$ 由另一个神经网络输出，其网络结构给定，x 为网络输入，ϕ 为待求网络参数，记分布为 $q_\phi(z|x)$。

变分下界 ELBO：

$$\begin{aligned}
L(\theta, \phi) &= \log p_\theta(x^{(i)}) - D_{KL}\left(q_\phi(z|x^{(i)}) \| p_\theta(z|x^{(i)})\right) \\
&= \mathbb{E}_{z \sim q_\phi(z|x^{(i)})}\left[\log p_\theta(x^{(i)}, z)\right] + H\left(q_\phi(z|x^{(i)})\right) \\
&= \mathbb{E}_{z \sim q_\phi(z|x^{(i)})}\left[\log p_\theta(x^{(i)}|z)\right] - D_{KL}\left(q_\phi(z|x^{(i)}) \| p_0(z)\right)
\end{aligned}$$

目标是：$\max_{\theta,\phi} L(\theta, \phi)$

E 步：

学成近似推断，即从数据中学得推断 q_ϕ (也就是学习 ϕ)，而不是像变分推断那样，求解的是一个泛函问题 (变分推断不假设 q 的参数形式，只是在诸如平均场假设的条件下，解析求解分布 q)。注意这里的数据并不是从真实环境中采集的，而是从我们构建的模型中采集的，实际上就是一种蒙特卡洛法。我们选用的学成近似推断算法为随机梯度变分推断 (SGVI；wake-sleep 算法也是一种学成近似推断算法)。

我们要 $\min_\phi L$，使用梯度下降法，需要计算梯度，经过一番推导：

$$\nabla_\phi L = \mathbb{E}_{z\sim q_\phi(z|x^{(i)})}\left[\left(\nabla_\phi \log q_\phi(z|x^{(i)})\right)\left(\log p_\theta(x^{(i)}, z) - \log q_\phi(z|x^{(i)})\right)\right]$$

再使用蒙特卡洛法估计上述梯度值：

$$z^{(l)} \sim q_\phi(z|x^{(i)}), \;\; l = 1, \cdots, L$$
$$\nabla_\phi L \approx \frac{1}{L}\sum_{l=1}^{L}\left[\left(\nabla_\phi \log q_\phi(z^{(l)}|x^{(i)})\right)\left(\log p_\theta(x^{(i)}, z^{(l)}) - \log q_\phi(z^{(l)}|x^{(i)})\right)\right]$$

M 步：

我们要 $\min_\theta L \Leftrightarrow \min_\theta \mathbb{E}_{z\sim q_\phi(z|x^{(i)})}\left[\log p_\theta(x^{(i)}|z)\right]$，使用梯度下降法：

$$\begin{aligned}\nabla_\theta L &= \nabla_\theta \mathbb{E}_{z\sim q_\phi(z|x^{(i)})}\left[\log p_\theta(x^{(i)}|z)\right] \\ &= \mathbb{E}_{z\sim q_\phi(z|x^{(i)})}\left[\nabla_\theta \log p_\theta(x^{(i)}|z)\right]\end{aligned}$$

和上面类似，使用蒙特卡洛法估计梯度：

$$z^{(l)} \sim q_\phi(z|x^{(i)}), \;\; l = 1, \cdots, L$$
$$\nabla_\theta L \approx \frac{1}{L}\sum_{l=1}^{L}\nabla_\theta \log p_\theta(x^{(i)}|z^{(l)})$$

VAE 再回首

上面介绍了 VAE 的构建和训练，

- VAE 是可微生成器网络 (Differentiable Generator Nets, DGN；见 《Deep Learning》20.10.2 节) 的一个具体例子，它采用方案 2，即使用神经网络输出分布的参数。VAE 存在两个神经网络，一个是用于 VAE 的构建，一个是用于 VAE 的训练 (近似推断)；
- VAE 顾名思义是一种 AE，即自编码器。从 VAE 的构建可以看到，我们实际上就是假设了一个概率模型 $p_\theta(x)$，要使其尽可能地符合真实采样得到的数据分布，但是所有的概率模型都是这么做的呀，既然说它是一种自编码器，那它体现在哪里呢？这需要结合训练一起来看。因为直接使用 MLE 很难优化，我们选用 EM 算法，为此基于 NN 构建了分布 $q_\phi(z|x)$ (该分布实际上就是在拟合 $p_\theta(z|x)$)；结合 VAE 模型构建时我们基于 NN 构建的分布 $p_\theta(x|z)$，可以看到，通过学习这两个神经网络 (即参数 θ, ϕ) 我们实际上完成了一个 AE 的构建：

  $$x \overset{q_\phi(z|x)}{\longrightarrow} z \overset{p_\theta(x|z)}{\longrightarrow} x'$$

  此外，考虑我们的优化目标：

  $$\max_{\theta,\phi}\; L(\theta,\phi) = \max_{\theta,\phi}\; \mathbb{E}_{z\sim q_\phi(z|x^{(i)})}\left[\log p_\theta(x^{(i)}|z)\right] - D_{KL}\left(q_\phi(z|x^{(i)}) \| p_0(z)\right)$$

  可以看到，**最大化第一项 $\mathbb{E}_{z\sim q_\phi(z|x^{(i)})}\left[\log p_\theta(x^{(i)}|z)\right]$ 实际上就是在训练一个自编码器 (最小化重构误差)，即给定 $x^{(i)}$，由此，编码器 $q_\phi(z|x^{(i)})$ 先生成 z，基于得到的 z 再由解码器 $p_\theta(x|z)$ 生成 x'，而生成的 x' 是 $x^{(i)}$ 的概率越高越好；第二项 $D_{KL}\left(q_\phi(z|x^{(i)}) \| p_0(z)\right)$ 实际上就是一个正则化项，即 $p_0(z)$ 为人为设定的先验信息，我们希望后验分布或者说近似推断 $q_\phi(z|x^i)$ 与之越相近越好。**

- VAE 的训练不属于变分推断，而属于学成近似推断，但之所以叫"变分"AE，是因为其训练是在优化变分下界 ELBO，即采用的 EM 算法。
