变分推断与变分自编码器

作者：凯鲁嘎吉 - 博客园 http://www.cnblogs.com/kailugaji/

本文主要介绍变分自编码器(Variational Auto-Encoder, VAE)及其推导过程，但变分自编码器涉及一些概率统计的基础知识，因此为了更好地理解变分自编码器，首先介绍变分推断(Variational Inference)与期望最大化(Expectation-Maximization, EM)算法，进而介绍变分自编码器，并给出另一种理解方法(参考文献[3])。

1. 变分推断

变分推断

- 参数估计
  - 根据样本所提供的信息，对总体分布中的未知参数θ进行估值
- 贝叶斯估计

极大似然估计

最大后验估计

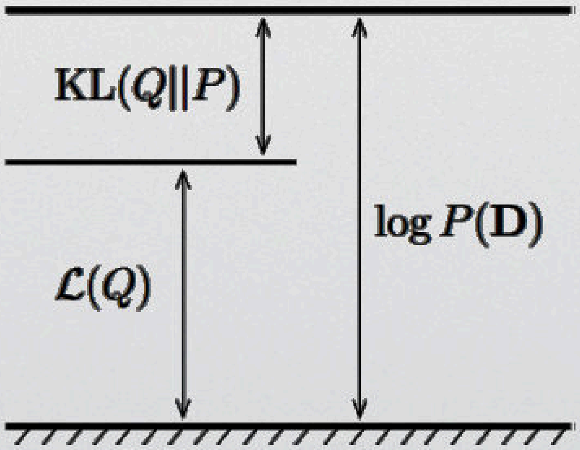
$$p(\theta \mid X)=\frac{p(X \mid \theta)p(\theta)}{p(X)}=\frac{\boxed{\boxed{p(X \mid \theta)}\,p(\theta)}}{\int p(X,\theta)d\theta}$$

- 贝叶斯估计中分母p(X)往往很难求，于是找一个简单的函数 $q(\theta)\approx p(\theta|X)$
- 如何评价$q(\theta)$与$p(\theta|X)$之间的近似程度呢？——Kullback-Leibler散度
- 目标函数：$\min KL(q(\theta)||p(\theta|X))$

变分推断

- 目标函数：min $KL(q(\theta)\|p(\theta|X))$

$$KL(q\|p)=\int q(\theta)\ln\frac{q(\theta)}{p(\theta|X)}d\theta$$
$$=\int q(\theta)\ln\frac{q(\theta)}{p(X,\theta)}d\theta+\ln p(X)$$



$$\ln p(X)=KL(q\|p)+\int q(\theta)\ln\frac{p(X,\theta)}{q(\theta)}d\theta=KL(q\|p)+L(q)$$

- $\ln p(X)=KL(q\|p)+L(q)$， 而$\ln p(X)$是与θ无关的常量，不变。
- min $KL(q\|p)$ ⇔ max $L(q)$，变分贝叶斯学习通过$q(\theta)$的迭代实现$L(q)$的最大化

$$\max L(q)=\int q(\theta)\ln p(X,\theta)d\theta-\int q(\theta)\ln q(\theta)d\theta$$

变分推断

➢平均场理论

- 根据平均场理论，变分分布q(θ)可以因式分解为M个互不相交的部分

$$q(\theta)=\prod_{i=1}^{M}q_i(\theta_i)$$

$$\max L(q)=\int q\ln p(X,\theta)d\theta-\int q\ln q d\theta=\int q_j\left\{\int \ln p(X,\theta)\prod_{i\neq j}q_i d\theta_i\right\}d\theta_j-\int q_j\ln q_j d\theta_j+c$$

$$=\int q_j\ln\tilde{p}(X,\theta_j)d\theta_j-\int q_j\ln q_j d\theta_j+c=-KL(q_j\,\|\,\tilde{p}(X,\theta_j))+c$$

其中$\ln\tilde{p}(X,\theta_j)=E_{i\neq j}[\ln p(X,\theta)]+c=\int \ln p(X,\theta)\prod_{i\neq j}q_i d\theta_i+c$

$q_j=\tilde{p}(X,\theta_j),L(q)$最大　　$\therefore \ln q_j^*=\ln\tilde{p}(X,\theta_j)=E_{i\neq j}[\ln p(X,\theta)]+c$

$$\boxed{q_j^*(\theta_j)=\frac{\exp\left(E_{i\neq j}[\ln p(X,\theta)]\right)}{\int \exp\left(E_{i\neq j}[\ln p(X,\theta)]\right)d\theta_j}}$$

变分推断

➢EM算法

$$\ln p(X;\theta) = KL(q(Z) \| p(Z \mid X;\theta)) + \int q(Z) \ln \frac{p(X,Z;\theta)}{q(Z)} dZ$$
$$= KL(q(Z) \| p(Z \mid X;\theta)) + L(q, X;\theta)$$

- E-step: 固定θ, 求q(Z)

