人工智能概论-对抗学习

赵亚伟

zhaoyw@ucas.ac.cn

中国科学院大学 大数据分析技术实验室

2018.6.29

目录

齐贤易财

- （北宋）戚里有分财不均者，更相讼。齐贤曰：“是非台府所能决，臣请自治之。”齐贤坐相府，召讼者问曰：“汝非以彼分财多，汝分少乎?”曰：“然。”具款，乃召两吏，令甲家入乙舍，乙家入甲舍，货财无得动，分书则交易，明日奏闻，上曰：“朕固知非君不能定也。”

纳什均衡

定义

- 在博弈$G=\{S_1,\ldots, S_n: u_1,\ldots, u_n\}$中，如果由各个博弈方的各一个策略组成的某个策略组合$(s_1^*,\ldots, s_n^*)$中，任一博弈方i的策略s_i^*，都是对其余博弈方策略的组合$(s_1^*,\ldots, s^*_{i-1}, s^*_{i+1},\ldots, s_n^*)$的最佳对策，也即$u_i(s_1^*,\ldots s^*_{i-1}, s_i^*, s^*_{i+1},\ldots, s_n^*)\geq u_i(s_1^*,\ldots, s^*_{i-1}, s_{ij}^*, s^*_{i+1},\ldots, s_n^*)$对任意$s_{ij}\in S_i$都成立，则称$(s_1^*,\ldots, s_n^*)$为$G$的一个纳什均衡。

解释

- 假设有n个局中人参与博弈，如果某情况下无一参与者可以独自行动而增加收益（即为了自身利益的最大化，没有任何单独的一方愿意改变其策略的），则此策略组合被称为纳什均衡。
- 所有局中人策略构成一个策略组合（Strategy Profile）。本质上，纳什均衡是一种**非合作博弈状态**。

约翰·纳什，生于1928年6月13日。著名经济学家、博弈论创始人、《美丽心灵》男主角原型。1994年获得诺贝尔经济学奖。2015年5月23日，约翰·纳什夫妇遇车祸，在美国新泽西州逝世。

零和博弈思想（Zero-sum Game）

- 指参与博弈的各方，在严格竞争下，一方的收益必然意味着另一方的损失，博弈各方的收益和损失相加总和永远为“零”，双方不存在合作的可能。属非合作博弈，纳什均衡。
- 博弈双方的利益之和是一个常数，比如两个人掰手腕，假设总的空间是一定的，你的力气大一点，那你就得到的空间多一点，相应的我的空间就少一点，相反我力气大我就得到的多一点，但有一点是确定的就是，我俩的总空间是一定的，这就是二人博弈，但是总利益是一定的。
- 例子
  - 赌博
  - 掰手腕
  - 打扑克
  - 微信红包
  - ……

经典问题

- **无监督学习数据质量问题**
  - 数据很多，质量很差，挑挑拣拣后剩下的是小样本，出现样本不足问题
- **样本不足问题：**
  - 无监督学习中的著名问题：给定一批样本，如何设计一个系统，并能够**生成(generate)**类似的新样本。如何做到系统产生的“新样本”可以作为训练样本使用？即可以做到“以假乱真”！
- **无监督学习是否可以转化为有监督学习?**

Generative Adversarial Nets

- Train two models: a generative model *G* that captures the data distribution, and a discriminative model *D* that estimates the probability that a sample came from the training data rather than *G*.
- 目标是为了获得生成模型*G*（Generative Model）
- 当固定生成网络 *G* 的时候，对于判别网络 *D* 的优化，可以这样理解：
  - 输入来自于真实数据，*D* 优化网络结构使自己输出 1，输入来自于生成数据，*D* 优化网络结构使自己输出 0；
- 当固定判别网络 *D* 的时候，*G* 优化自己的网络使自己输出尽可能和真实数据一样的样本（损失最小），并且使得生成的样本经过 *D* 的判别之后，*D* 输出高概率。

