

日期： /

latent variable models

隐变量模型有以下动机：

① 如果选择合适的隐变量 z , 建模 $P(x|z)$ 要易于 $P(x)$

② 通过正确训练, 从 z 中可以提取特征

隐变量模型的思想是：

$- \underline{z \sim N(0, I)}$ - 一个相对随意的先验

隐变量 z 的维数 $- P(x|z) = N(\mu_\theta(z), \Sigma_\theta(z))$

通常远小于 x 的维度 GMM的思想

$$-\mu_\theta(z) = \phi(Az + c) = (\phi(a_1 z + c_1), \phi(a_2 z + c_2)) = (\mu_1(z), \mu_2(z))$$

$$-\Sigma_\theta(z) = \text{diag}(\exp(\phi(Bz + d))) = \begin{pmatrix} \exp(\phi(b_{11}z + d_{11})) & 0 \\ 0 & \exp(\phi(b_{22}z + d_{22})) \end{pmatrix}$$

$$-\theta = (A, B, c, d)$$

此时 x 的生成并不依赖自回归结构

但这种结构并非端到端的可微分问题

[混合高斯分布模型(GMM)]

GMM是一种基于概率的聚类方法, 广泛应用于无监督学习中。归纳偏置认为数据的真实分布由多个高斯分布的线性组合生成。

$$P(x, \theta) = \sum_{k=1}^K \alpha_k \phi(x, \mu_k, \Sigma_k)$$

日期： /

GMM 通过两个步骤更新参数

① E step: $y_{ik} = P(z_i=k | x_i, \theta) = \frac{\alpha_k \phi(x_i; \mu_k, \Sigma_k)}{\sum_{j=1}^K \alpha_j \phi(x_i; \mu_j, \Sigma_j)}$

② M step: $\alpha_k = N_k / N$

$$\mu_k = \sum_{i=1}^N y_{ik} x_i / N_k$$

$$\Sigma_k = \sum_{i=1}^N y_{ik} (x_i - \mu_k)(x_i - \mu_k)^T / N_k$$

对于 GMM，给出 K 的情况下也可以在特定的分布中采样

$$p(x) = \sum_z p(x, z) = \sum_z p(z) p(x|z) = \sum_{k=1}^K p(z=k) \mathcal{N}(x; \mu_k, \Sigma_k)$$

VAE 天然适合无监督任务，同时也更难于训练

在 VAE 中，直接通过 $\int p_\theta(x|z) p(z) dz$ 计算边缘分布 $p(x)$ 的成本是极大的。这是由于 z 是未被观测到的，若使用极大似然的方式：

$$\log \prod_{x \in D} p(x, \theta) = \sum_{x \in D} \log p(x, \theta) = \sum_{x \in D} \log \sum_z p(x, z, \theta)$$

一种自然的简化方法是使用 Monte-Carlo 代价极大
由其对于 z 取值极多或其分布
副情况

$$P_\theta(x) = \sum_z p_\theta(x, z) = |Z| \sum_{z \in Z} \frac{1}{|Z|} p_\theta(x, z) = |Z| \mathbb{E}_{z \sim q_{\text{prior}}(z)} [p_\theta(x, z)]$$

但是无意义，因为 z 并非均匀分布

日期： /

取而代之的是 Importance Sampling，即我们不再均匀采样，而是集中采样。

变分推理(VI) $P_{\theta}(x) = \sum_z P_{\theta}(x, z) = \sum_{z \in Z} \frac{q(z)}{q(z)} P_{\theta}(x, z) = E_{z \sim q(z)} \left[\frac{P_{\theta}(x, z)}{q(z)} \right]$

此后，我们并不均匀采样 z ，而是由 $q(z)$ 采样 z 。
则有 $P_{\theta}(x) \approx \frac{1}{K} \sum_{k=1}^K \frac{P_{\theta}(x, z^{(k)})}{q(z^{(k)})}$

现在的问题是，如何选择 $q(z)$ ？

$q(z)$ 的选择依赖于 x

在 f 非线然而我们实际关心的是

性能，即使 $E(x)$ 为 $\log(P(x)) = \log(\sum_z P(x, z))$

$E(x)$ 的无偏估计。这使得原本的无偏估计成为了有偏估计

$+(\hat{E}(x))$ 也 [Jensen 不等式] $\log(E_{z \sim p(x)}[f(z)]) = \log(\sum_z q(z)f(z)) \geq \sum_z q(z)\log f(z)$

