笔记：深度变分信息瓶颈*

李宇豪†

日期：2023 年 2 月 26 日

1 自编码器（AutoEncoder, AE）

自动编码器是一种无监督的神经网络模型，它可以学习到输入数据的隐含特征，称为编码 (coding)，同时用学习到的新特征可以重构出原始输入数据，称之为解码 (decoding)。

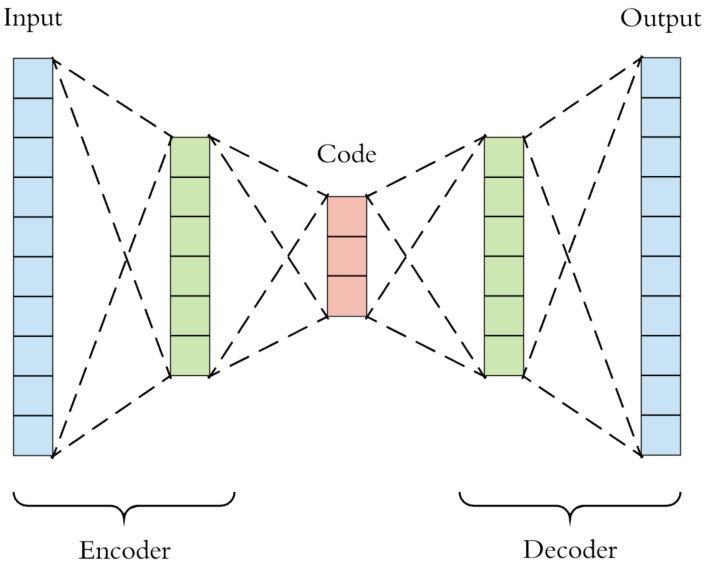

图 1: 自编码器的结构

自编码器的结构如图 1 所示，编码过程可以表示为

$$\boldsymbol{h} = g_1(\boldsymbol{X}) = \sigma(\boldsymbol{W}_1\boldsymbol{X} + \boldsymbol{b}_1) \tag{1}$$

解码的过程可以表示为

$$\boldsymbol{X}^R = g_2(\boldsymbol{h}) = \sigma(\boldsymbol{W}_2\boldsymbol{h} + \boldsymbol{b}_2) \tag{2}$$

模型的优化目标为

$$\min \operatorname{dist}(\boldsymbol{X}, \boldsymbol{X}^R) \tag{3}$$

*Key Words Talking，PMI 组会，2023.2.25

†liyh536@mail2.sysu.edu.cn

2 变分自编码器（Variational Auto-Encoders, VAE）

2.1 模型概述

变分自编码器是 Kingma 等人于 2014 年提出的基于变分贝叶斯（Variational Bayes）推断的生成式网络结构。标准的自编码器的 Encoder 产生的是一个确定的、由单值组成的向量 $\boldsymbol{h}$，因而 Decoder 只能基于 $\boldsymbol{h}$ 重构出原有的输入数据，不能生成任意的未知的数据。变分自编码器通过对编码器增加约束，使其产生服从高斯分布的潜在变量，然后从这个分布中随机采样产生特征，再进行解码。用 $\boldsymbol{X}$ 表示我们要训练的数据，并假设其是由不可观测的隐变量 $\boldsymbol{z}$ 生成的。$\boldsymbol{Z} \to \boldsymbol{X}$ 的生成模型即解码器，为似然分布 $p(x|z)$；编码器为后验分布 $p(z|x)$。

这里还存在一个问题：对于 $\boldsymbol{X}$ 中的一个样本 x_k，编码后表示为 z 服从的高斯分布，解码器从高斯分布中采样得到特征后进行还原，那么如何保证还原出来的 x'_k 是与 x_k 对应的呢？为了解决这个问题，解码器不能从先验分布 $p(z)$ 中采样，而是应该从后验分布 $p(z|x)$ 中采样，即为 $\boldsymbol{X}$ 中的每一个样本 x_k 构造一个分布 $p(z|x_k)$，然后从 $p(z|x_k)$ 中采样进行解码得到 x'_k。变分自编码器的完整结构如图 2 所示。

