非参数统计
第十章 密度估计

制作：邓明华

2023 秋季学期

本章目录

本节目录

密度估计问题

- 问题：给定 $X_1, X_2, \cdots, X_n \sim F(x)$, 如何从样本估计出密度函数 $f(x)$?
- 对于参数问题，我们需要事先假定密度的函数形式 (比如正态分布只包含两个参数：均值和方差); 从而密度估计问题转化为参数估计问题;
- 非参数密度估计直接从样本估计出密度的函数形式

直方图

- 直方图方法先将区间等分，用落入区间的样本数来近似函数在区间上的取值，即

$$f_n(x) = \begin{cases} \dfrac{\sum_{i=1}^{n} I_k(X_i)}{nh_n}, & \text{如果} x \in I_k, k = 1, 2, \cdots, K. \\ 0, & \text{其它}. \end{cases}$$

其中 $h_n = (b-a)/K$ 称为窗宽，I_k 为区间，

$$I_k(x) = \begin{cases} 1, & \text{如果} x \in I_k. \\ 0, & \text{其它}. \end{cases}$$

直方图

- 设有密度函数 $f(x)$,

$$\begin{aligned} p(x \in I_k) &= \int_{I_k} f(x)\,dx \approx f(x) \times |I_k|, \quad |I_k| \to 0 \\ &= \frac{n_i}{n} \end{aligned}$$

- 所以

$$f(x) \approx \frac{n_i}{nh}, \quad h \text{ 为窗宽}$$

定理 10.1

- **定理 10.1：** 设 $t_i^{(n)}$ 是区间 $[a,b]$ 的窗宽为 h_n 的分割点 $h_n = t_{i+1}^{(n)} - t_i^{(n)}$,

$$f_n(x) = \frac{1}{nh_n}\sum_{i=1}^{n} I_{(t_k^{(n)}, t_{k+1}^{(n)})}(x_i), \quad x \in [a,b]$$

若 $f(x)$ 在 x 点连续，且 $\lim_{n\to+\infty} h_n = 0$, $\lim_{n\to+\infty} nh_n = +\infty$, 则对 $\forall \epsilon > 0$ 有：

$$\lim_{n\to+\infty} Pr(|f_n(x) - f(x)| \geq \epsilon) = 0$$

定理证明

- 证明：因为 $f(x)$ 在 x 点连续，$\exists \delta > 0, K_0 > 0$ 使得

$$\forall y \in [x-\delta, x+\delta], \quad f(y) \leq K_0,$$

由条件 $h_n \to 0$, 因此 $\exists N$, 当 $n > N$ 时, $h_n < \delta$, 使得 $x \in (t_k^{(n)}, t_{k+1}^{(n)})$, 且包含 x 的整个 Bin 落入区间 $[x-\delta, x+\delta]$. 由切比雪夫不等式

$$Pr(|f_n(x) - f(x)| \geq \epsilon) \leq \frac{1}{\epsilon^2} E(f_n(x) - f(x))^2$$

定理证明

- 右边可以改写为

$$\frac{1}{\epsilon^2}\left[E(f_n(x)-Ef_n(x))^2+(Ef_n(x)-f(x))^2\right]$$

- 一方面

$$f_n(x)-Ef_n(x)=\frac{1}{nh_n}\sum_{i=1}^{n}g_n(x_i)$$

$$g_n(x_i)=I_{[t_k^{(n)},t_{k+1}^{(n+1)}]}(x_i)-EI_{[t_k^{(n)},t_{k+1}^{(n+1)}]}(x_i)$$

$$Eg_n(x_i)=0$$

$$Eg_n^2(x_i)\le\int_{t_k^{(n)}}^{t_{k+1}^{(n)}}f(y)\,dy\le K_0h_n$$

定理证明

- 于是

$$E(f_n(x) - Ef_n(x))^2 \leq \frac{nK_0h_n}{n^2h_n^2} = \frac{K_0}{nh_n} \to 0$$

- 另一方面

$$Ef_n(x) = \frac{1}{h_n}\int_{t_k^{(n)}}^{t_{k+1}^{(n)}} f(y)\,dy \to f(x)$$

- 只要 h_n 足够小即可.

