生成模型概述

什么是生成

- 根本问题：给定海量数据样本（如人脸图片），如何学习其背后那个极其复杂的高维概率分布 P(image)？
- 一旦学到这个分布，我们就能评估新样本的“真实性”（密度估计），更能从中采样，创造出全新的、逼真的样本（生成）

为什么要生成模型？

- 生成模型
  - 给定一组真实数据样本 $\{x_1, x_2, \ldots, x_n\}$，它们被假设是从一个未知的真实数据分布 $P_{data}(x)$ 中采样得到的
  - 生成模型 (Generative Model) 的任务是学习一个模型 $P_{model}(x)$，使其尽可能地逼近 $P_{data}(x)$
- 为什么要生成
  - 理解数据。 一个好的生成模型抓住了数据的本质结构与变化规律
  - 创造数据。一旦学到了 P_{model} ，就可以从中采样，生成全新的、与真实数据风格一致的数据。“创造力”
  - 应用: 图像合成、风格迁移、数据增强、超分辨率、药物发现等
- 传统方法的挑战
  - 直接对高维数据的 $P_{data}(x)$ 进行最大似然估计（MLE）通常是极其困难或不可行

生成的技术路线

- 显式密度建模 (Explicit Density Modeling - VAE)
  - 尝试构建一个带有参数 θ 的模型$\mathrm{p}_\theta(\mathrm{x})$，并直接优化它，使其逼近真实的数据分布（通常通过最大化对数似然 $\log \mathrm{p}_\theta(\mathrm{x})$。但由于 $\mathrm{p}(\mathrm{x})$ 形式复杂，我们转而优化其一个可计算的下界（ELBO）。
- 隐式密度建模 (Implicit Density Modeling - GAN)
  - 放弃直接定义和优化 $\mathrm{p}_\theta(\mathrm{x})$ 。而是构建一个“生成器”，它能直接从一个简单分布中采样并变换成目标样本。不关心 $\mathrm{p}_\theta(\mathrm{x})$ 的具体形式，只要求它生成的样本“看起来是真的”。通过一个“判别器”进行对抗训练，间接地让生成分布逼近真实分布。
- 基于马尔可夫链的采样 (Markov Chain-based Sampling - Diffusion)
  - 将“一步生成”分解为一系列极其简单的“小步去噪”问题。模型学习的不是直接生成，而是从纯噪声开始，通过一个逐步去噪的马尔可夫链，最终采样得到一个清晰的样本

数学基础

概率与分布

- 大写字母：随机变量
  - 如，图像随机变量X, 表示抽象图像的概念, 而非某一张具体的图像
- $P(X)$ ：随机变量 X 的概率分布
- 小写字母：随机变量的一个具体实现
  - 如， x 表示某一张具体图片，看作随机变量 X 的一次具体采样的结果
- $P(X = x)$ ：随机变量 X 采样得到 x 的概率数值
  - 概率密度函数 $P(X)$ 代入具体数值 x 的计算结果
- $\mathcal{N}(\mu, \sigma^2)$ 表示均值为 μ, 方差为 σ^2 的高斯分布
  - 请注意，不是$\mathcal{N}(\mu, \sigma)$

概率

- 事件：一个特定的结果或一组结果
- 样本空间：所有可能结果的集合
- 概率：表示某个事件发生的可能性

$$P(A) = \frac{\text{事件 A 发生的方式数}}{\text{样本空间的总方式数}}$$

  - 概率是一个数值，通常在 0 到 1 之间

概率

- 条件概率
- $P(A,B) = P(A \mid B)P(B) = P(B \mid A)P(A)$
- 边缘概率

$$p(B) = \int_A p(A,B)dA$$

- 求期望，去除无关变量的操作： $\mathrm{E}_{p(x_1,x_2,\ldots,x_T)} f(x_i) = \mathrm{E}_{p(x_i)} f(x_i)$