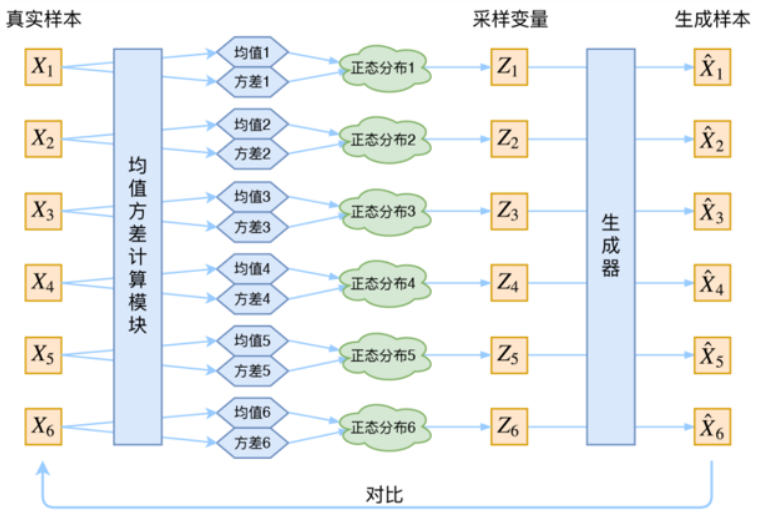

图 2: 变分自编码器的完整结构

Bayesian 公式给出了先验、似然和后验分布之间的关系

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)} \tag{4}$$

但是使用式（4）求解后验分布 $p(z|x)$ 的困难在于，$p(x)$ 是未知的，并且在 Bayesian 推断中难以求解[1]。因此接下来我们采用变分推断的近似方法来求 $p(z|x)$[2]。

[1]解析求解的方法是 $\int p(x,z)\mathrm{d}z$，对于一个高维向量 $\boldsymbol{z}$，这个多重积分的时间复杂度是指数级的 $O(k^n)$。

[2]从另一个角度来说，对于一个难以求解的分布 $p(z|x)$，我们可以使用神经网络进行拟合。但是在神经网络的训练过程中也要考虑损失函数的形式（如果选取的损失函数不好，可能会使网络在训练过程中退化为 AE），因此我们直接从数学的角度来考虑。事实上，使用神经网络拟合采用的损失函数与使用变分推断得到的优化目标是相同的，这也是 VAE 中编码器又被称为“推断网络”的原因。

2.2 变分推断（Variational Inference, VI）

变分推断是贝叶斯统计中常用的两种后验分布的推断方法之一。[3] 尝试使用一个可解的分布 $q(z|x)$ 来近似 $p(z|x)$，并用 KL 散度来衡量二者之间的差异。二者的 KL 散度为

$$\begin{aligned}\mathrm{KL}\left[q(z|x)||p(z|x)\right] &\equiv \int q(z|x)\log\frac{q(z|x)}{p(z|x)} \\ &= \int q(z|x)\log\frac{q(z|x)}{\frac{p(x|z)p(z)}{p(x)}} \\ &= \int q(z|x)\log q(z|x)\mathrm{d}z + \int q(z|x)\log p(x)\mathrm{d}z - \int q(z|x)\log[p(x|z)p(z)]\mathrm{d}z \\ &= \log p(x) + \int q(z|x)\log q(z|x)\mathrm{d}z - \int q(z|x)\log[p(x|z)p(z)]\mathrm{d}z\end{aligned} \tag{5}$$

其中，第三行到第四行利用了 $\int q(z|x)\mathrm{d}z$。我们的目标是使 KL 散度最小化，对于给定的训练集，$\log p(x)$ 是固定的，优化时不需要考虑。我们的优化目标为式（5）的后两项，记为

$$\begin{aligned}L &\equiv \int q(z|x)\log q(z|x)\mathrm{d}z - \int q(z|x)\log[p(x|z)p(z)]\mathrm{d}z \\ &= \int q(z|x)\log\frac{q(z|x)}{p(z)}\mathrm{d}z - \int q(z|x)\log p(x|z)\mathrm{d}z \\ &= \mathrm{KL}\left[q(z|x)||p(z)\right] - \mathbb{E}_{z\sim q(z|x)}\left[p(z|x)\right]\end{aligned} \tag{6}$$

因此使用 $q(z|x)$ 来近似 $p(z|x)$ 的优化目标为

$$\min L = \max \mathbb{E}_{z\sim q(z|x)}\left[p(z|x)\right] \quad \text{and} \quad \min \mathrm{KL}\left[q(z|x)||p(z)\right] \tag{7}$$

如果假设先验分布 $p(z)$ 为标准高斯分布 $p(z) \sim \mathcal{N}(0,1)$，后验分布 $p(z|x)$ 为高斯分布 $p(z|x) \sim \mathcal{N}(\mu, \sigma^2)$。两个正态分布的 KL 散度为（见附录 A.3 的推导）

$$\mathrm{KL}\left[q(z|x)||p(z)\right] = \frac{1}{2}\sum_{j=1}^{J}\left[1 + \log\left((\sigma_j)^2\right) - (u_j)^2 - (\sigma_j)^2\right] \tag{8}$$