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

损失函数

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

- 上述这个公式说白了就是一个最大最小优化问题，其实对应的也就是上述的两个优化过程。
- 这个公式既然是最大最小的优化，那就不是一步完成的，其实对比我们的分析过程也是这样的，先优化D，再优化G，本质上是两个优化问题，把拆解就如同下面两个公式：
- 优化D：

$$\max_D V(D,G) = E_{x\sim p_{data}(x)}[log(D(x))] + E_{z\sim p_z(z)}[log(1 - D(G(z)))]$$

- 优化G：

$$\min_G V(D,G) = E_{z\sim p_z(z)}[log(1 - D(G(z)))]$$

优化D

判别真样本，越大越好　　判别假样本，越小越好

$$\max_{D} V(D,G) = E_{x\sim p_{data}(x)}[log(D(x))] + E_{z\sim p_z(z)}[log(1-D(G(z)))]$$

- 优化D的时候，**G固定**，后面的G(z)就相当于得到的假样本。优化D的公式的第一项，使的真样本*x*输入的时候，得到的结果越大越好，可以理解，因为需要真样本的预测结果越接近于1越好。
- 对于假样本，需要优化是的其结果越小越好，也就是D(G(*z*))越小越好，因为它的标签为0。但是，要求第一项越大，同时要求第二项是越小，这就矛盾了
- 所以，把第二项改成1-D(G(*z*))，这样就是越大越好，两者合起来就是越大越好。