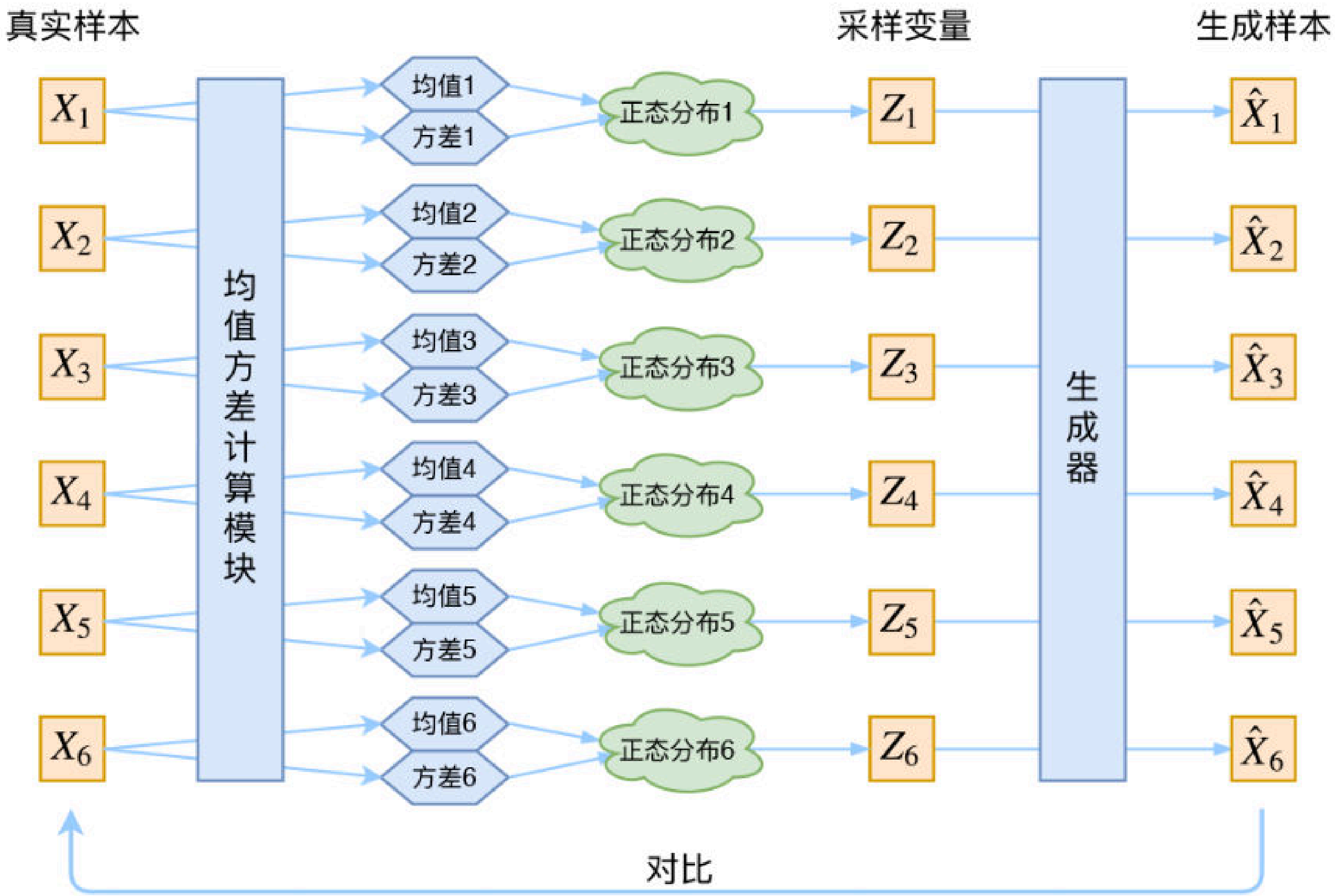
$$q_{t+1}(Z) = \arg\max_{q} L(q, X;\theta_t)$$

✓ 若p(Z|X; θ)好求，则q(Z)= p(Z|X; θ)
✓ 否则，用变分推断近似估计q(Z)

- M-step: 固定q(Z), 求θ

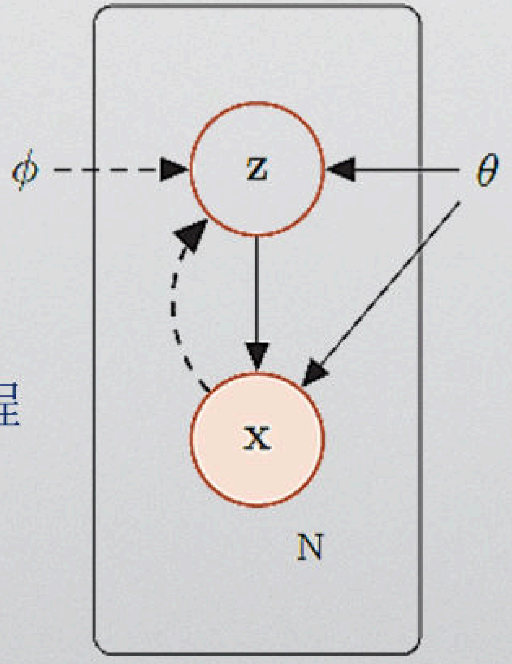
$$\theta_{t+1} = \arg\max_{\theta} L(q_{t+1}, X;\theta)$$

