讲诉比较清楚的网站：<https://blog.csdn.net/c9Yv2cf9I06K2A9E/article/details/79675832>

**变分:**指的是泛函的变分。打个比方，从A点到B点有无数条路径，每一条路径都是一个函数吧，这无数条路径，每一条函数（路径）的长度都是一个数，那你从这无数个路径当中选一个路径最短或者最长的，这就是求泛函的极值问题。有一种老的叫法，函数空间的自变量我们称为**宗量（自变函数），**当**宗量**变化了一点点而导致了**泛函**值变化了多少，这其实就是变分

总：变分是一个过程

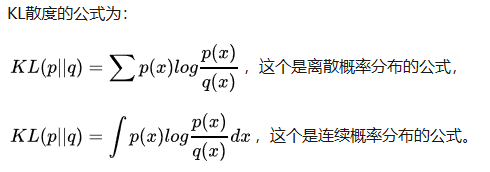
泛函就是函数的函数，因此泛函就包含了变分这个过程

下列中F(y(x))是泛函，他的自变量为y(x)为一个函数，成为宗量

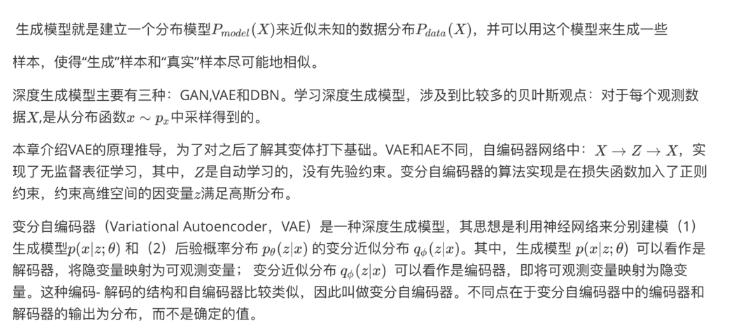
对于普通的函数f(x)，我们可以认为f是一个关于x的一个实数算子，其作用是将实数x映射到实数f(x)。那么类比这种模式，假设存在函数算子F，它是关于f(x)的函数算子，可以将f(x)映射成实数F(f(x)) 。对于f(x)我们是通过改变x来求出f(x)的极值，而在变分中这个x会被替换成一个函数y(x)，我们通过改变x来改变y(x),最后使得F(y(x))求得极值。

KL散度：

在Variational Inference中，我们希望能够找到一个相对简单好算的概率分布q，使它尽可能地近似我们待分析的后验概率p(z|x)，其中z是隐变量，x是显变量。在这里我们的“loss函数”就是KL散度，他可以很好地测量两个概率分布之间的距离。如果两个分布越接近，那么KL散度越小，如果越远，KL散度就会越大。