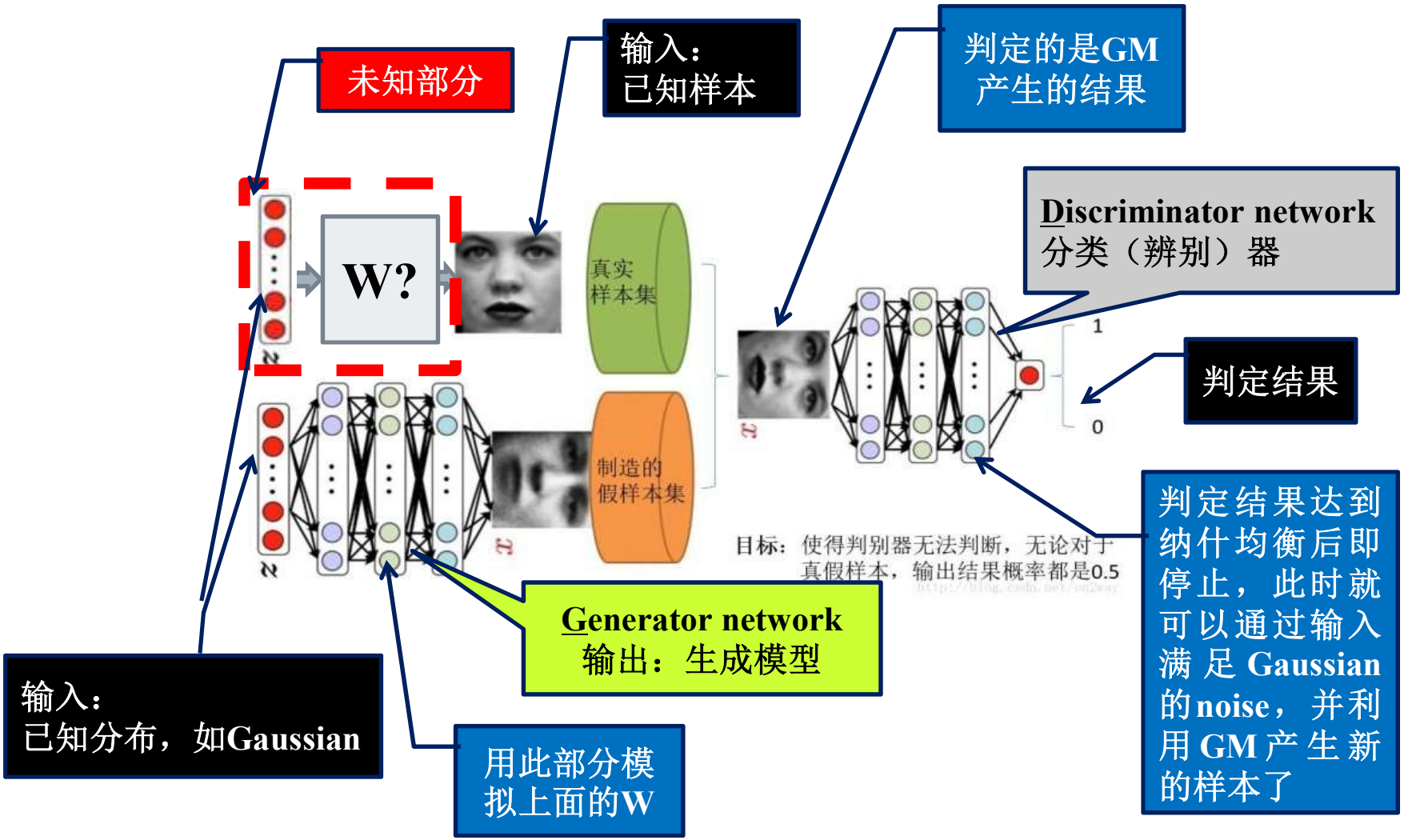
优化G

$$\min_G V(D,G) = E_{z\sim p_z(z)}[log(1-D(G(z)))]$$

判别假样本，越小越好！与D(x)真样本无关，所以，D(x)去掉

- 优化G的时候，这个时候没有真样本什么事，所以把第一项直接去掉了。
- 这个时候只有假样本，但是我们说这个时候希望假样本的标签是1，所以是D(G(z))越大越好，但是为了统一成1-D(G(z))的形式，那么只能是最小化1-D(G(z))，本质上没有区别，只是为了形式的统一。
- 之后这两个优化模型可以合并起来写，就变成了最大最小目标函数了。

图解GAN


未知部分
输入：
已知样本
判定的是GM
产生的结果
Discriminator network
分类（辨别）器
W?
真实
样本集
制造的
假样本集
1
0
判定结果
目标：使得判别器无法判断，无论对于
真假样本，输出结果概率都是0.5
判定结果达到
纳什均衡后即
停止，此时就
可以通过输入
满 足 Gaussian
的noise，并利
用 GM 产 生 新
的样本了
Generator network
输出：生成模型
输入：
已知分布，如Gaussian
用此部分模
拟上面的W

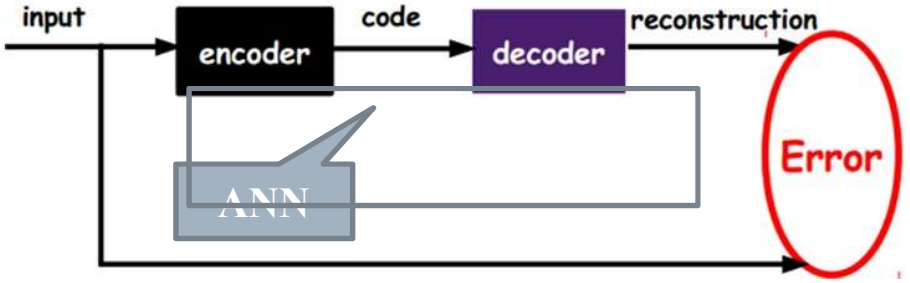
目录

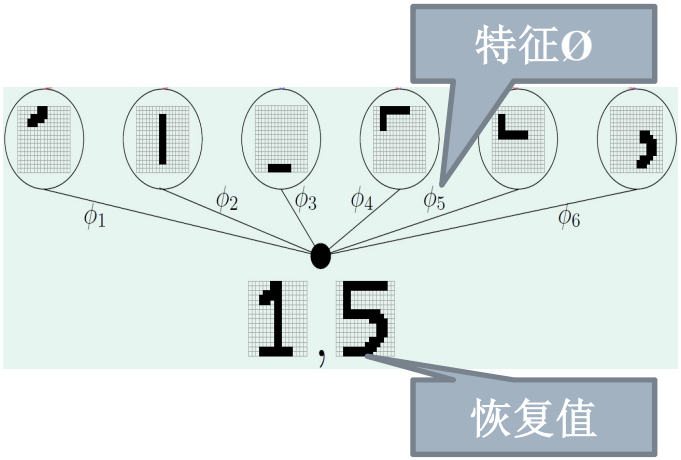
Adversarial Autoencoders（AAE)

