|  |  |
| --- | --- |
| **文章信息** | **动机及方法概述** |
| **阅读程度**：  精读  **文章标题：**  Generative Modeling by Estimating Gradients of the  Data Distribution  **中文标题：**  通过估计数据分布的梯度进行生成建模  **发表于：**  NIPS2019  **作者：**  Yang Song  **单位：**  斯坦福 | 文章中最终要的两个东西：1.郎之万动力学采样 2.分数匹配，和两个挑战。  **郎之万动力学采样**：，这里只需要x0从一个给定的先验分布中采样，如标准正态分布中采样得到的高斯白噪音。其中表示概率密度函数在处的梯度，表示标准正态分布，在>0但比较小，但比较大的情况下，可以认为是从分布中采样到的一个样本。就叫分数score，如果能得到一个数据集分布密度的分数，那么就可以使用郎之万动力学采样这一过程得到符合这个分布密度的样本，这就是基于分数的生成模型所以来的原理，使用神经网络去学习输出一个分数这一过程就称为**分数匹配**。  而分数匹配无非就是进行一个优化：，这个公式很好理解，就是参数化一个神经网络，给一个样本，预测这个概率分布的分数。但是**概率分布本身是未知的**，更别说求得梯度，但是可以有一个**近似的优化目标**，（但这个优化目标成立的前提是the support of the data distribution is the whole space ，这就涉及了分数生成模型的**两个挑战之一The manifold hypothesis）**。通过加噪满足这个近似后，其中迹的计算是比较复杂的，这里对迹的计算又进行两种近似：**第一种是Denoising score matching**，直接绕过迹的计算，通过构建一个  条件分布对原始数据进行加噪，在加噪后的边缘分布中进行抽样，最小化网络输出与条件分布的之间的欧式距离，并且理论证明这个优化的理论最优结果就是边缘分布的分数，即，而当噪声较小时，边缘分布的分数近似等于原始分布的分数，因此可以认为网络预测结果与原始分布相近似；**第二种是Sliced score matching，**优化目标改为，加入了一个新的分布pv，自己定义的一个分布可以是一个多元标准正态分布，从中抽样向量v，然后最小化期望值即可。这两种近似都是基于第一个近似优化目标的再近似。  下面关于退火郎之万动力学采样的训练与预测过程才是真正的妙：  **基本设置：**，取L个标准差，保证从1到L逐渐递减，将原始分布使用高斯噪声进行扩散，这就是加噪后样本的条件分布。条件分布属于高斯分布，而高斯分布的分数可以通过计算得到准确值，作为标签。  **训练：**  我们训练一个网络，输入样本和噪声大小就可以得到这个样本在这个噪声大小下的条件分布的分数。。上面只是对一个固定的噪声大小的损失，我们总共有L个噪声大小，那总的损失是对这L个损失的加权平均，权重的取值是（因为在训练时我们统计发现了这么一个比例关系，因此当，），因此当乘以权重后我们得到前者正比与1，后者是一个标准正态分布，那么此时损失值的数量级与噪声大小无关，这意味着每一个取值的损失都保持在一个相当的数量级，这自然使得训练更加均衡。显然当训练均衡后，我们得到了一个很好的条件概率密度分数的估计。  **预测：**  当我们以标准高斯噪声为起点，使用条件下的分数网络结合郎之万采样过程迭代T次，此时可以认为最终的样本是从条件概率分布中采样到的一个样本，然后取，采用同样过程获得从中采样的一个样本，以此类推最终得到一个从采样的一个样本，由于此时很小，那么这个样本与t就差别不大了，其中t是一个从数据集分布中采样的一个结果。（这是我的理解）。  原文的理解是**Denoising score matching**对条件概率密度分数进行优化，其实最优解也等于边缘分布密度的分数，而在噪声较小的情况下，边缘概率密度分数约原始概率分布的分数，因此，以标准高斯噪声为起点，使用条件下的分数网络结合郎之万采样过程迭代T次，此时可以认为最终的样本是从边缘概率分布中采样到的一个样本，最终的分数近似等于原始分布的分数，得到一个符合原始分布采样的样本。以从中采样的得到样本为初始使用郎之万采样去获得一个符合分布的样本，显然比直接从高斯噪声为起点去获得一个符合分布的样本更为准确，因此，前L-1步迭代相当于逐步优化郎之万采样过程的起点，当得到一个符合分布的样本时，以此为起点进行最后一步郎之万采样自然效果很好。关于郎之万采样的步长是根据信噪比设置的：，，而，即采样的信噪比的数量级与无关。 |
| **摘要及贡献** | **实验** |
