数据生成问题

简单的随机数生成

生成对抗网络的思想

网络结构

优化目标函数

训练算法

理论分析

数据生成问题
之前介绍的机器学习算法都是在解决分类、回归、聚类或者数据降维之类的数据预测问题
另外还存在一类称为数据生成的问题，它的目标是生成服从某种概率分布的数据

实际例子-生成手写数字图像
数字图像可以看做是一个随机向量，如何生成这个向量？
人在学习写字时是通过练习得到的，刚开始写的不好，不断收到反馈并进行改进，最后学会写字
用机器学习解决该问题时，可以采用类似的思路

目前常见的深度生成模型
VAE-变分自动编码器
GAN-生成对抗网络。本课程介绍的重点

生成均匀分布的随机数

编程语言中常用的随机数函数就是一种随机数据生成算法，它可以生成符合某种概率分布的随机数（实际上是伪随机数而不是真正意义上的随机数），如均匀分布和正态分布的随机数

生成均匀分布随机数的经典算法是线性同余法

$$x_{i+1} = \left(a \cdot x_i + b\right) \bmod m$$

生成正态分布的随机数

Box-Muller算法，借助于均匀分布的随机数

假设随机变量 u_1 和 u_2 服从 $[0,1]$ 内的均匀分布，则随机数 z_1 和 z_2

$$z_1 = \sqrt{-2\ln u_1}\cos 2\pi u_2$$

$$z_2 = \sqrt{-2\ln u_1}\sin 2\pi u_2$$

相互独立，并且服从正态分布 $N(0,1)$

这两个例子都是已知要生成的数据所服从的概率分布，如均匀分布、正态分布；并且分布的参数也是已知的，比如正态分布的均值和方差，这称为显式的建模

对于实际应用中的很多问题，只有一些样本，算法需要从这些样本来估计它们服从的分布，且概率分布非常复杂，无法得到概率密度函数的精确表达式，但要估计出概率密度函数或者直接根据一个模型生成想要的随机数，这称为隐式建模

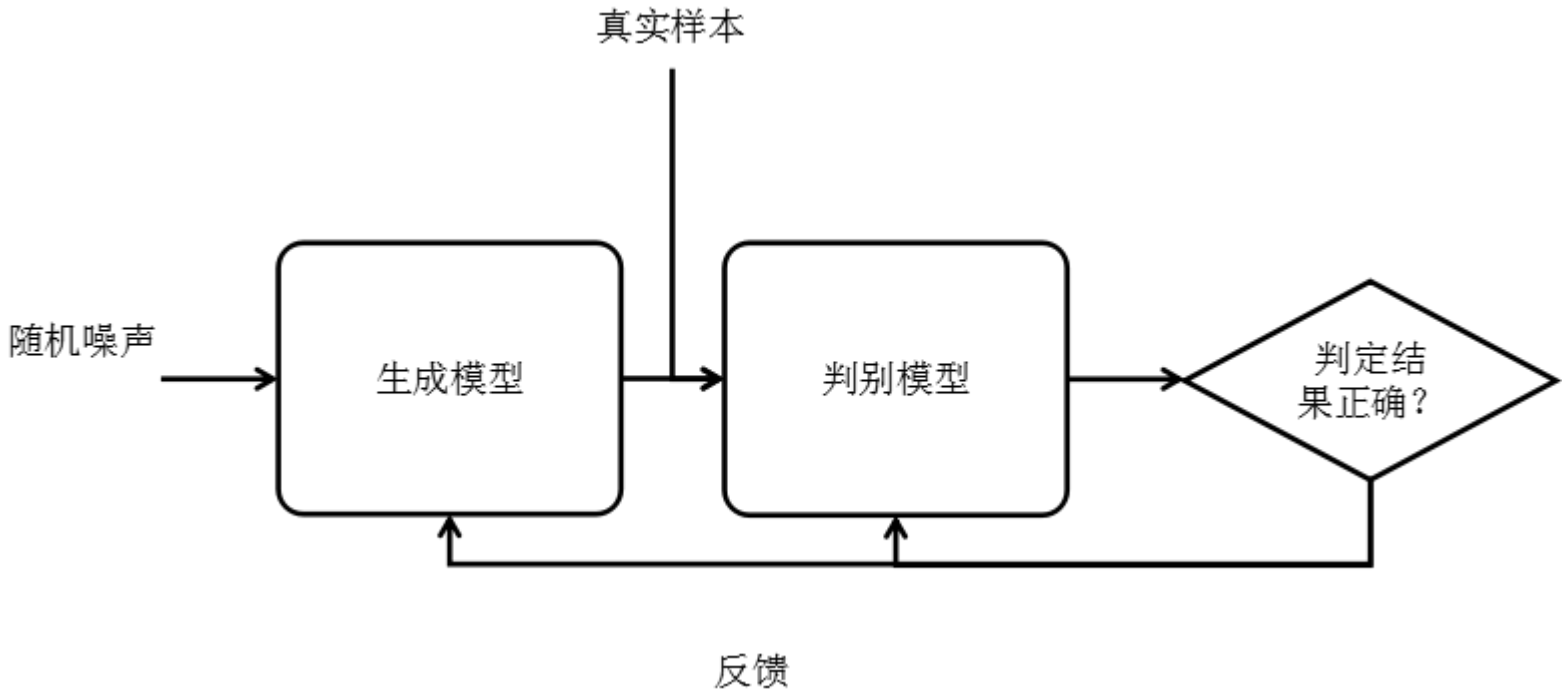
对于写出0-9之间的数字的问题，算法无法得知这些数字的图像所服从的概率分布。对于每个类型的数字，假设要生成32x32的黑白数字图像，将图像拼接成向量为1024维的随机向量，每种数字服从某种概率分布