说明

- 进一步可以证明：
- **定理 10.2：** 若 $f(x)$ 在 $(-\infty, +\infty)$ 上一致连续, 且 $\exists \delta > 0$, 使得

$$\int_{-\infty}^{+\infty} |x|^{\delta} f(x)\, dx < +\infty,$$

又 $\lim_{n \to +\infty} h_n = 0$, $h_n \geq (\ln n)^2 / n$, 则

$$Pr\left(\lim_{n \to +\infty} \mathrm{Sup} |f_n(x) - f(x)| = 0 \right) = 1$$

直方图高维推广

- 将空间分割成若干小区域 $R, V = |R|$, 若有 n 个观察值 $x_1, x_2, \cdots, x_n$, 设 n_i 是落入空间 R 内的样本个数

$$p(x) \approx \frac{n_i}{nV}$$

- 说明：窗宽影响很大
  1. 太大：抹杀了细节，多峰分布可能会被抹平。
  2. 太小：不连续，波动大

举例

- 生成两个正态混合样本, 若混合比例为 p, 可以先生成 $(0,1)$ 上的均匀随机数 $u \in U(0,1)$, 若 $u \leq \pi$, 取第一个正态总体，反之取第二个总体；
- 选取不同的窗宽做直方图，看得到什么结果

本节目录

Rosenblatt 方法

- 从经验分布函数出发

$$f(x) \approx \frac{F(x+h) - F(x-h)}{2h}$$

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x), \quad F_n(x) \to F(x)$$

- 故定义

$$\begin{aligned} f_n^R(x) &= \frac{F_n(x+h) - F_n(x-h)}{2h} \\ &= \frac{1}{2nh}\sum_{i=1}^{n} I(x-h \le X_i \le x+h) \\ &= \frac{1}{nh}\sum_{i=1}^{n} K_0(\frac{x-X_i}{h}), \quad K_0(u) = \frac{1}{2} I(|u| \le 1) \end{aligned}$$

Rosenblatt 方法

- 从形式上看，Rosenblatt 方法和直方图方法似乎差不了多少。
- 其不同点在于直方图的分割区间是事先给定的，而 Rosenblatt 方法不把分割先固定，而让区间随着估计点 x 跑，使 x 始终处在区间的中心位置。
- 实际上可以证明：当 $h_n \to 0$, $nh_n \to +\infty$ 时，有

$$\int E(f_n^R(x) - f(x))^2 dx = O(n^{-\frac{4}{5}})$$

$$\int E(f_n(x) - f(x))^2 dx = O(n^{-\frac{2}{3}})$$

即 Rosenblatt 方法优于直方图方法。

本节目录

核估计

- 如果把 Rosenblatt 方法中 $[-1,1]$ 上的均匀分布替换为一般的概率分布函数

$$K(u) \geq 0, \quad \int_{-\infty}^{+\infty} K(u)du = 1$$

- 核估计

$$f_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)$$

- $K(x)$ 称为 核函数，h_n 称为 窗宽。

常用核函数选取

- Parzen 窗 (Rosenblatt 方法)

$$\frac{1}{2} I(|u| \leq 1)$$

- 三角核函数

$$(1 - |u|) I(|u| \leq 1)$$

- Epanechikov 核函数

$$\frac{1}{4}(1 - u^2) I(|u| \leq 1)$$

- 四次 (quartic) 核函数

$$\frac{15}{16}(1 - u^2)^2 I(|u| \leq 1)$$

常用核函数选取

- 三权 (Triweight) 核函数

$$\frac{35}{32}(1-u^2)^3 I(|u| \leq 1)$$

- 余弦核函数

$$\frac{\pi}{4}\cos(\frac{\pi}{2}u) I(|u| \leq 1)$$

- 高斯核函数

$$\frac{1}{\sqrt{2\pi}}\exp(-\frac{u^2}{2})$$

- 指数核函数

$$\frac{1}{2}\exp(-|u|)$$

核估计

- 后面我们会看到，核函数的形式不是影响密度估计的关键因素，窗宽才是影响密度估计的关键

- 问题：如何选择合适的窗宽？
- 首要问题：如何评价密度估计的好坏？

评价标准

- 任意固定 x, 均方误差 (Mean squared error)

$$\begin{aligned} MSE(f_n(x)) &= E_F\left[f_n(x) - f(x)\right]^2 \\ &= \left[E_F(f_n(x)) - f(x)\right]^2 + Var_F(f_n(x)) \end{aligned}$$