求解对数似然期望也是一个麻烦事，因此使用 MC 将其等价为[4]

$$\mathbb{E}_{z\sim q(z|x)}\left[p(z|x)\right] \approx \frac{1}{L}\sum_{i=1}^{L}\log p(x|z^{(l)}) \tag{9}$$

其中，$z^{(l)} \sim q(z|x)$。

通过优化式（8）（9），可以得到最优的 $q_\Phi(z|x)$，作为 $p(z|x)$ 的一个很好的近似，这就是推断网络（即编码器）的作用。

3 信息瓶颈（Information Bottleneck, IB）

所谓的信息瓶颈，是这样一种朴素的思想：面对一个任务，我们应当试图用最少的信息来完成它。最少的信息意味着最低的成本，同时也意味着完成这个任务的模型找到了一些普适的规律和特性，因而泛化能力更好。

[3] 另一种是马尔可夫链蒙特卡罗方法（Markov Chain Monte Carlo, MCMC）

[4] 至于怎么做到的……

对于如下的一个输入为 X，中间层（隐藏特征）为 Z，输出为 Y 的神经网络：

$$X \to Z \to Y$$

好的模型或算法意味着 Z 是尽可能地对 X 的压缩，并且尽可能包含 Y 的最大信息。用互信息表示两个变量之间的相关性，则信息瓶颈的数学表示为

$$\max I(Z,Y;\theta) \quad \text{s.t. } \min I(X,Z;\theta) \tag{10}$$

其中，θ 是神经网络的超参数。引入拉格朗日乘子 β，可以将式（10）改写为

$$R_{IB} = I(Z,Y;\theta) - \beta I(X,Z;\theta) \tag{11}$$

R_{IB} 即所谓的信息瓶颈，模型的优化目标就是 $\max R_{IB}$。

4 变分信息瓶颈（Variational Information Bottleneck, VIB）

两个互信息分别展开为

$$I(Z,Y) = \iint p(z,y)\log\frac{p(z,y)}{p(z)p(y)}\mathrm{d}z\mathrm{d}y = \iint p(z,y)\log\frac{p(y|z)}{p(y)}\mathrm{d}y\mathrm{d}z \tag{12}$$

$$I(X,Z) = \iint p(x,z)\log\frac{p(x,z)}{p(x)p(z)}\mathrm{d}x\mathrm{d}z = \iint p(x,z)\log\frac{p(z|x)}{p(z)}\mathrm{d}x\mathrm{d}z \tag{13}$$

4.1 $I(Z,Y)$ 的变分近似

假设 X,Y,Z 满足如下的 Markov 链：

$$Y \leftrightarrow X \leftrightarrow Z$$

则联合概率分布 $p(x,y,z)$ 可以分解为[5]

$$p(x,y,z) = p(x)p(y|x)p(z|x) \tag{14}$$

考虑到 $p(y|z)$ 并不容易直接计算，设其变分近似为 $q(y|z)$，则由 $\mathrm{KL}[p(y|z),q(y|z)] \geq 0$ 可得

$$\int p(y|z)\log p(y|z)\mathrm{d}y \geq \int p(y|z)\log q(y|z)\mathrm{d}y \tag{15}$$

因此

$$\begin{aligned}
I(Z,Y) &= \iint p(z,y)\log\frac{p(y|z)}{p(y)}\mathrm{d}y\mathrm{d}z \\
&\geq \iint p(z,y)\log\frac{q(y|z)}{p(y)}\mathrm{d}y\mathrm{d}z \\
&= \iint p(z,y)\log q(y|z)\mathrm{d}y\mathrm{d}z - \int\left[\int p(z,y)\mathrm{d}z\right]\log p(y)\mathrm{d}y \\
&= \iint p(z,y)\log q(y|z)\mathrm{d}y\mathrm{d}z - \int p(y)\log p(y)\mathrm{d}y \\
&= \iint p(z,y)\log q(y|z)\mathrm{d}y\mathrm{d}z + H(y)
\end{aligned} \tag{16}$$

其中 $H(y)$ 是标签 y 的熵，与优化过程无关，可以忽略。

[5] why?