$$\mathrm{E}_{p(x_1,x_2,\ldots,x_T)} f(x_i)$$

$$= \int_{x_1,x_2,\ldots,x_T} p(x_1,x_2,\ldots,x_T) f(x_i) dx_1 dx_2 \ldots dx_T$$

$$= \int_{x_i} \int_{x_1,\ldots,x_{i-1},x_{i+1}\ldots,x_T} p(x_1,x_2,\ldots,x_T) \underbrace{dx_1 \ldots dx_{i-1} dx_{i+1} \ldots dx_T}_{\text{exclude } x_i} f(x_i) dx_i$$

$$= \int_{x_i} p(x_i) f(x_i) dx_i = \mathrm{E}_{p(x_i)} f(x_i)$$

条件概率

- $p(z \mid x)$
  - 函数映射
    - 给定x，得到的z概率$p(z \mid x)$
  - 推断/推理（本质/原因x，推断现象/结果z）
    - 先验概率$p(x)$
    - 根据观察现象/结果z，推断本质/原因x。后验概率$p(x \mid z)$
  - 类条件概率函数
    - 类别c 的特征xe的概率密度函数， $p(x \mid \mathrm{c})$
  - 编码
    - 给定x，得到编码z的概率$p(z \mid x)$
  - 解码/生成
    - 给定编码z ，解码得到x的概率$p(x \mid z)$

期望和采样

期望

- 概率密度函数 $p(x)$，x的期望

$$\mathbb{E}[x] = \int xp(x)dx$$

- （近似）数值计算时，选若干代表性的点 $x_0 < x_1 < x_2 < \cdots < x_n$，得

$$\mathbb{E}[x] \approx \sum_{i=1}^{n} x_i\, p(x_i)(x_i - x_{i-1})$$

- 如果根据 $p(x)$采样若干个点 $x_1, x_2, \ldots, x_n$，那么

$$\mathbb{E}[x] \approx \frac{1}{n}\sum_{i=1}^{n} x_i\,, \quad x_i \sim p(x)$$

- 记号
  - 假设有一个连续型随机变量 x ,其概率密度表示为 $p(x)$ 。此时,函数 $f(x)$ 的期望值可以表示为

$$\mathbb{E}_{p(x)}[f(x)] = \int f(x)p(x)\mathrm{d}x$$

    - 这种表示方法清晰地表明它是关于 $p(x)$ 的期望值
  - 以此类推,关于概率分布 $q(x)$ 的期望值表示为

$$\mathbb{E}_{q(x)}[f(x)] = \int f(x)q(x)\mathrm{d}x$$

期望 - 蒙特卡洛模拟

- 计算方式的主要区别
  - 其一包含了概率的计算
  - 另一个仅有x的计算
    - x_i从$p(x)$中依概率采样而来，概率大的x_i出现的次数也多，采样的结果已经包含了$p(x)$信息

$$\mathbb{E}_{x\sim p(x)}[f(x)] = \int f(x)p(x)dx \approx \frac{1}{n}\sum_{i=1}^{n} f(x_i), \quad x_i \sim p(x)$$

KL 散度

- 衡量两个概率分布之间差异的一种方法是 KL 散度。当给定两个概率分布 $p(x)$ 和 $q(x)$ 时,当 x 为连续型随机变量时,KL 散度

$$D_{\mathrm{KL}}(p \parallel q) = \int p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x$$

  - 当 x 为离散型随机变量时, 数学式如下所示。

$$D_{\mathrm{KL}}(p \parallel q) = \sum_{x} p(x) \log \frac{p(x)}{q(x)}$$

- KL 散度具有以下特性:
  - 两个概率分布的差异越大, KL 散度的值就越大;
  - KL 散度的值大于或等于 0 ,且仅当两个概率分布相同时,其值才为 0 ;
  - KL 散度是非对称的衡量指标,因此 $D_{\mathrm{KL}}(p \parallel q)$ 和 $D_{\mathrm{KL}}(q \parallel p)$ 的值不同。

“最大似然估计” (Maximum Likelihood Estimation)