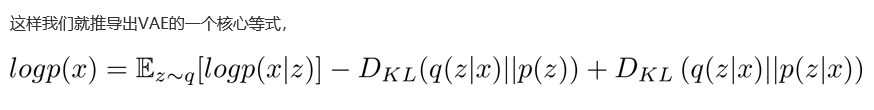
VAE概述：X->Z->X z是隐变量，x是显变量

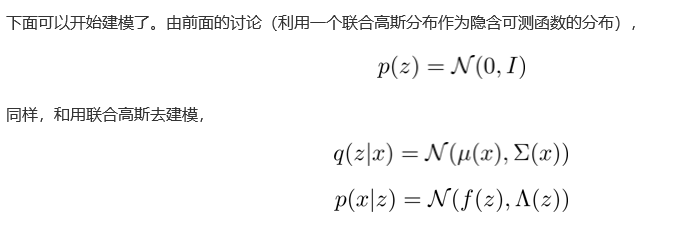


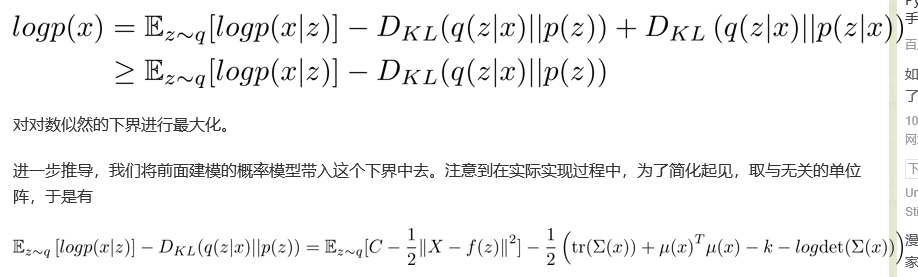
如何学习得到解码器：

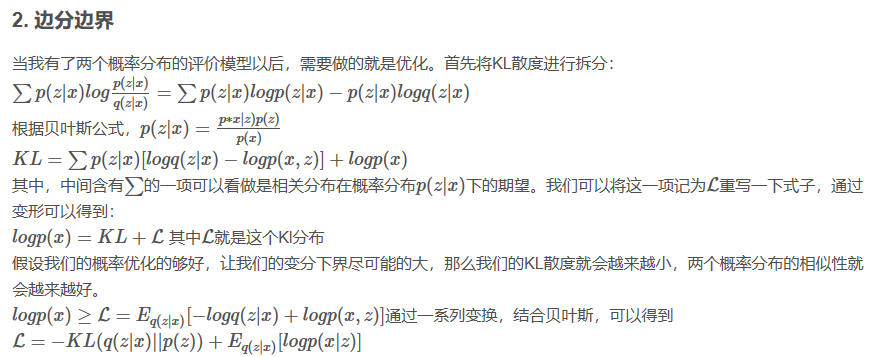
我们仅仅有些样本（学习到的参数只适合样本，并不包含总体），但希望生成一些具有多样性类似的样本。那么自然会想到利用极大似然法来估计可学习的参数（适合全局）

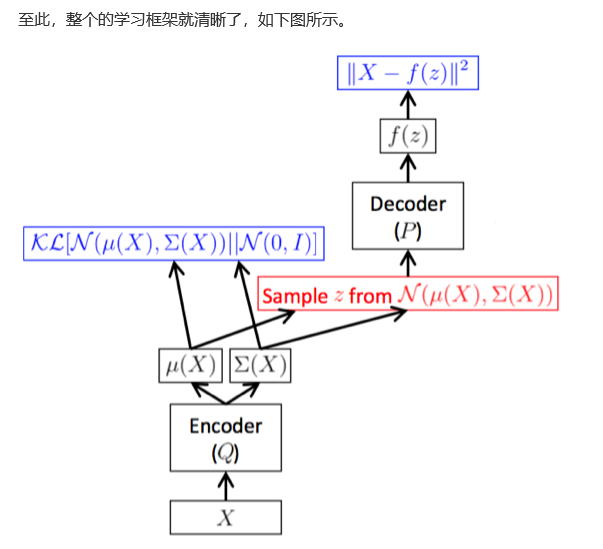
推到过程见：<https://blog.csdn.net/omnispace/article/details/73123913>