2．变分自编码器


真实样本
采样变量
生成样本
X_1
X_2
X_3
X_4
X_5
X_6
均值方差计算模块
均值1
方差1
均值2
方差2
均值3
方差3
均值4
方差4
均值5
方差5
均值6
方差6
正态分布1
正态分布2
正态分布3
正态分布4
正态分布5
正态分布6
Z_1
Z_2
Z_3
Z_4
Z_5
Z_6
生成器
$\hat{X}_1$
$\hat{X}_2$
$\hat{X}_3$
$\hat{X}_4$
$\hat{X}_5$
$\hat{X}_6$
对比

变分自编码器

- 深度生成模型
  - 就是利用神经网络来建模条件分布p(x|z;θ)。
  - 对抗生成式网络（Generative Adversarial Network，GAN）
  - 变分自编码器（Variational Autoencoder，VAE)
- 生成模型
  - 指一系列用于随机生成可观测数据的模型。生成数据x的过程可以分为两步进行：
  - 根据隐变量的先验分布p(z;θ)进行采样，得到样本z;
  - 根据条件分布p(x|z;θ)进行采样，得到x。

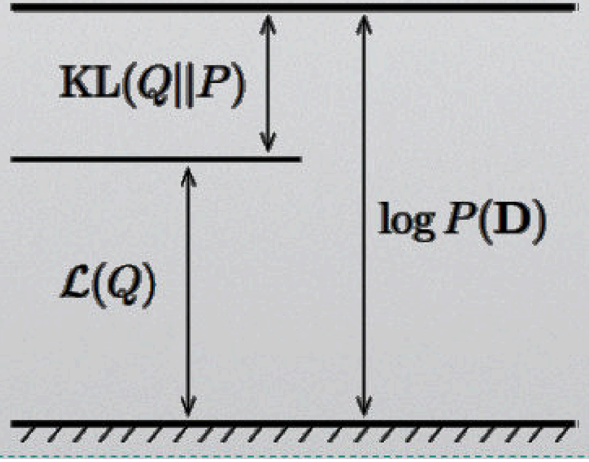


$$\boxed{p(X)=\int p(X,Z)dZ=\int p(X\mid Z)p(Z)dZ}$$

变分自编码器

- 给定一个样本x，其对数边际似然log p(x;θ)可以分解为

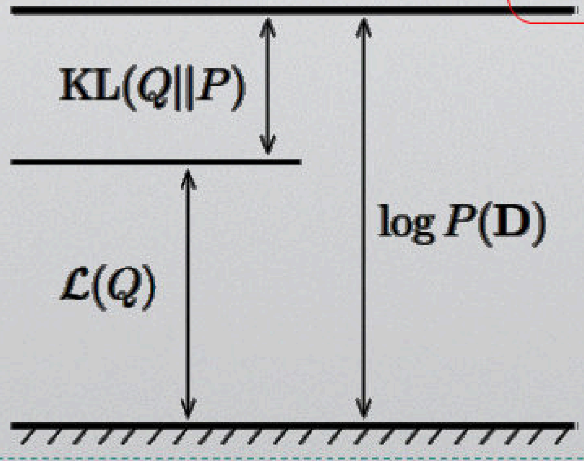
$$
\begin{aligned}
\log p(\boldsymbol{x};\theta) &= \int q(z;\phi)\log\frac{p(\boldsymbol{x},z;\theta)}{q(z;\phi)}dz - \int q(z;\phi)\log\frac{p(z\mid\boldsymbol{x};\theta)}{q(z;\phi)}dz \\
&= L(q,\boldsymbol{x};\theta,\phi) + KL(q(z;\phi)\,\|\,p(z\mid\boldsymbol{x};\theta)) \\
&= E_{z\sim q(z;\phi)}\left[\log\frac{p(\boldsymbol{x},z;\theta)}{q(z;\phi)}\right] + KL(q(z;\phi)\,\|\,p(z\mid\boldsymbol{x};\theta)) \\
&= E_{z\sim q(z;\phi)}\left[\log\frac{p(\boldsymbol{x}\mid z;\theta)p(z;\theta)}{q(z;\phi)}\right] + KL(q(z;\phi)\,\|\,p(z\mid\boldsymbol{x};\theta))
\end{aligned}
$$