| 我们引入了一个新的生成模型，其中样本是通过朗之万动力学（使用梯度的数据分布估计）与分数匹配。  由于当数据位于低维流形上时，梯度可能定义不明确且难以估计，因此我们用不同水平的高斯噪声扰动数据，并联合估计相应的分数，即所有噪声水平的“被扰动数据分布”的梯度向量场。  对于采样，我们提出了一个退火的朗之万动力学，其中我们使用梯度对应于逐渐降低的噪声水平，因为采样过程越来越接近数据流形。  我们的框架允许灵活的模型架构，在训练期间不需要抽样或使用对抗性的方法，提供了一个学习目标，可用于进行模型比较的原则。  我们的模型产生的样本与GANs在MNIST、CelebA和CIFAR-10数据集上相当，在CIFAR-10上获得了最高水平的8.87分（inception score）。  此外，我们通过图像插入绘画实验证明了我们的模型学习到了有效的representations。 | **第一个挑战The manifold hypothesis的展示实验**：  这个假设认为，一个图片是一个D维的数据，如一个100个像素的rgb图片是一个300维的数据，而真实世界的数据分布集中在这个300维的线性空间的一个子空间中（超平面，原文叫ambient space），这样就有两个问题，第一是分数的定义是不明确的（这样理解，如果一个二的数据分布几乎全集中在二维空间的一条线上，那么这个数据分布的梯度很难求解），第二是分数匹配的近似优化目标是不成立的。解决方法是we perturb the data with a small Gaussian noise (such that the perturbed data distribution has full support over RD)，证据是使用原始CIfar10优化第一个近似目标（使用**Sliced score matching**）时，损失先下降再抖动，使用加噪数据优化第一个近似目标时，损失一直下降：    第二个挑战**Low data density regions**的展示实验：  **Inaccurate score estimation with score matching：**  ，即在高维空间中有很多地方会有一大块空间中的数据点的分布密度约等于0（这里不是子空间的概念，真子空间指嵌在高维空间中的低维线性空间，这里一大块空间高维空间中连续的一块而已，如三维空间中一个面是一个子空间，一个正方体包含的空间只是一个所谓的一大块空间，并不是一个线性空间，所以这里第二个挑战跟第一个挑战并不是一个概念，第一个挑战是分布蜷缩在一个子空间，是维度层面上的挑战。）这意味着训练时采集的数据样本中很少有样本来自于这块空间，此时这块空间的样本点处的分数预测是不准确的。一个小实验，作者设计了一个二元正态分布，在(-5,-5)与(5,5)这两个样本点附件的样本的概率密度值较高，其他区域的概率密度接近0，作者比较了真实梯度图与预测梯度图发现概率密度值较高的样本点附近的分数估计比概率密度值低的样本点附件的分数估计更为准确。（密度函数：    **Slow mixing of Langevin dynamics：**  这个问题的原文描述为：When two modes of the data distribution are separated by low density regions, Langevin dynamics will not be able to correctly recover the relative weights of these two modes in reasonable time, and therefore might not converge to the true distribution.例子是：当一个分布是两个分布累加得到的，并且两个概率分布的随机变量的取值范围没有交集的情况下，在p1所覆盖的取值范围集合中，这意味着，在这片取值范围内，概率分布密度的梯度与两个分布的权重完全没有关系，此时就会出现郎之万动力学无法还原权重的现象：    如图真实分布是，普通郎之万动力学采样在两个数据集中区域的密度是几乎没有区别的，而退火郎之万采样显然在(5,5)区域附近有更高的采样密度。关于退火郎之万采样，从大噪声逐步退火到小噪声，小噪声就可以解决**The manifold hypothesis问题，**而大噪声把更多小概率密度空间联通起来，能够解决**Low data density regions问题**  **关于最终实验：**  作者在三个数据集上进行了训练与图像生成，还使用郎之万动力学进行图像补全实验，我猜是以残块图像为起点，进行郎之万动力采样，最后获得完整样本。  ***优雅，永不过时！！*** |