- 其中前者是偏差平方，后者是估计的方差
- 由于密度本身是函数，要对所有 x 都考虑，一种方法是再对 x 积分，即所谓的 积分均方误差 (Integral mean squared error)

评价标准

- 积分均方误差

$$
\begin{aligned}
& MISE(f_n(x)) \\
= & \int E\left[f_n(x) - f(x)\right]^2 dx \\
= & \int MSE(f_n(x))\, dx \\
= & \int \left[E(f_n(x)) - f(x)\right]^2 dx + \int Var(f_n(x))\, dx
\end{aligned}
$$

前者为积分偏差平方和，后者是积分方差

核估计期望计算

- **命题 10.3：** 核估计的均值为

$$E_F(f_n(x)) = \int K(y)f(x - h_n y)\,dy$$

- 证明：

$$\begin{aligned}
\frac{1}{h_n} E_F K(\frac{x - X}{h_n}) &= \int_{-\infty}^{+\infty} K(\frac{x - t}{h_n}) f(t)\,dt \\
&\overset{\frac{x-t}{h_n}=y}{=} \frac{1}{h_n} \int_{+\infty}^{-\infty} k(y)f(x - h_n y)(-h_n)\,dy \\
&= \int_{-\infty}^{+\infty} k(y)f(x - h_n y)\,dy
\end{aligned}$$

- 所以

$$E_F(f_n(x)) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_n} E_F K(\frac{x - X}{h_n}) = \int K(y)f(x - h_n y)\,dy$$

核估计方差计算

- **命题 10.4：** 核估计的方差为

$$Var(f_n(x)) = \frac{1}{nh_n}\int K^2(y)f(x-h_ny)\,dy - \frac{1}{n}\left[\int K(y)f(x-h_ny)\,dy\right]^2$$

- 证明：因为 X_i 相互独立，

$$\begin{aligned} Var(f_n(x)) &= \frac{1}{n^2h^2}\sum_{i=1}^{n}\left\{EK^2(\frac{x-X}{h}) - [EK(\frac{x-X}{h})]^2\right\} \\ &= \frac{n}{n^2h^2}\left[h\int K^2(y)f(x-hy)\,dy - (h\int K(y)f(x-hy)\,dy)^2\right] \\ &= \frac{1}{nh}\int K^2(y)f(x-hy)\,dy - \frac{1}{n}\left[\int K(y)f(x-hy)\,dy\right]^2 \end{aligned}$$

最佳窗宽选取

- 基于 MISE

$$h_{opt} = \underset{h}{\text{Argmin}}\, MISE(f_n(x))$$

- 基于 MSE

$$h_{opt} = \underset{h}{\text{Argmin}}\, MSE(f_n(x))$$

条件

- 以下不妨设核函数 $K(t)$ 满足

$$\int tK(t)\,dt = 0$$

$$K_2 = \int t^2 K(t)\,dt \neq 0$$

- $f''(x)$ 绝对连续, 且 $\int |f''(x)|^2\,dx < \infty$,

$$h = h_n \to 0 \quad (n \to +\infty)$$

定理

- **定理 10.5：** 在上述条件成立时，对核估计有

$$
\begin{aligned}
MSE(f_n(x)) &= E[f_n(x) - f(x)]^2 \\
&= \frac{1}{4}K_2^2 h_n^4 [f''(x)]^2 + \frac{f(x)\int K^2(x)\,dx}{nh_n} + O(\frac{1}{n}) + O(h_n^6)
\end{aligned}
$$

$$
\begin{aligned}
MISE(f_n(x)) &= \int E[f_n(x) - f(x)]^2\,dx \\
&= \frac{1}{4}K_2^2 h_n^4 \int [f''(x)]^2\,dx + \frac{\int K^2(x)\,dx}{nh_n} + O(\frac{1}{n}) + O(h_n^6)
\end{aligned}
$$

定理证明

- 首先看偏差

$$
\begin{aligned}
& E_F[f_n(x)] - f(x) \\
=& \int_{-\infty}^{+\infty} K(y)[f(x - h_n y) - f(x)]\, dy \\
=& \int_{-\infty}^{+\infty} K(y)[f(x - h_n y) - f(x) - h_n y f'(x)]\, dy \\
=& \frac{h_n^2}{2} \int_{-\infty}^{+\infty} K(y) y^2 f''(x - \theta h_n y)\, dy
\end{aligned}
$$

其中 $0 \le \theta = \theta(h_n) \le 1$.