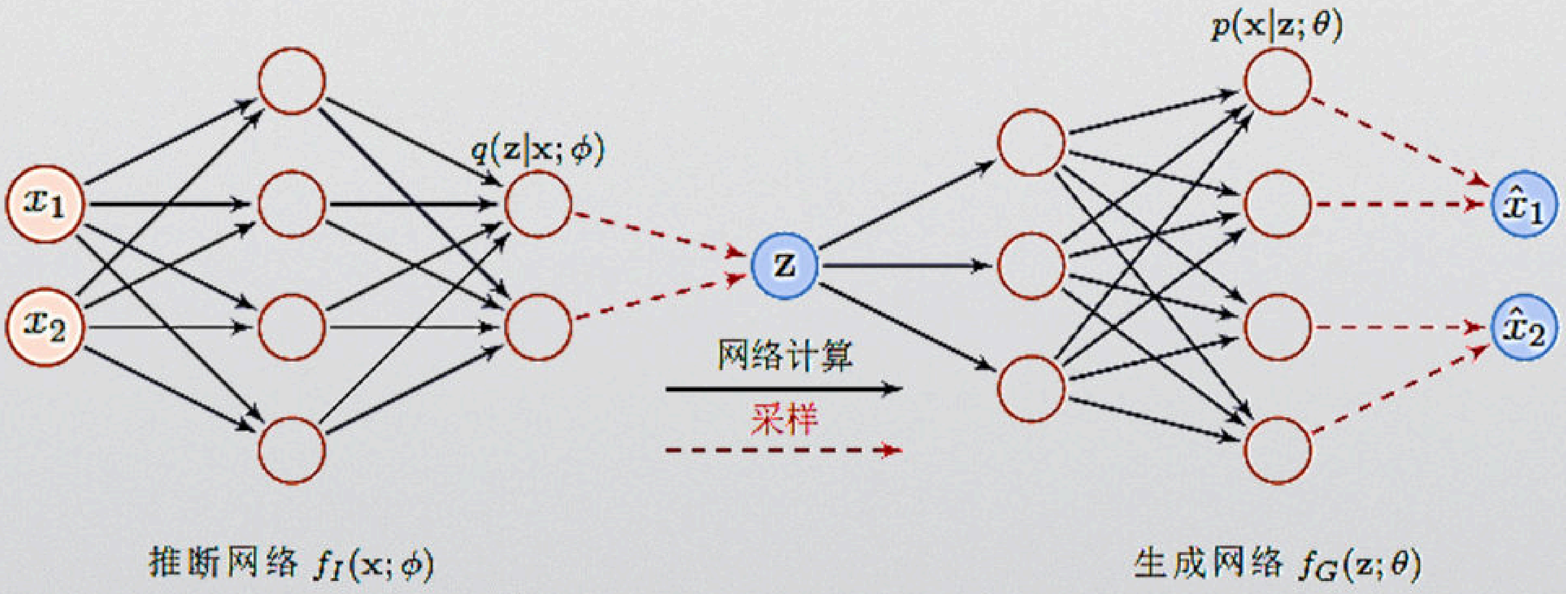
变分自编码器

- 变分自编码器目标函数：

$$
\begin{aligned}
&\max_{\theta,\phi} L(q,\boldsymbol{x};\theta,\phi) \\
&= E_{z\sim q(z;\phi)}\left[\log \frac{p(\boldsymbol{x}\mid \boldsymbol{z};\theta)\,p(\boldsymbol{z};\theta)}{q(\boldsymbol{z};\phi)}\right] \\
&= \underbrace{E_{z\sim q(z\mid \boldsymbol{x};\phi)}\left[\log p(\boldsymbol{x}\mid \boldsymbol{z};\theta)\right]}_{\text{生成}} - \underbrace{KL(q(\boldsymbol{z}\mid \boldsymbol{x};\phi)\,\|\,p(\boldsymbol{z};\theta))}_{\text{推断}}
\end{aligned}
$$



变分自编码器



- 变分自编码器的模型结构可以分为两个部分：
  - ✓ 寻找后验分布p(z|x;θ)的变分近似q(z|x;ϕ*)(即: q(z;ϕ*));
    - 变分推断：用简单的分布q去近似复杂的分布p(z|x;θ)
  - ✓ 在已知q(z|x;ϕ*)的情况下，估计更好的生成p(x|z;θ)。

变分自编码器

- 变分自编码器的模型结构分为两个部分：
  - ✓ 推断网络：q(z|x;ϕ)尽可能接近p(z|x;θ) (不用平均场理论，而是直接假设分布，用神经网络训练参数)

$$\phi^* = \arg\min_{\phi} KL(q(\boldsymbol{z} \mid \boldsymbol{x};\phi) \,\|\, p(\boldsymbol{z} \mid \boldsymbol{x};\theta)) = \operatorname*{argmax}_{\phi} L(q, \boldsymbol{x};\theta,\phi)$$

$$\boxed{q(\boldsymbol{z} \mid \boldsymbol{x};\phi) = \mathcal{N}(\boldsymbol{z};\boldsymbol{\mu}_I, \sigma_I^2 \boldsymbol{I})}$$

$$\boldsymbol{h} = \sigma(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}), \boldsymbol{\mu}_I = \boldsymbol{W}^{(2)}\boldsymbol{h} + \boldsymbol{b}^{(2)}, \sigma_I = softplus(\boldsymbol{W}^{(3)}\boldsymbol{h} + \boldsymbol{b}^{(3)})$$

  - ✓ 生成网络：p(x, z;θ)=p(x|z;θ)p(z;θ)

$$\theta^* = \operatorname*{argmax}_{\theta} L(q, \boldsymbol{x};\theta,\phi)$$

$$p(\boldsymbol{z};\theta) = \mathcal{N}(\boldsymbol{z};\boldsymbol{0},\boldsymbol{I})$$

  - ✓ 总体目标函数：

$$\max_{\theta,\phi} L(q,\boldsymbol{x};\theta,\phi) = \max_{\theta,\phi} E_{\boldsymbol{z}\sim q(\boldsymbol{z};\phi)}\left[\log \frac{p(\boldsymbol{x} \mid \boldsymbol{z};\theta)p(\boldsymbol{z};\theta)}{q(\boldsymbol{z};\phi)}\right]$$

$$= \max_{\theta,\phi} E_{\boldsymbol{z}\sim q(\boldsymbol{z}|\boldsymbol{x};\phi)}[\log p(\boldsymbol{x} \mid \boldsymbol{z};\theta)] - KL(q(\boldsymbol{z} \mid \boldsymbol{x};\phi) \,\|\, p(\boldsymbol{z};\theta))$$