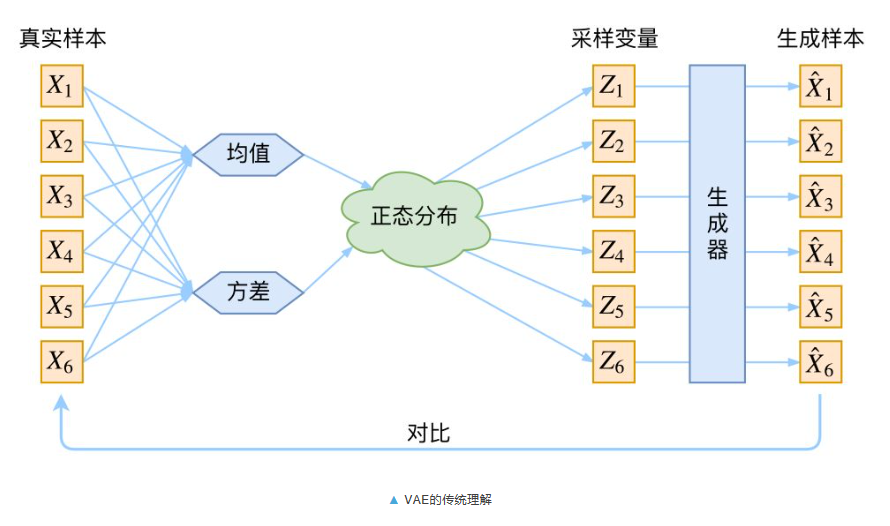




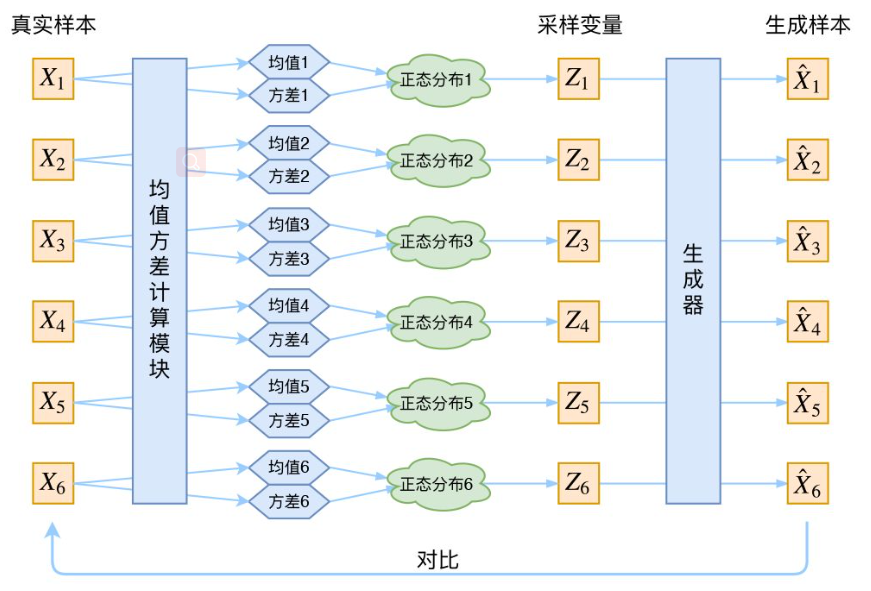




总结起来，整个训练框架就是在对样本进行编解码。是将样本编码为隐变量（变分近似分布），而又将隐含变量解码成（生产模型），进而最小化重构误差。训练的目的是学习出编码器的映射函数和解码器的映射函数，所以训练过程实际上是在进行变分推断，即寻找出某一个函数来优化目标。因此取名为变分编码器VAE(Variational Auto-encoder).



如上图：不确定经过重新采样出来的 *Zk*，是否还对应着原来的 *Xk*，所以我们如果直接最小化 *D*(*X̂ k*,*Xk*)^2（这里 *D*代表某种距离函数）是很不科学的



事实上，VAE 是为每个样本构造专属的正态分布，然后采样来重构。

总结：

样本X不是单位为1的量，是包含很多数据，数据的总个数为样本空间

1. 构造专属于样本Xk的正态分布，这个步骤就是encoder

为什么是专属？

是为了保证从Z中重采样后通过生成器出来的Xk~能够与原始样本Xk一一对应。

为什么是正态分布而不是其他分布？

通过实验表面，正太分布要好于均匀分布等分布函数，正太分布能固定均值的时候改变方差（噪声），均匀分布不能固定均值的时候改变方差，正态分布的均值和方差是两个独立的参数。

怎么构造？

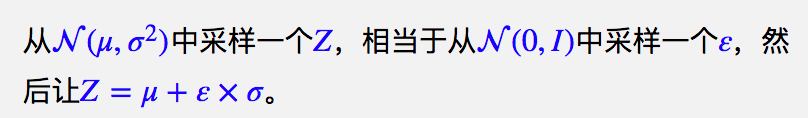
样本Xk中是一个高维的数据，我们不知道这些数据满足什么分布，但我们可以通过神经网络的强大去不断优化参数（类似权重和偏置，初始值随便设置），使得输出满足正态分布，均值和方差分别在两个神经网络中训练。

1. 让输出的正态分布N(μ，σ^2)不断去靠近标准正态分布N（0,1）,通过散度KL来衡量两个分布的相似度。对散度KL值采用随机梯度下降算法来优化构造的正太分布N(μ，σ^2)，使u🡪0, logσ^2🡪0（等价于σ^2🡪1)，选择拟合 log*σ*^2 而不是直接拟合 *σ*^2，是因为 *σ*^2 总是非负的，需要加激活函数处理，而拟合 log*σ*^2 不需要加激活函数，因为它可正可负。拟合均值和方差的时候，实质上是拟合神经网络中的参数，因为均值方差是由这些参数决定的

为什么要不断靠近标准正态分布？

我们重构 X的时候，也就是最小化 D(X̂k,Xk)^2，其结果要收到噪声的影响（方差），因为是在生成模型的输入参数Z是在构造的正态分布中重采样得到的，不是直接encoder计算出来的，（但其实我们的目的就是要加入噪声得以让模型又生成能力），而如果我们要重构得好的XK，必须减少噪声（方差为0无噪声），但方差为0时，也就没有随机性了，所以不管怎么采样其实都只是得到确定的结果（也就是均值），模型会慢慢退化成普通的 AutoEncoder，噪声不再起作用，因此，标准正态分布避免了噪声为0得情况，既能使模型有生成能力，又可以尽量使重构得XK与原样本中得XK得以相似。

1. 我们用重参数技巧，就是我们要从 *p*(*Z*|*Xk*) 中采样一个 *Zk* 出来，尽管我们知道了 *p*(*Z*|*Xk*) 是正态分布，但是均值方差都是靠模型算出来的，我们要靠这个过程反过来优化均值方差的模型，但是“采样”这个操作是不可导的，而采样的结果是可导的，于是我们利用了一个事实：（其实是为了优化均值方差模型）



最后我们让Z通过生成模型输出重构得X，利用损失函数去不断优化前面两个模型得参数值。

标准正太分布得值怎么采？

从eps中采样