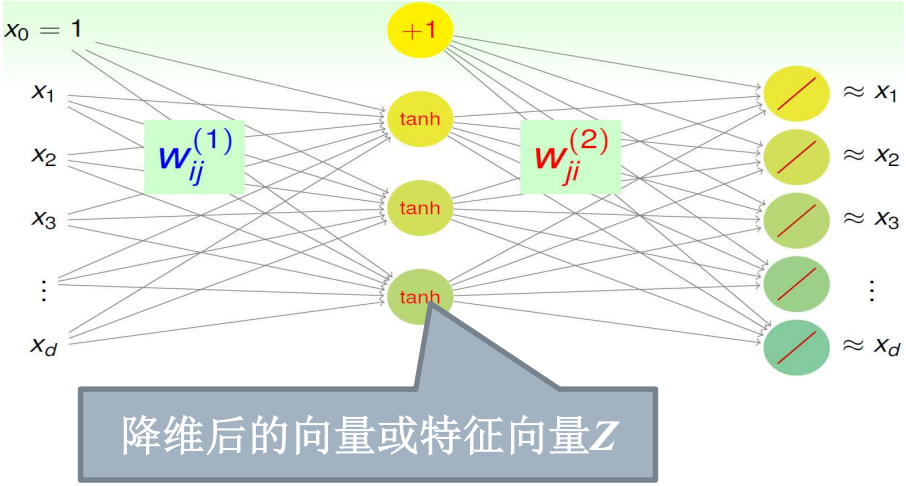
- **解决问题： Autoencoder编码结果不可控，如何使编码结果符合预期分布（即如何获取特定分布的编码层）？**
  - 如何将自动编码机编码转换为一个生成模型问题？
  - 模型中两个标准：
    - ① 传统重构误差标准（AE的问题）
    - ② 对抗训练标准（满足集成前部分布，对任意先验分布自动编码机的隐含表示）？（GAN的问题）
- **目标：**
  - 采用对抗学习的方法获得符合预期的隐向量分布

关于自动编码机

- Autoencoder是一种无监督的学习算法，主要用于数据的降维或者特征的抽取，在深度学习中，Autoencoder可用于在训练阶段开始前，确定权重矩阵W的初始值。
- 所有自动编码机，目标都是样本重构