变分自编码器

- 总体目标函数：

$$\max_{\theta,\phi} L(q,\boldsymbol{x};\theta,\phi) = E_{z\sim q(\boldsymbol{z}|\boldsymbol{x};\phi)}\left[\log p(\boldsymbol{x}\mid \boldsymbol{z};\theta)\right] - KL(q(\boldsymbol{z}\mid \boldsymbol{x};\phi)\,\|\,p(\boldsymbol{z};\theta))$$

$$\stackrel{(1)}{\approx} \frac{1}{K}\sum_{k=1}^{K}\log p(\boldsymbol{x}\mid \boldsymbol{z}^{(k)};\theta) - KL(q(\boldsymbol{z}\mid \boldsymbol{x};\phi)\,\|\,p(\boldsymbol{z};\theta))$$

$$\stackrel{(2)}{=} \sum_{n=1}^{N}\left(\frac{1}{K}\sum_{k=1}^{K}\log p(\boldsymbol{x}^{(n)}\mid \boldsymbol{z}^{(n,k)};\theta) - KL(q(\boldsymbol{z}\mid \boldsymbol{x}^{(n)};\phi)\,\|\,N(\boldsymbol{z};\boldsymbol{0},\boldsymbol{I}))\right)$$

(1): 对于样本x, 根据$q(z|x;\phi)$采集K个z, $1\leq k\leq K$

(2): $p(\boldsymbol{z};\theta) = \mathcal{N}(\boldsymbol{z};\boldsymbol{0},\boldsymbol{I})$

一般情况下，只采一个样即可，即K=1。详见参考文献[3]。

变分自编码器

- 总体目标函数：

$$\max_{\theta,\phi} L(q,\boldsymbol{x};\theta,\phi) = \max_{\theta,\phi} E_{\boldsymbol{z}\sim q(\boldsymbol{z}|\boldsymbol{x};\phi)}\left[\log p(\boldsymbol{x}\mid\boldsymbol{z};\theta)\right] - KL(q(\boldsymbol{z}\mid\boldsymbol{x};\phi)\,\|\,p(\boldsymbol{z};\theta))$$

若$p(\boldsymbol{x}\mid\boldsymbol{z};\phi)=\mathcal{N}(\boldsymbol{z};\boldsymbol{\mu}_G,\boldsymbol{I}), q(\boldsymbol{z}\mid\boldsymbol{x};\phi)=\mathcal{N}(\boldsymbol{z};\boldsymbol{\mu}_I,\sigma_I^2\boldsymbol{I})$, 目标函数简化为：

$$\boxed{\max_{\theta,\phi}\ -\left\|\boldsymbol{x}-\boldsymbol{\mu}_G\right\|^2 - KL(N(\boldsymbol{\mu}_I,\sigma_I)\,\|\,N(\boldsymbol{0},\boldsymbol{I}))}$$

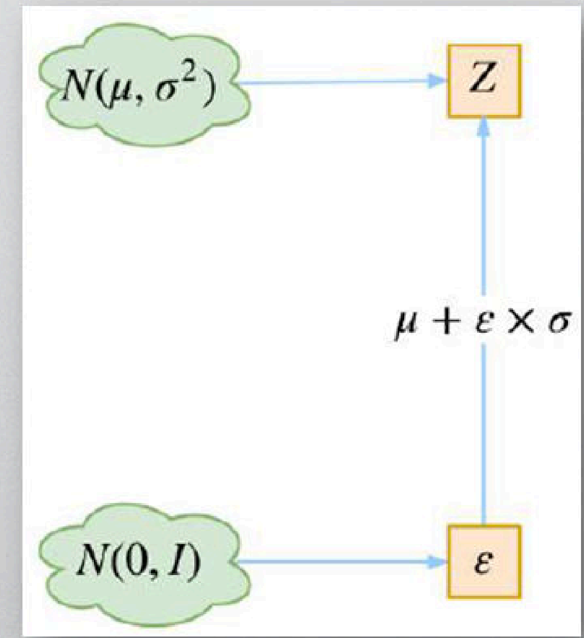
若$p(\boldsymbol{x}\mid\boldsymbol{z};\phi)=\boldsymbol{\mu}_G^x(1-\boldsymbol{\mu}_G)^{1-x}, q(\boldsymbol{z}\mid\boldsymbol{x};\phi)=\mathcal{N}(\boldsymbol{z};\boldsymbol{\mu}_I,\sigma_I^2\boldsymbol{I})$, 目标函数简化为：

$$\boxed{\max_{\theta,\phi}\ \boldsymbol{x}\log(\boldsymbol{\mu}_G)+(1-\boldsymbol{x})\log(1-\boldsymbol{\mu}_G) - KL(N(\boldsymbol{\mu}_I,\sigma_I)\,\|\,N(\boldsymbol{0},\boldsymbol{I}))}$$

