Score Matching

Tags

1. Score的定义

对于一个概率分布 p(x)，其得分函数是指对数似然的梯度，记为：

$$\nabla_x \log p(x)$$

2. Matching的意义

在Score Matching中，目标是训练一个神经网络模型$s_\theta(x)$，使得它的输出尽可能接近真实数据分布的得分函数 $s_\theta(x)$

3. Score Matching的损失函数

初始损失函数：

$$\mathcal{L}(\theta) = \frac{1}{2}\mathbb{E}[\|s_\theta(x) - \nabla_x \log p(x)\|_2^2]$$

噪声加成和损失函数的修改

由于我们通常无法直接获得真实数据分布 $p(x)$，因此通过加噪声得到一个易于处理的条件分布。将数据点 x 加上高斯噪声，得到一个扰动的样本 $\tilde{x} \sim q_\sigma(\tilde{x}|x) = \mathcal{N}(x, \sigma^2\mathbf{I})$

然后，损失函数可以改写为：

$$\mathcal{L}(\theta;\sigma) = \frac{1}{2}\mathbb{E}_{p_{\text{data}}(x)}\mathbb{E}_{\tilde{x}\sim\mathcal{N}(x;\sigma^2\mathbf{I})}\left[\|s_\theta(\tilde{x},\sigma) + \frac{\tilde{x}-x}{\sigma^2}\|_2^2\right]$$

关于噪声的规模和权重系数的引入

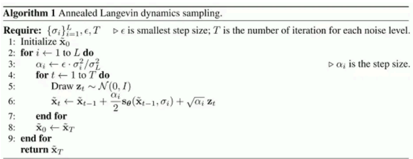
在实验中，发现

$$\|s_\theta(x,\sigma)\|_2 \propto \frac{1}{\sigma}$$

由于不同噪声水平的 σ会导致模型输出的波动较大，进而影响损失函数的整体值。为了解决这个问题，引一个权重系数 $\lambda(\sigma_i)$，并对不同噪声水平 σ_i的损失进行加权平均：

$$\mathcal{L}(\theta;\{\sigma_i\}_{i=1}^L) = \frac{1}{L}\sum_{i=1}^{L}\lambda(\sigma_i)\mathcal{L}(\theta;\sigma_i)$$

4.模型算法



Algorithm 1 Annealed Langevin dynamics sampling.

Require: $\{\sigma_i\}_{i=1}^L, \epsilon, T$ ▷ ϵ is smallest step size; T is the number of iteration for each noise level.

Initialize $\tilde{\mathbf{x}}_0$
for $i \leftarrow 1$ to L **do**
  $\alpha_i \leftarrow \epsilon \cdot \sigma_i^2/\sigma_L^2$ ▷ α_i is the step size.
  for $t \leftarrow 1$ to T **do**
    Draw $\mathbf{z}_t \sim \mathcal{N}(0, I)$
    $\tilde{\mathbf{x}}_t \leftarrow \tilde{\mathbf{x}}_{t-1} + \frac{\alpha_i}{2}\mathbf{s}_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}_{t-1}, \sigma_i) + \sqrt{\alpha_i}\,\mathbf{z}_t$
  end for
  $\tilde{\mathbf{x}}_0 \leftarrow \tilde{\mathbf{x}}_T$
end for
return $\tilde{\mathbf{x}}_T$

5.深入理解模型

朗之万动力学采样

布朗运动方程

$$m\frac{dv(t)}{dt} = -\gamma v(t) + \eta, \eta \sim \mathcal{N}(0, \sigma^2\mathbf{I})$$

其中，γ是动摩擦因数，η是随机力。布朗运动描述的是粒子的运动。

玻尔兹曼分布：

$$p(x) = \frac{e^{-U(x)}}{\mathcal{Z}}$$

其中$\mathcal{Z}$是归一化因子，$\mathcal{U}(x)$是势能。

玻尔兹曼分布变换：

$$\nabla_x \log p(x) = -\nabla_x \mathcal{U}(x) \tag{1}$$

根据动能定理有$F = \frac{dE}{dx} = \nabla_x E$，根据能量守恒有$E_\omega = E + U$。

$\nabla_x E = \nabla_x[E_\omega - U] = -\nabla_x U = F$

$F = -\nabla_x \mathcal{U}(x) = -\gamma v(t) + \eta$

结合（1）式：

但是物理中的扩散过程，粒子是在向着概率密度低的方向移动，可见$\nabla_x logp(x)$前面的符号为负号。朗之万动力学采样其实是一个逆的扩散过程：

为什么要加噪声?