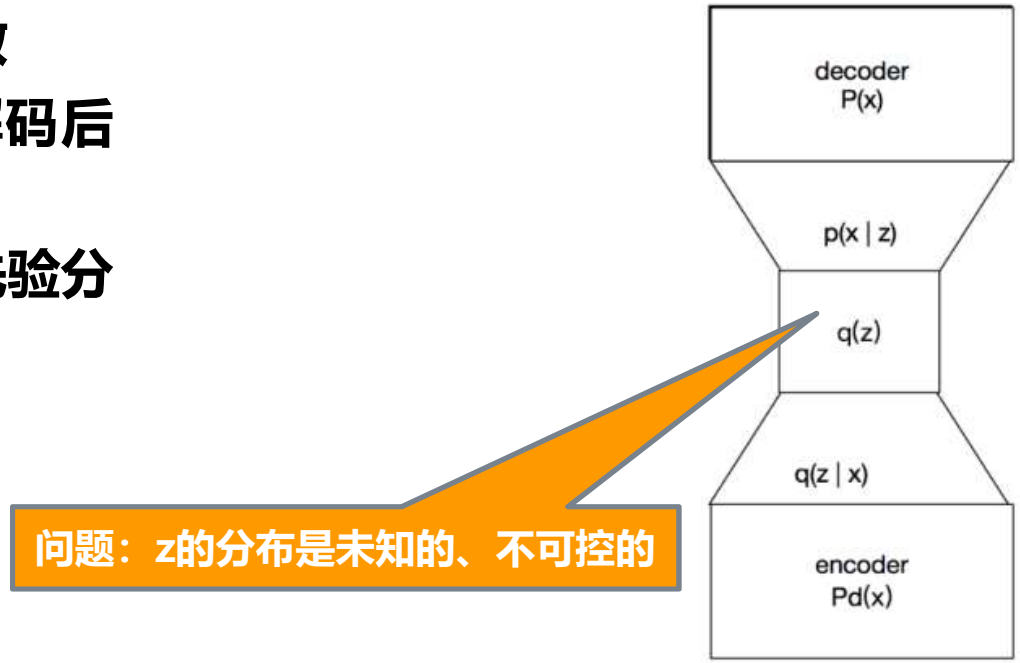


自动编码机的结构

- x 是样本输入
- z 是隐编码向量（隐单元）
- $p_d(x)$ 表示真实数据x分布
- $q(z|x)$ 是编码 z 分布函数
- $q(z)$ 聚合（编码 z）后验分布
- $p(x|z)$ 是解码 x 分布函数
- $p(x)$ 表示模型数据（解码后的x）分布
- $p(z)$是隐编码 z 的任意先验分布

The encoding function of the autoencoder $q(z|x)$ defines an aggregated posterior distribution of $q(z)$ on the hidden code vector of the autoencoder as follows:

$$q(\mathbf{z}) = \int_{\mathbf{x}} q(\mathbf{z}|\mathbf{x}) p_d(\mathbf{x}) d\mathbf{x}$$