变分自编码器

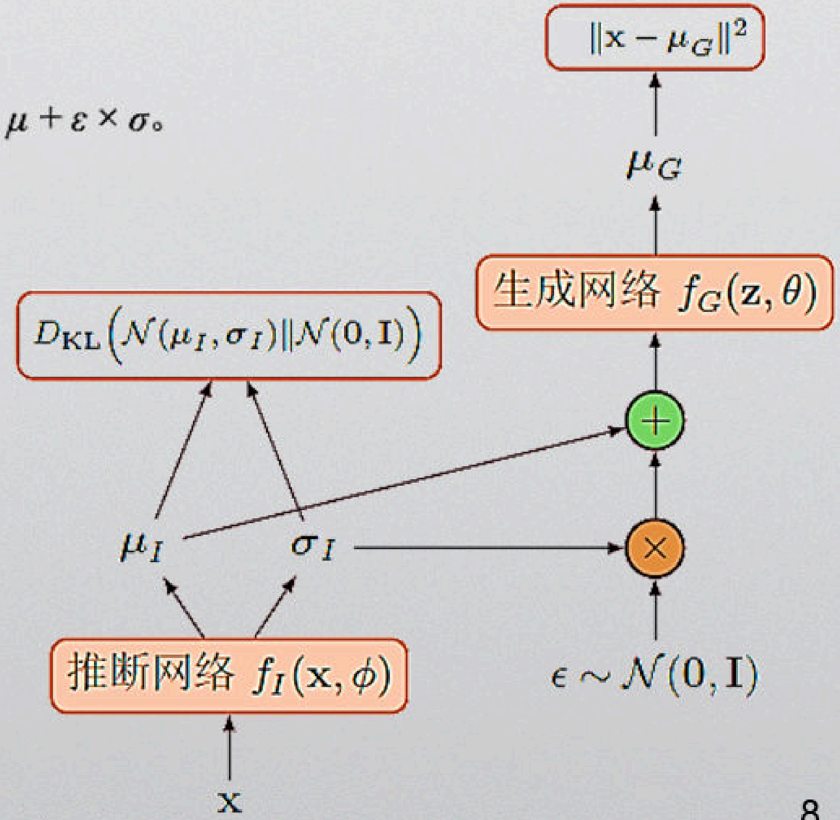
- 分布q(z|x,ϕ)依赖于参数ϕ，采样无法刻画z与ϕ函数关系，无法求导
- 重参数化（reparameterization）是实现通过随机变量实现反向传播的一种重要手段。将采样关系->函数关系。

从$\mathcal{N}(\mu,\sigma^2)$中采样一个Z，相当于从$\mathcal{N}(0,I)$中采样一个$\varepsilon$，然后让$Z=\mu+\varepsilon\times\sigma$。