根本原因：我们不知道$p_{data}(x)$的表示形式，也就无法给出损失函数的表示形式，无法进行估计。所以我们需要对原始数据$x \sim p_{data}(x)$进行加噪，得到一个已知的分布$\tilde{x} \sim q_\sigma(\tilde{x}|x) = \mathcal{N}(x, \sigma^2\mathbf{I})$。

主要原因：

1. **流形假设**：真实数据分布通常位于高维空间中的低维流形上。通过加噪声，数据的分布变得更加均匀，这使得模型能在整个空间内进行训练，从而打破了流形假设，提升了模型的泛化能力。
2. **数据增广**：加上了多种σ，使得训练数据在空间中分布的更为均匀。由于真实数据分布很集中，一开始采样很大几率采样在距离数据中心很远的地方，所以一开始加上比较大方差的噪声，步长也比较大，使得样本快速向真实数据分布移动，后面逐渐减小方差和步长，使得样本逐渐接近真实数据分布，最后加上的方差几乎可以忽略不计，使得最后采样得到的样本落在真实数据分布里（但是最后得到的数据还是带小噪声的，这可能也是Score Matching效果不好的原因）。

加噪声后的条件损失函数是否等价?

$$\mathcal{L} = \mathbb{E}_{q_\sigma(\tilde{x})}\left[\frac{1}{2}\|s_\theta(\tilde{x}, \sigma) - \nabla_{\tilde{x}} \log q_\sigma(\tilde{x})\|_2^2\right]$$

$$\mathcal{L}' = \mathbb{E}_{q_\sigma(\tilde{x}|x), q_\sigma(x)}\left[\frac{1}{2}\|s_\theta(\tilde{x}, \sigma) - \nabla_{\tilde{x}} \log q_\sigma(\tilde{x}|x)\|_2^2\right]$$

证明$\mathcal{L}' = \mathcal{L}$

$$\frac{1}{2}||s_\theta(\tilde{x},\sigma) - \nabla_{\tilde{x}} logq_\sigma(\tilde{x}|x)||^2$$
$$= ||s_\theta(\tilde{x},\sigma)||^2 - 2 < s_\theta(\tilde{x},\sigma), \nabla_{\tilde{x}} logq_\sigma(\tilde{x}|x) > + ||\nabla_{\tilde{x}} logq_\sigma(\tilde{x}|x)||^2$$

最后一项与模型参数θ无关，所以只看前两项

对于第一项$||s_\theta(\tilde{x},\sigma)||^2$

$$\mathbb{E}_{q_\sigma(\tilde{x})}[||s_\theta(\tilde{x},\sigma)||^2] = \int_{\tilde{x}} q_\sigma(\tilde{x})||s_\theta(\tilde{x},\sigma)||^2 d\tilde{x} = \int_{\tilde{x}}\int_x q_\sigma(\tilde{x}|x)q_\sigma(x)dx||s_\theta(\tilde{x},\sigma)||^2 d\tilde{x} = \mathbb{E}_{q_\sigma(\tilde{x}|x),q_\sigma(x)}[||S_\theta(\tilde{x},\sigma)||^2]$$

对于第二项$< s_\theta(\tilde{x},\sigma), \nabla_{\tilde{x}} logq_\sigma(\tilde{x}) >$

$$\begin{aligned}
\mathbb{E}_{q_\sigma(\tilde{x})}\left[\langle s_\theta(\tilde{x},\sigma), \nabla_{\tilde{x}} \log q_\sigma(\tilde{x})\rangle\right] = \;& \int_{\tilde{x}} q_\sigma(\tilde{x})\langle s_\theta(\tilde{x},\sigma), \nabla_{\tilde{x}} \log q_\sigma(\tilde{x})\rangle \\
&= \int_{\tilde{x}} q_\sigma(\tilde{x}) \left\langle s_\theta(\tilde{x},\sigma), \frac{\partial \log q_\sigma(\tilde{x})}{\partial \tilde{x}} \right\rangle d\tilde{x} \\
&= \int_{\tilde{x}} q_\sigma(\tilde{x}) \left\langle s_\theta(\tilde{x},\sigma), \frac{1}{q_\sigma(\tilde{x})}\frac{\partial q_\sigma(\tilde{x})}{\partial \tilde{x}} \right\rangle d\tilde{x} \\
&= \int_{\tilde{x}} \left\langle s_\theta(\tilde{x},\sigma), \frac{\partial q_\sigma(\tilde{x})}{\partial \tilde{x}} \right\rangle d\tilde{x} \\
&= \int_{\tilde{x}} \left\langle s_\theta(\tilde{x},\sigma), \frac{\partial \int_x q_\sigma(\tilde{x}|x)q_\sigma(x)dx}{\partial \tilde{x}} \right\rangle d\tilde{x} \\
&= \int_{\tilde{x}} \left\langle s_\theta(\tilde{x},\sigma), \int_x q_\sigma(x)\frac{\partial q_\sigma(\tilde{x}|x)}{\partial \tilde{x}} dx \right\rangle d\tilde{x} \\
&= \int_{\tilde{x}}\int_x q_\sigma(x) \left\langle s_\theta(\tilde{x},\sigma), \frac{\partial q_\sigma(\tilde{x}|x)}{\partial \tilde{x}} \right\rangle dx d\tilde{x} \\
&= \int_{\tilde{x}}\int_x q_\sigma(x)q_\sigma(\tilde{x}|x) \left\langle s_\theta(\tilde{x},\sigma), \frac{\partial \log q_\sigma(\tilde{x}|x)}{\partial \tilde{x}} \right\rangle dx d\tilde{x} \\
&= \mathbb{E}_{q_\sigma(\tilde{x}|x),q_\sigma(x)}\left[\langle s_\theta(\tilde{x},\sigma), \nabla_{\tilde{x}} \log q_\sigma(\tilde{x}|x)\rangle\right]
\end{aligned}$$