对抗自动编码机（AAE）

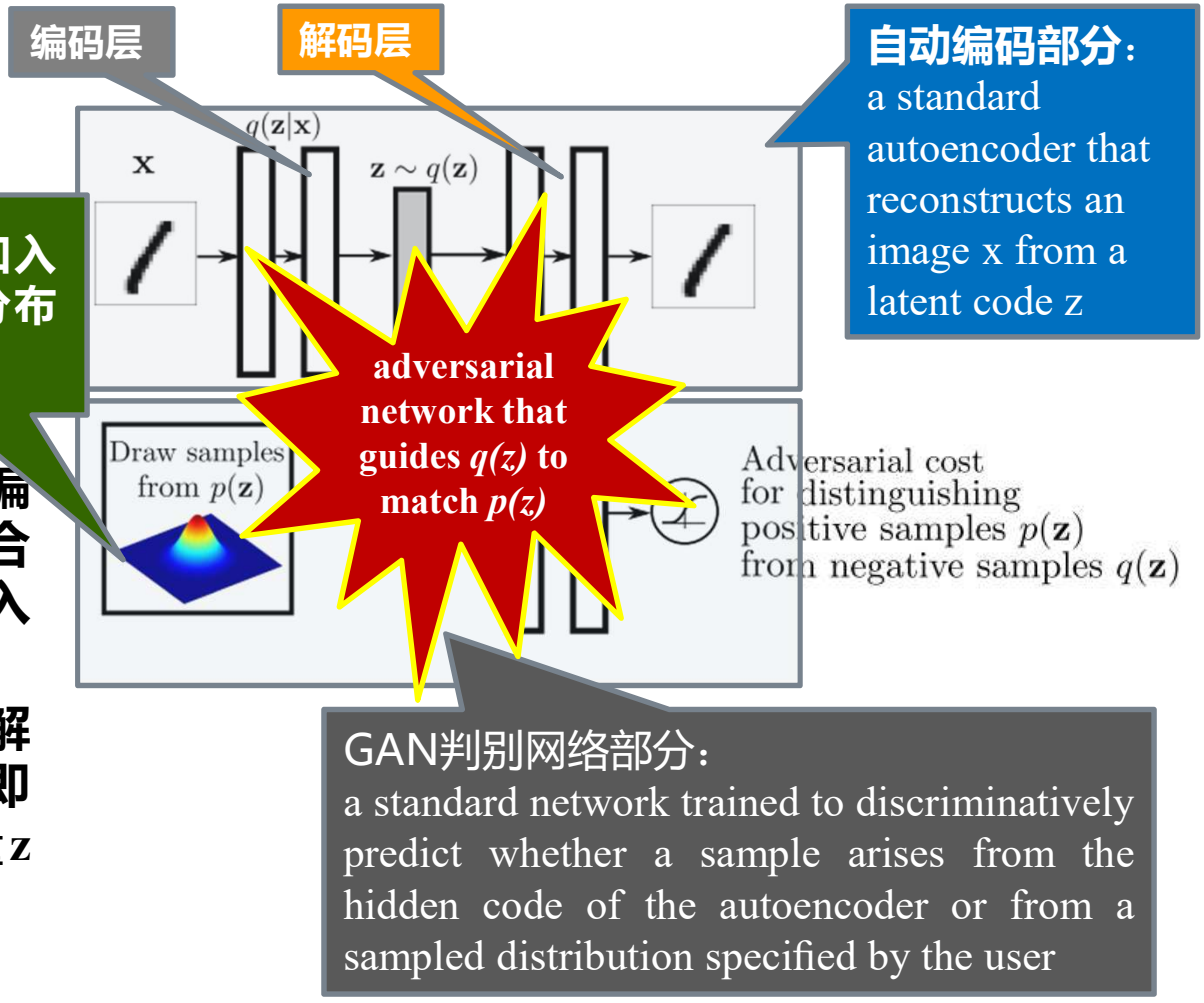
- AAE就是将对抗的思想融入到Autoencoders中来，来干什么呢?
  - 获得指定分布的隐编码 z，从而学得能够获得 $q(z)$ 的AE输入层（x是不可变的，所以实际上优化的是边权 $W_{ij}^{(1)}$）。
- Once the training procedure is done, the decoder of the autoencoder will define a generative model that maps the imposed prior of p(z) to the data distribution.（一旦AAE训练过程完成，编码机的编码层将定义一个产生模型，该模型可强制将先验分布的$p(z)$（通过$W_{ij}^{(1)}$）映射到数据分布）

基本思想

- the *adversarial network* and the *autoencoder* are trained jointly with SGD in two phases
  - the *reconstruction* phase
    - the autoencoder updates the encoder and the decoder to minimize the reconstruction error of the inputs. （最小化输入和输出的error）
  - the *regularization* phase – executed on each mini-batch（梯度下降的一种）
    - the adversarial network first updates its discriminative network to tell apart the true samples (generated using the prior) from the generated samples (the hidden codes computed by the autoencoder). （D更新）
    - The adversarial network then updates its generator (which is also the encoder of the autoencoder) to confuse the discriminative network. （G更新）

AAE模型框架

- For GAN：
  - Positive Samples：*p(z)*
  - Negative Samples：*q(z)*
- For AE：
  - 输入：*x*
  - 输出：$z \sim q(z)$
- **训练完成后：AE的编码层便可以学到符合输出为$q(z)$分布的输入分布$p_d(x)$**
- **引申：也可以通过解码层输出x'（即$p(x'|z)$），且隐向量z符合$q(z)$分布**



目录

小结

- 生成式对抗网络（GAN, Generative Adversarial Networks ）是一种深度学习模型，是近年来复杂分布上无监督学习最具前景的方法之一。
- 模型通过框架中（至少）两个模块：
  - 生成模型（Generative Model）
  - 判别模型（Discriminative Model）
- 两个互相博弈学习产生预期的输出。
- 原始 GAN 理论中，并不要求 G 和 D 都是神经网络，只需要是能拟合相应生成和判别的函数即可。但实用中一般均使用深度神经网络作为 G 和 D 。
- 一个优秀的GAN应用需要有良好的训练方法，否则可能由于神经网络模型的自由性而导致输出不理想。