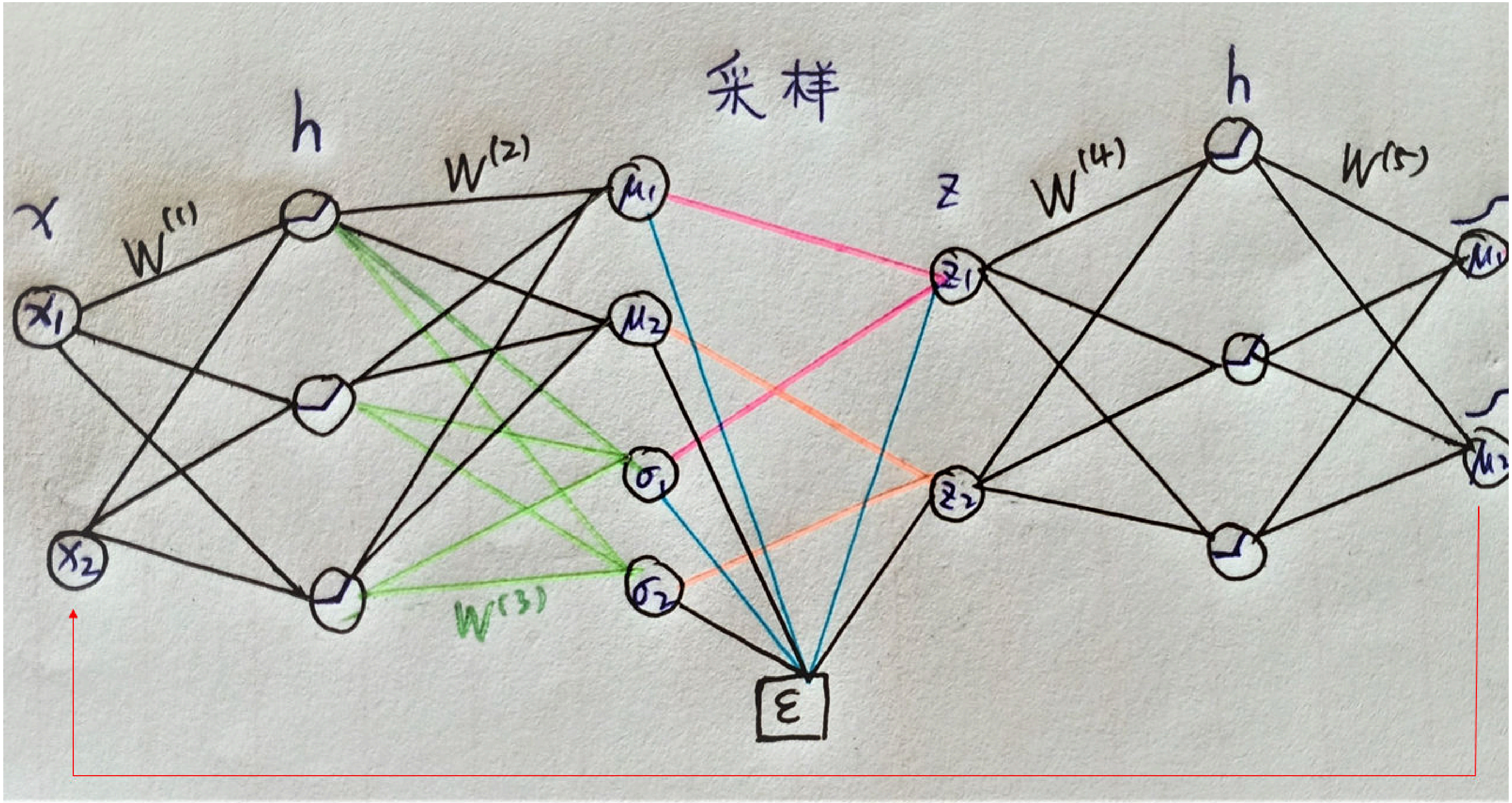


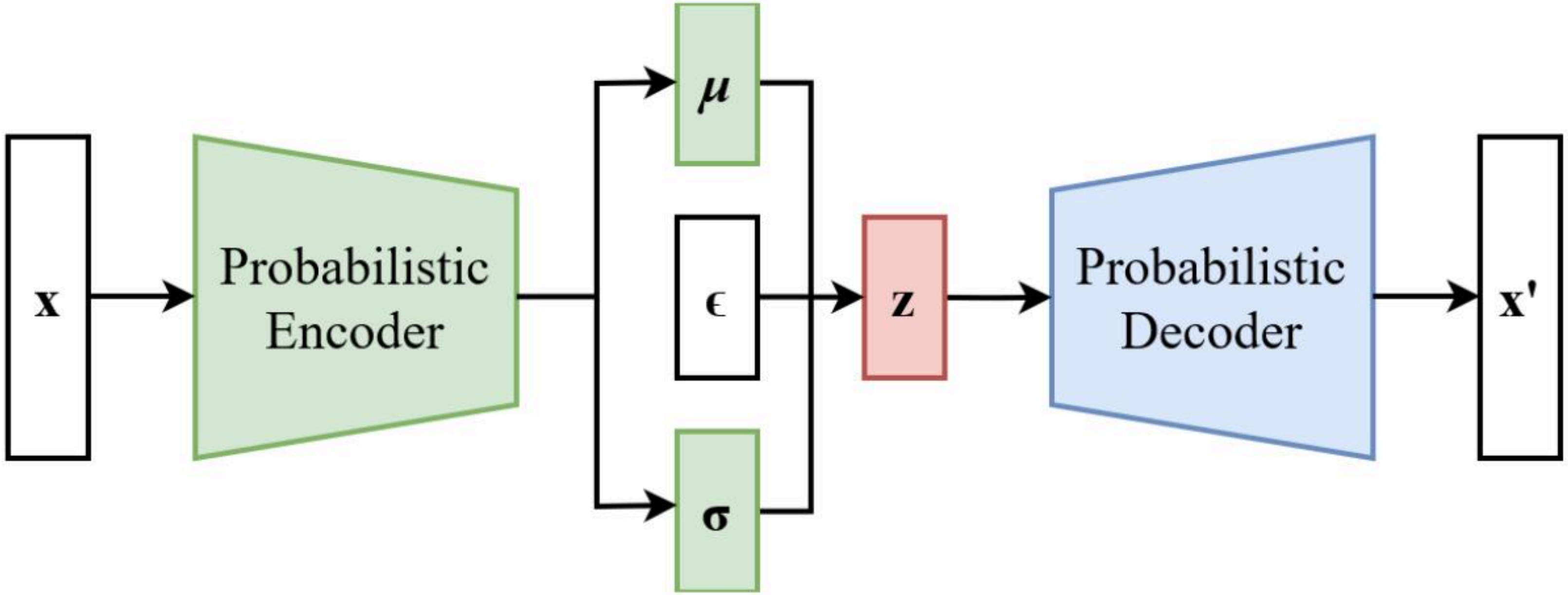
$$z \sim N(\mu,\sigma^2)$$

$$\varepsilon = \frac{z-\mu}{\sigma} \sim N(0,1)$$



代码里z是这样的来的：$z \sim \mu +\varepsilon {{e}^{0.5\ln ({{\sigma }^{2}})}}=\mu +\varepsilon \sigma $