不一定为 $+(\bar{E}(x))$ 那么 $\log(E_{z \sim q(x)}[\frac{P_{\theta}(x, z)}{q(z)}]) \geq E_{z \sim q(x)}[\log(\frac{P_{\theta}(x, z)}{q(z)})]$

的无偏估计 也就是说不可对 $P_{\theta}(x, z)/q(z)$ 采样后取对数 这是一个下界

对 q 的选择影响了下界的紧密程度

当 $q(z) \geq P(z|x)$ 时，近似是紧致的 变分下界 (ELBO)

ELBO 可以写作 $E_{q_{\phi}(z|x)}[\log P_{\theta}(x|z)] - D_{KL}(q_{\phi}(z|x) || p(z))$

重建损失

KL 散度

日期： /

有另一种推导 ELBO 的方式：

$$KL(q(z) \parallel p(z|x))$$

这是难以计算的，因此以 $q(z)$ 代替。
VAE 处理为 $p(z|x) = N(\mu_\theta(x), \Sigma_\theta(x))$

$\mu_\theta, \Sigma_\theta$ 是由神经网络产生的变分参数

$$= - \int_z q(z) \log \left[\frac{p(z|x)}{q(z)} \right] dz$$

$$= \int_z q(z) \log q(z) dz - \int_z q(z) \log p(z|x) dz$$

$$= E_{z \sim q(z)} [\log q(z)] - E_{z \sim q(z)} [\log p(z|x)]$$

$$= E_{z \sim q(z)} [\log q(z)] - E_{z \sim q(z)} [\log p(x, z)] + E_{z \sim q(z)} [\log p(x)]$$

$$= E_q [\log q(z)] - E_q [\log p(x, z)] + \log p(x)$$

- ELBO

对数似然

当 $q(z) = p(z|x; \theta)$ 时取等

$$\Rightarrow KL(q(z) \parallel p(z|x)) \geq 0, \text{ 故 } \log p(x) \geq \textcircled{3} \text{ ELBO}$$

$$\Rightarrow ELBO = E_q [\log p(x|z)] + E_q [\log p(z)] - E_q [\log q(z)]$$

$$= E_q [\log p(x|z)] + \int_z q(z) \log \frac{p(z)}{q(z)} dz$$

$$= E_q [\log p(x|z)] - KL(q(z) \parallel p(z))$$

在数据集上， $\ell(\theta, D) = \sum_{x \in D} \log p(x_i; \theta) \geq \sum_{x \in D} L(x_i; \theta, \phi)$

因此优化目标为 $\max \ell(\theta, D) \geq \max_{\theta, \phi_1, \phi_2, \dots} \sum_{x \in D} L(x_i; \theta, \phi_i)$

由于 $L(x_i; \theta, \phi) = E_q [\log p(x, z) - \log q(z)]$

$\nabla_\theta L$ 和 $\nabla_\phi L$ 是不可解析求出的，需使用 Monte Carlo 采样估计

日期： /

$$E_{q(z; \phi)} [\log p(z, x; \theta) - \log q(z; \theta)] \approx \frac{1}{k} \sum_k \log p(z^k, x; \theta) - \log q(z^k; \phi)$$

$$\text{那么 } \nabla_{\phi} E_{q(z; \phi)} [\log p(z, x; \theta) - \log q(z; \phi)] = E_{q(z; \phi)} [\nabla_{\phi} \log p(z, x; \theta)] \\ \approx \frac{1}{k} \sum_k \nabla_{\phi} \log p(z^k, x; \theta)$$

∇_{ϕ} 更加复杂，因为采样空间本身依赖 ϕ ，而采样过程是不可微的。当选择 $q_\phi(z)$ 为正态分布或特定连续分布时，一种可能的方式是重参数化。

以采样 $\epsilon \sim N(0, 1)$ 并计算 $z = \mu + \epsilon \sigma$ 代替

采样 $z \sim N(\mu, \sigma)$ ，则 $E_{z \sim q_\phi} [r(z)] = E_{\epsilon \sim N(0, 1)} [r(g(\epsilon; \phi))]$

则 $\nabla_{\phi} E_{q(z; \phi)} [r(z)] = E_{\epsilon} [\nabla_{\phi} r(g(\epsilon; \phi))]$
此即采样不依赖于 ϕ

使用过于复杂的解码器 $p_\theta(x|z)$ 可能导致 $q_\phi(z|x)$ 过于接近先验分布 $p(z)$ 以至于不能充分利用 x 中的信息