定理证明

- 由控制收敛定理

$$E_F(f_n(x)) - f(x) = \frac{1}{2} f''(x) K_2 h_n^2 + O(h_n^3)$$

- 于是，偏差 (bias) 可表示为

$$[E_F(f_n(x)) - f(x)]^2 = \frac{1}{4} (f''(x))^2 K_2^2 h_n^4 + O(h_n^6)$$

定理证明

- 前面已经证明

$$Var(f_n(x)) = \frac{1}{nh_n}\int K^2(y)f(x-h_ny)\,dy - \frac{1}{n}\left[\int K(y)f(x-h_ny)\,dy\right]^2$$

- 当 $n \to +\infty$ 时，$f(x-h_ny) \to f(x)$, 于是

$$Var(f_n(x)) = \frac{1}{nh_n}f(x)\int K^2(y)\,dy + O(\frac{1}{n})$$

定理证明

- 将上两式合在一起

$$
\begin{aligned}
MSE(f_n(x)) &= [E(f_n(x)) - f(x)]^2 + Var(f_n(x)) \\
&= \frac{1}{4}K_2^2 h_n^4 [f''(x)]^2 + \frac{f(x)\int K^2(x)\,dx}{nh_n} + O\left(\frac{1}{n}\right) + O(h_n^6)
\end{aligned}
$$

- 再对 x 积分得

$$
MISE(f_n(x)) = \frac{1}{4}K_2^2 h_n^4 \int [f''(x)]^2\,dx + \frac{\int K^2(x)\,dx}{nh_n} + O\left(\frac{1}{n}\right) + O(h_n^6)
$$

最优窗宽选取

- 令

$$A = \frac{1}{4} K_2^2 \int [f''(x)]^2 dx$$

$$B = \int K^2(x) dx$$

$$G(h) = Ah^4 + \frac{B}{nh}$$

- 于是对 MISE 的优化等价于对 $G(h)$ 进行优化

最优窗宽选取

- 对 $G(h)$ 求导

$$G'(h) = 4Ah^3 - \frac{B}{nh^2}$$

- 由 $G'(h) = 0$ 得

$$h_{opt} = \left(\frac{B}{4An}\right)^{\frac{1}{5}}$$

- 代入得

$$G(h_{opt}) = \frac{5}{4}(4A)^{\frac{1}{5}} B^{\frac{4}{5}} \left(\frac{1}{n}\right)^{\frac{4}{5}}$$

最优窗宽选取

- **定理 10.6：** 在上述条件下，最优窗宽为

$$\begin{aligned} & h_{opt} \\ =& K_2^{-\frac{2}{5}} \left[\int K^2(u)\,du\right]^{\frac{1}{5}} \left[\int f''(x)^2\,dx\right]^{-\frac{1}{5}} n^{-\frac{1}{5}} \end{aligned}$$

相应的最优积分均方误差为

$$\begin{aligned} & \mathrm{MISE}(h_{opt}) \\ =& \frac{5}{4}\left(K_2^2 \int f''(x)^2\,dx\right)^{\frac{1}{5}} \left(\int K^2(u)\,du\right)^{\frac{4}{5}} n^{-\frac{4}{5}} \end{aligned}$$

本节目录

K 近邻密度估计

- Rosenblatt 方法是每个点都选用相同体积内的样本，如果窗宽过大，这分布较密的区域受到多个样本点的支持，可能过于光滑; 而分别稀疏的位置受到很少样本点的支持, 估计不光滑。一种替代的方法是 K-近邻估计。其基本想法是让每个位置都用邻近的 K 个样本来估计密度