采样
x
h
W(1)
W(2)
W(3)
W(4)
W(5)
z
ε



3. 变分自编码器另一种理解——直面联合分布

概率论基础知识

$$E(x)=\int xp(x)dx=c,x\sim p(x)$$

$$E(f(x,\xi))=\int f(x,\xi)p(x)dx=g(\xi),x\sim p(x)$$

$$E(f(x))=\int f(x)\cdot p(x)dx=c,x\sim p(x)$$

$$\int p(z\mid x)dz=1$$

变分自编码器另一种理解

- 变分自编码器另一种理解——直面联合分布

$$\boxed{p(X)=\int p(X,Z)dZ=\int p(X\mid Z)p(Z)dZ}$$

$$\boxed{q(X)=\int q(X,Z)dZ=\int q(X\mid Z)q(Z)dZ}$$

✓ 用q(X)≈p(X) ⇔ q(X,Z)≈p(X,Z)

$$KL(q(x,z)\,\|\,p(x,z))=\iint q(x,z)\log\frac{q(x,z)}{p(x,z)}dzdx$$

$$=\int q(x)\left[\int q(z\mid x)\log\frac{q(x)q(z\mid x)}{p(x,z)}dz\right]dx=E_{x\sim q(x)}\left[\int q(z\mid x)\log\frac{q(x)q(z\mid x)}{p(x,z)}dz\right]$$

$$=E_{x\sim q(x)}\left\{\int\left[q(z\mid x)\log q(x)+q(z\mid x)\log\frac{q(z\mid x)}{p(x,z)}\right]dz\right\}$$

变分自编码器另一种理解

$$KL(q(x,z)\,\|\,p(x,z)) = E_{x\sim q(x)}\left[\int q(z\mid x)\log\frac{q(z\mid x)}{p(x,z)}dz\right]+c$$

$$= E_{x\sim q(x)}\left[\int q(z\mid x)\log\frac{q(z\mid x)}{p(x\mid z)p(z)}dz\right]+c$$

$$= E_{x\sim q(x)}\left[-\int q(z\mid x)\log p(x\mid z)dz+\int q(z\mid x)\log\frac{q(z\mid x)}{p(z)}dz\right]+c$$

$$= E_{x\sim q(x)}\left[E_{z\sim q(z\mid x)}\left[-\log p(x\mid z)\right]+E_{z\sim q(z\mid x)}\left[\log\frac{q(z\mid x)}{p(z)}\right]\right]+c$$

VAE的损失函数

$$= E_{x\sim q(x)}\left[\boxed{E_{z\sim q(z\mid x)}\left[-\log p(x\mid z)\right]+KL(q(z\mid x)\,\|\,p(z))}\right]+c$$

KL(q(x, z)||p(x, z))等价于VAE的损失函数

2

4．KL散度公式推导

$\mathcal{KL}$散度公式推导

令$f(x;\mu,\sigma^2)=\dfrac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

$$KL(N(\mu,\sigma)\,\|\,N(0,I))$$

$$=\int f(x;\mu,\sigma^2)\log\left(\frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}\cdot\frac{\sqrt{2\pi}}{e^{-\frac{x^2}{2}}}\right)dx=\int f(x;\mu,\sigma^2)\log\left(\frac{1}{\sqrt{\sigma^2}}e^{\frac{x^2-\frac{(x-\mu)^2}{\sigma^2}}{2}}\right)dx$$

$$=\frac{1}{2}\int f(x;\mu,\sigma^2)\left(-\log\sigma^2+x^2-\frac{(x-\mu)^2}{\sigma^2}\right)dx\overset{(1)}{=}\frac{1}{2}\left(-\log\sigma^2+\mu^2+\sigma^2-1\right)$$

(1) 用到了正态分布的二阶矩公式，详见下一页

$\mathcal{KL}$散度公式推导

$$令 f(x;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$若 X\sim N(\mu,\sigma^2)$$

$$\because E(X)=\mu, D(X)=E(X^2)-(E(X))^2=\sigma^2$$

$$\therefore E(X^2)=\int x^2\cdot f(x;\mu,\sigma^2)dx=(E(X))^2+D(X)=\mu^2+\sigma^2$$

$$\because E(X-\mu)^n=\int (x-\mu)^n\cdot f(x;\mu,\sigma^2)dx$$

$$=\begin{cases} 0, & n \text{ is odd}, \\ \sigma^n(n-1)!!, & n \text{ is even}. \end{cases}$$

$$\therefore E(X-\mu)^2=\int (x-\mu)^2\cdot f(x;\mu,\sigma^2)dx=\sigma^2$$

5. 参考文献

[1] 变分贝叶斯 - 凯鲁嘎吉 - 博客园

[2] 邱锡鹏, 神经网络与深度学习[M]. 2019.

[3] 标签 vae 下的文章 - 科学空间|Scientific Spaces

[4] Kingma D P , Welling M . Auto-Encoding Variational Bayes[J]. 2013.

[5] 变分推断详细请参考：华俊豪博客-变分推理、变分贝叶斯算法理解与推导

[6] Tutorial - What is a variational autoencoder? – Jaan Altosaar https://jaan.io/what-is-variational-autoencoder-vae-tutorial/

[7] CS 285, Variational Inference and Generative Models, http://rail.eecs.berkeley.edu/deeprlcourse-fa20/static/slides/lec-18.pdf

[8] Ankush Ganguly, Samuel W. F. Earp, An Introduction to Variational Inference, 2021. https://arxiv.org/pdf/2108.13083.pdf