$$p\left(\mathrm{x}\middle|c\right), c = 0, \ldots, 9$$

我们并不知道这个概率分布的具体形式。需要通过机器学习的手段直接产生一个映射函数，给定输入数据如噪声和数字的类别，直接产生出服从此概率分布的样本

生成对抗网络的基本思想
由一个生成模型和一个判别模型组成
生成模型用于学习真实样本数据的概率分布，并直接生成符合这种分布的数据
判别模型的任务是指导生成模型的训练，判断一个输入样本数据是真实样本还是由生成模型生成的
在训练时，两个模型不断竞争，从而分别提高它们的生成能力和判别能力

生成对抗网络的结构



生成模型

以随机噪声或类别之类的控制变量作为输入，一般用多层神经网络实现，其输出为生成的样本数据，这些样本数据和真实样本一起送给判别模型进行训练

让生成的数据尽可能与真实数据相似，最小化判别模型的判别准确率

从训练样本学习到它们所服从的概率分布 p_g ，假设随机噪声变量服从的概率分布为 $p_z(\mathrm{z})$

生成模型将这个随机噪声映射到样本数据空间

$$G\left(\mathrm{z},\theta_g\right)$$

这个映射根据随机噪声变量构造出服从某种概率分布的随机数

判别模型
是一个二分类器，判定一个样本是真实的还是生成的，一般也用神经网络实现
训练目标是最大化判别准确率，即区分样本是真实数据还是由生成模型生成的
当这个样本被判定为真实数据时标记为1，判定为来自生成模型时标记为0

$$D(\mathrm{x},\theta_d)$$

在训练时，两个模型不断竞争，从而分别提高它们的生成能力和判别能力
随着训练的进行，生成模型产生的样本与真实样本几乎没有差别，判别模型也无法准确的判断出一个样本是真实的还是生成模型生成的，此时的分类错误率为0.5，系统达到平衡，训练结束

优化目标函数
训练的目标是让判别模型能够最大程度的正确区分真实样本和生成模型生成的样本；同时要让生成模型生成的样本尽可能的和真实样本相似

判别模型要尽可能将真实样本判定为真实样本，将生成模型产生的样本判定为生成样本
生成模型要尽量让判别模型将自己生成的样本判定为真实样本

对于生成模型，要最小化如下目标函数

$$\log\left(1-D\left(G\left(\mathrm{z}\right)\right)\right)$$

对于判别模型，要让真实样本尽量被判定为真实的，即最大化 $\log D\left(\mathrm{x}\right)$ ，这意味着 $D\left(\mathrm{x}\right)$ 的值尽量接近于1；对于生成模型生成的样本，尽量被判别为0，即最大化 $\log\left(1-D\left(G\left(\mathrm{z}\right)\right)\right)$

由此得到目标函数为

$$\min_G \max_D V\left(D,G\right)=\mathrm{E}_{\mathrm{x}\sim p_{data}\left(\mathrm{x}\right)}\left[\log D\left(\mathrm{x}\right)\right]+\mathrm{E}_{\mathrm{z}\sim p_{\mathrm{z}}\left(\mathrm{z}\right)}\left[\log\left(1-D\left(G\left(\mathrm{z}\right)\right)\right)\right]$$

类比logistic回归的目标函数，二者都要解决二分类问题

$$\sum_{i=1}^{l}\left(y_i \log h_{\mathrm{w}}\left(\mathrm{x}_i\right)+\left(1-y_i\right)\log\left(1-h_{\mathrm{w}}\left(\mathrm{x}_i\right)\right)\right)$$

如果将正样本和负样本拆开，得到如下的形式

$$\sum_{i=1,y_i=1}^{l}\log h_{\mathrm{w}}\left(\mathrm{x}_i\right)+\sum_{i=1,y_i=0}^{l}\log\left(1-h_{\mathrm{w}}\left(\mathrm{x}_i\right)\right)$$

logistic回归在训练达到最优点处时负样本的预测输出值接近于0，而在生成对抗网络中判别模型对生成样本的输出概率值在最优点处接近于0.5，与生成模型达到均衡

训练时采用分阶段优化策略进行优化，交替的优化生成模型和判别模型，最终达到平衡的状态，训练终止
完整的训练算法如下
循环，对t = 1, ..., max_iter
第一阶段：训练判别模型
循环，对i = 1, ..., k
根据噪声服从的概率分布产生m个噪声数据 $\mathrm{z}_1,\ldots,\mathrm{z}_m$
根据样本数据服从的概率分布采样出m个样本 $\mathrm{x}_1,\ldots,\mathrm{x}_m$
用随机梯度上升法更新判别模型，判别模型参数梯度的计算公式为

$$\nabla_{\theta_d}\frac{1}{m}\sum_{i=1}^{m}\left[\log\left(D\left(\mathrm{x}_i\right)\right)+\log\left(1-D\left(G\left(\mathrm{z}_i\right)\right)\right)\right]$$