K 近邻密度估计

- 设简单随机样本 $X_1, \cdots, X_n \sim F(x)$, $f(x)$ 是分布 $F(x)$ 的密度函数，对于任意固定点 x, 记 $a_n(x)$ 为最小的正数 a, 使得区间 $[x-a, x+a]$ 中恰好包含 K 个样本点, 那么该点 x 的 K 近邻密度估计为

$$f_n(x) = \frac{K}{2a_n(x)n}$$

- 落入区间的样本个数固定，但区间长度随机

定理 10.7

- **定理 10.7：** 对于固定 n 和样本 $X_1, cdots, X_n$, K 近邻密度估计满足

$$\int_{-\infty}^{+\infty} f_n(x)\,dx = +\infty$$

- 分析：关键在于区间长度 $a_n(x)$ 在 x 充分大时和 x 是同阶的。原因是样本固定了，只能覆盖有限的区间。

定理 10.7 证明

- 证明：记 $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$, 那么

$$
\begin{aligned}
a_n(x) &= X_{(k)} - x \qquad && x < X_{(1)} \\
a_n(x) &= x - X_{(n-k+1)} && x > X_{(n)}
\end{aligned}
$$

于是

$$
\begin{aligned}
&\int_{-\infty}^{+\infty} f_n(x)\,dx \\
&\ge \int_{x \le X_{(1)}} f_n(x)\,dx + \int_{x \ge X_{(n)}} f_n(x)\,dx \\
&= \int_{-\infty}^{X_{(1)}} \frac{K}{2n(X_{(k)} - x)}\,dx + \int_{X_{(n)}}^{+\infty} \frac{K}{2n(x - X_{(n-k+1)})}\,dx \\
&\to +\infty
\end{aligned}
$$

附注

- 从证明过程可以看出，

$$f_n(x) = O\left(\frac{1}{|x|}\right), \quad |x| \to +\infty$$

- 与待估 $f(x)$ 的尾部特征无关，因而 $f_n(x)$ 不能作为 $f(x)$ 的整体估计。

定理 10.8

- **定理 10.8：** 当 $n \to +\infty$ 时，若 $K_n \to +\infty$, $K_n/n \to 0$, 则对任意固定点 x, 若 $x \in Supp(f)$, 且函数 $f(x)$ 在 x 点处连续，有

$$f_n(x) \xrightarrow{P} f(x), \quad n \to +\infty.$$

即 $\forall \epsilon > 0$,

$$Pr(|f_n(x) - f(x)| \geq \epsilon) \to 0, \quad n \to +\infty$$

- 说明：$f_n(x)$ 是连续点 x 的密度 $f(x)$ 的一个不错的估计.

定理 10.8 证明

- 证明：固定 $x \in Supp(f), \forall \epsilon > 0,$

$$
\begin{aligned}
& Pr(|f_n(x) - f(x)| \geq \epsilon) \\
=& Pr\left(\frac{K}{2na_n(x)} > f(x) + \epsilon\right) + Pr\left(\frac{K}{2na_n(x)} < f(x) - \epsilon\right) \\
=& Pr\left(a_n(x) < \frac{K}{2n(f(x) + \epsilon)}\right) + Pr\left(a_n(x) > \frac{K}{2n(f(x) - \epsilon)}\right)
\end{aligned}
$$

定理 10.8 证明

- 令

$$p_n = \int_{x+\frac{K}{2n(f(x)+\epsilon)}}^{x-\frac{K}{2n(f(x)+\epsilon)}} f(t)\,dt$$

- 考虑二项分布随机变量 $y_n \sim B(n, p_n)$,
- 由 $a_n(x)$ 的定义可得

$$Pr\left(a_n(x) < \frac{K}{2n(f(x)+\epsilon)}\right) = Pr(y_n \geq K)$$

定理 10.8 证明

- 另一方面，由于 $K/n \to 0$, 即 n 足够大时，区间

$$\left[x - \frac{K}{2n(f(x)+\epsilon)}, x + \frac{K}{2n(f(x)+\epsilon)}\right]$$