- 假设有一个概率分布,其形状由参数 θ 决定。当参数为 θ 时,获得数据 x 的概率密度为 $p(x;\theta)$ 。
- 对于正态分布,可以将参数视为 $\theta = \{\mu, \sigma\}$ 。

“最大似然估计” (Maximum Likelihood Estimation)

- 样本 $\mathcal{D} = \{x^{(1)}, x^{(2)}, \cdots, x^{(N)}\}$ 每个数据都是基于概率分布 $p(x;\theta)$ 独立生成的。此时，获得样本 $\mathcal{D}$ 的概率密度

$$p(\mathcal{D};\theta) = p(x^{(1)};\theta)p(x^{(2)};\theta)\cdots p(x^{(N)};\theta) = \prod_{n=1}^{N} p(x^{(n)};\theta)$$

- 由于假定每个数据是独立生成的,因此获得 N 个数据的概率密度等于每个数据的概率密度的乘积。
- $p(\mathcal{D};\theta)$ 表示当参数为 θ 时获得样本 $\mathcal{D}$ 的概率密度。这个 $p(\mathcal{D};\theta)$ 也可以被看作以 θ 为参数的函数,其定义式如下所示。

$$L(\theta) = p(\mathcal{D};\theta)$$

“最大似然估计” (Maximum Likelihood Estimation)

- $L(\theta) = p(\mathcal{D};\theta) = p\left(x^{(1)};\theta\right)p\left(x^{(2)};\theta\right)\cdots p\left(x^{(N)};\theta\right) = \prod_{n=1}^{N} p\left(x^{(n)};\theta\right)$
  - $L(\theta)$ 称为似然 (likelihood) 或似然函数 (likelihood function)
  - 以参数 θ 为参数的函数,表示在给定参数 θ 的情况下,样本 $\mathcal{D}$ 出现的概率密度。
- 最大似然估计，找到使似然 $p(\mathcal{D};\theta)$ 最大的参数 θ
  - 如果使似然最大的参数是 $\hat{\theta}$,那么当参数为 $\hat{\theta}$ 时,观测到样本的概率最大
  - 也就是说,当参数为 $\hat{\theta}$ 时,模型最拟合样本。
- 其实最大似然估计进行的不是似然 $p(\mathcal{D};\theta)$ 的最大化,而是对数似然 $\log p(\mathcal{D};\theta)$ 的最大化

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_{n=1}^{N} \log p_{\theta}\left(x^{(n)}\right)$$

模型表达

模型编码或者解码的表示

- 函数
  - 直接对应关系
- 概率
  - 输入x，生成z的概率$p(z \mid x)$
- 神经网络，拟合
  - 函数
  - 概率中的参数

函数

- $z = f(x)$
  - 直接对应关系，根据输入得到确定的输出

- 缺点
  - 对已经有的数据的总结
  - 没有考虑数据之间的关联，所以并不能生成新的数据

- 如果希望生成新的数据，那么模型需要考虑样本推断相邻区域的能力，因此，对应不能只是一个点，应该是一片
  - 因此，映射机制发展成概率和采样的生成机制

概率生成模型 - 条件概率$p(z \mid x)$

- $p(z \mid x)$：由x得到z的概率。x参数，z为随机变量
- $p(z \mid x)$为x到各个z（ 没有唯一对应的z ）映射的概率（频率或者可能性）
  - $p(z \mid x)$值越大，z和x的相关性越高
  - 和映射$z = f(x)$形式对照

$$z_{max} = \arg\max_{z} p(z|x)$$

- 按照概率$p(z \mid x)$采样z，记号：$z \sim p(z|x)$
  - 采样得到z：先采样一个x，再依据概率$p(z \mid x)$采样得到z
  - 一个策略：采样1000次，得到1000个z_i，这些z_i符合概率$p(z_i \mid x)$分布，如果只要一个样本，随机挑选一个
- 如果$p(z \mid x)$形式为高斯分布。给定x，
  - z的一个采样 =（高斯函数的均值）+ （随机生成一个标准正态噪声）* 方差