结束循环
第二阶段：训练生成模型
根据噪声分布产生m个噪声数据 $\mathrm{z}_1,\ldots,\mathrm{z}_m$
用随机梯度下降法更新生成模型，生成模型参数的梯度计算公式为：

$$\nabla_{\theta_g}\frac{1}{m}\sum_{i=1}^{m}\log\left(1-D\left(G\left(\mathrm{z}_i\right)\right)\right)$$

结束循环

理论分析

如果生成模型固定不变，使得目标函数取得最优值的判别模型为

$$D_G^*(\mathrm{x}) = \frac{p_{data}(\mathrm{x})}{p_{data}(\mathrm{x}) + p_g(\mathrm{x})}$$

将数学期望按照定义展开，要优化的目标是

$$\begin{aligned} V(G,D) &= \int_{\mathrm{x}} p_{data}(\mathrm{x}) \log\left(D(\mathrm{x})\right) d\mathrm{x} + \int_{\mathrm{z}} p_z(\mathrm{z}) \log\left(1 - D\left(g(\mathrm{z})\right)\right) d\mathrm{z} \\ &= \int_{\mathrm{x}} p_{data}(\mathrm{x}) \log\left(D(\mathrm{x})\right) + p_g(\mathrm{x}) \log\left(1 - D(\mathrm{x})\right) d\mathrm{x} \end{aligned}$$

将$p_{data}(\mathrm{x})$和$p_g(\mathrm{x})$看作常数，上式为$D(\mathrm{x})$的函数。构造如下函数

$$a \log x + b \log(1-x)$$

对函数求导并令导数为0，解方程可以得到

$$x = a/(a+b)$$

函数在该点处取得极大值

将最优判别模型的值代入目标函数中消掉D，得到关于G的目标函数

$$C(G)=\max_D V(D,G)$$

$$=\mathrm{E}_{\mathrm{x}\sim p_{data}(\mathrm{x})}\left[\log D_G^*(\mathrm{x})\right]+\mathrm{E}_{\mathrm{z}\sim p_z(\mathrm{z})}\left[\log\left(1-D_G^*(G(\mathrm{z}))\right)\right]$$

$$=\mathrm{E}_{\mathrm{x}\sim p_{data}(\mathrm{x})}\left[\log D_G^*(\mathrm{x})\right]+\mathrm{E}_{\mathrm{z}\sim p_g(\mathrm{z})}\left[\log\left(1-D_G^*(\mathrm{x})\right)\right]$$

$$=\mathrm{E}_{\mathrm{x}\sim p_{data}(\mathrm{x})}\left[\log\frac{p_{data}(\mathrm{x})}{p_{data}(\mathrm{x})+p_g(\mathrm{x})}\right]+\mathrm{E}_{\mathrm{z}\sim p_g(\mathrm{z})}\left[\log\frac{p_g(\mathrm{x})}{p_{data}(\mathrm{x})+p_g(\mathrm{x})}\right]$$

当且仅当

$$p_g=p_{data}$$

时这个目标函数取得最小值，且最小值为-log4

如果有

$$p_g=p_{data}$$

$$D_G^*(\mathrm{x})=1/2$$

则有

$$C(G)=\log\frac{1}{2}+\log\frac{1}{2}=-\log 4$$

将$C(G)$减掉-log4，有

$$C(G) = -\log 4 + KL\left(p_{data} \middle\| \frac{p_{data} + p_g}{2}\right) + KL\left(p_g \middle\| \frac{p_{data} + p_g}{2}\right)$$

KL为Kullback-Leibler散度（简称KL散度）。KL散度用于衡量两个概率分布之间的距离。对于离散型随机变量定义为

$$KL(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

目标函数也可以写成

$$C(G) = -\log(4) + 2JSD\left(p_{data} \| p_g\right)$$

JSD为Jensen-Shannon散度。Jensen-Shannon散度衡量两个概率分布之间的相似度，定义为

$$JSD(p\|q) = \frac{1}{2}KL(p\|m) + \frac{1}{2}KL(q\|m)$$

其中

$$m = \frac{1}{2}(p+q)$$

两个概率分布之间的Jensen-Shannon散度非负，并且只有当两个分布相等时取值为0，因此结论成立
当生成模型生成的样本和真实样本充分相似时，判别模型无法有效区分二者，此时系统达到最优状态

AI学习与实践平台

www.sigai.cn