结合式（14）（16）可得 $I(Z,Y)$ 的下限为

$$\begin{aligned} I(Z,Y) &\geq \iint p(z,y)\log q(y|z)\mathrm{d}y\mathrm{d}z \\ &= \iint \left[\int p(x,y,z)dx\right]\log q(y|z)\mathrm{d}y\mathrm{d}z \\ &= \iiint p(x,y,z)\log q(y|z)\mathrm{d}x\mathrm{d}y\mathrm{d}z \\ &= \iiint p(x)p(y|x)p(z|x)\log q(y|z)\mathrm{d}x\mathrm{d}y\mathrm{d}z \end{aligned} \tag{17}$$

4.2 $I(X,Z)$ 的变分近似

$$I(X,Z) = \iint p(x,z)\log\frac{p(z|x)}{p(z)}\mathrm{d}x\mathrm{d}z = \iint p(x,z)\log p(z|x)\mathrm{d}x\mathrm{d}z - \iint p(x,z)\log p(z)\mathrm{d}x\mathrm{d}z \tag{18}$$

其中，$p(z)$ 同样不易计算，设 $p(z)$ 的变分近似为 $r(z)$，由 $\mathrm{KL}[p(z),r(z)]\geq 0$ 得

$$\int p(z)\log p(z)\mathrm{d}z \geq \int p(z)\log r(z)\mathrm{d}z \tag{19}$$

即

$$\begin{aligned} \int\left[\int p(x,z)\mathrm{d}x\right]\log p(z)\mathrm{d}z &\geq \int\left[\int p(x,z)\mathrm{d}x\right]\log r(z)\mathrm{d}z \\ \iint p(x,z)\log p(z)\mathrm{d}x\mathrm{d}z &\geq \iint p(x,z)\log r(z)\mathrm{d}x\mathrm{d}z \end{aligned} \tag{20}$$

因此 $I(X,Z)$ 的上限为

$$\begin{aligned} I(X,Z) &= \iint p(x,z)\log p(z|x)\mathrm{d}x\mathrm{d}z - \iint p(x,z)\log p(z)\mathrm{d}x\mathrm{d}z \\ &\leq \iint p(x,z)\log p(z|x)\mathrm{d}x\mathrm{d}z - \iint p(x,z)\log r(z)\mathrm{d}x\mathrm{d}z \\ &= \iint p(x,z)\log\frac{p(z|x)}{r(z)}\mathrm{d}x\mathrm{d}z \\ &= \iint p(x)p(z|x)\log\frac{p(z|x)}{r(z)}\mathrm{d}x\mathrm{d}z \end{aligned} \tag{21}$$

4.3 变分信息瓶颈

结合式（17）（21）可得在变分近似下信息瓶颈 R_{IB} 的下限，

$$\begin{aligned} R_{IB} &= I(Z,Y;\theta) - \beta I(X,Z;\theta) \\ &\leq \iiint p(x)p(y|x)p(z|x)\log q(y|z)\mathrm{d}x\mathrm{d}y\mathrm{d}z - \beta\iint p(x)p(z|x)\log\frac{p(z|x)}{r(z)}\mathrm{d}x\mathrm{d}z \equiv R_{VIB} \end{aligned} \tag{22}$$

R_{VIB} 即变分信息瓶颈。在实际计算时，一般用经验分布 $p(x,y)=\frac{1}{N}\sum_{i=1}^{N}\delta_{x_n}(x)\delta_{y_n}(y)$ 来代替 $p(x)p(y|x)=p(x,y)$，因此 R_{VIB} 改写为

$$\begin{aligned} R_{VIB} &= \frac{1}{N}\sum_{i=1}^{N}\int\left[p\left(z|x_n\right)\log q\left(y_n|z\right) - \beta p\left(z|x_n\right)\log\frac{p\left(z|x_n\right)}{r(z)}\right]\mathrm{d}z \\ &= \frac{1}{N}\sum_{i=1}^{N}\int\left[p\left(z|x_n\right)\log q\left(y_n|z\right) - \beta\mathrm{KL}[p\left(z|x_n\right),r(z)]\right]\mathrm{d}z \end{aligned} \tag{23}$$

假设编码器 $f(z|x)$ 具有类似 VAE 的结构 $p(z|x) = \mathcal{N}\left(z|f_e^{\mu}(x), f_e^{\Sigma}(x)\right)$，则可以利用重参数化技巧得到

$$p(z|x)\mathrm{d}x = p(\epsilon)\mathrm{d}\epsilon \tag{24}$$

因此实际计算时最大化 R_{VIB} 可以转为最小化

$$J_{IB} = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{\epsilon\sim p(\epsilon)}\left[-\log q\left(y_n \mid f\left(x_n, \epsilon\right)\right)\right] + \beta\,\mathrm{KL}\left[p\left(z \mid x_n\right) \mid r(z)\right] \tag{25}$$

附录

A KL 散度

A.1 定义

这肯定都会吧

A.2 性质

这肯定都知道吧

A.3 两个正态分布的 KL 散度

这个不知道可以上网查

B 互信息

这个也肯定都知道

C Markov 链

这个大概知道是个啥就行

D 重参数化技巧

这个以后再说吧