为 x 点附近的小邻域，由 $f(x)$ 连续，则在此区间上有

$$|f(t) - f(x)| < \frac{\epsilon}{2}$$

- 于是在上面的区间上,

$$f(t) < f(x) + \frac{\epsilon}{2}$$

定理 10.8 证明

- 那么

$$
\begin{aligned}
p_n &= \int_{x+\frac{K}{2n(f(x)+\epsilon)}}^{x-\frac{K}{2n(f(x)+\epsilon)}} f(t)\,dt \\
&< (f(x)+\frac{\epsilon}{2}) \times 2 \times \frac{K}{2n(f(x)+\epsilon)} \\
&= \frac{f(x)+\frac{\epsilon}{2}}{f(x)+\epsilon} \times \frac{K}{n} \\
&= c \times \frac{K}{n}
\end{aligned}
$$

- 其中

$$
c = \frac{f(x)+\frac{\epsilon}{2}}{f(x)+\epsilon}, \quad 0 < c < 1.
$$

定理 10.8 证明

- 于是

$$
\begin{aligned}
Pr(y_n \geq K) &= Pr\left(\frac{y_n}{n} - p_n \geq \frac{K}{n} - p_n\right) \\
&\leq Pr\left(\frac{y_n}{n} - p_n \geq \frac{K}{n} - c\frac{K}{n}\right) \\
&= Pr\left(\frac{y_n}{n} - p_n \geq (1-c)\frac{K}{n}\right)
\end{aligned}
$$

定理 10.8 证明

- 由切比雪夫不等式

$$Pr(y_n \geq K) \leq \frac{\frac{1}{n^2} np_n(1-p_n)}{(1-c)^2(\frac{K}{n})^2}$$

$$\leq \frac{n \times \frac{cK}{n} \times \frac{1}{n^2}}{(1-c)^2(\frac{K}{n})^2} = \frac{c}{(1-c)^2 K} \to 0$$

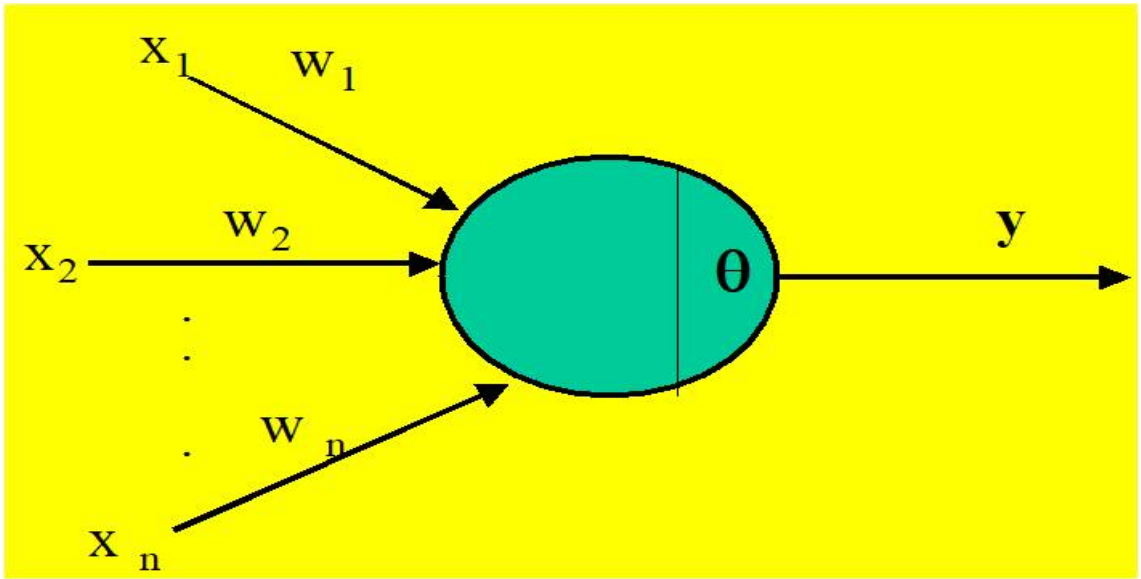
- 同理可证

$$Pr\left(a_n(x) > \frac{K}{2n(f(x)-\epsilon)}\right) \to 0.$$

本节目录

MP 模型

- McCulloch, Pitts 神经元模型 (1943)



$$y = g(\sum_k \omega_k x_k - \theta)$$

激活函数

- 符号函数

$$g(x) = \begin{cases} 1, & x > 0, \\ 0, & x \leq 0. \end{cases}$$

- Sigmoid 函数

$$g(x) = \frac{1}{1 + \exp(-ax)}$$

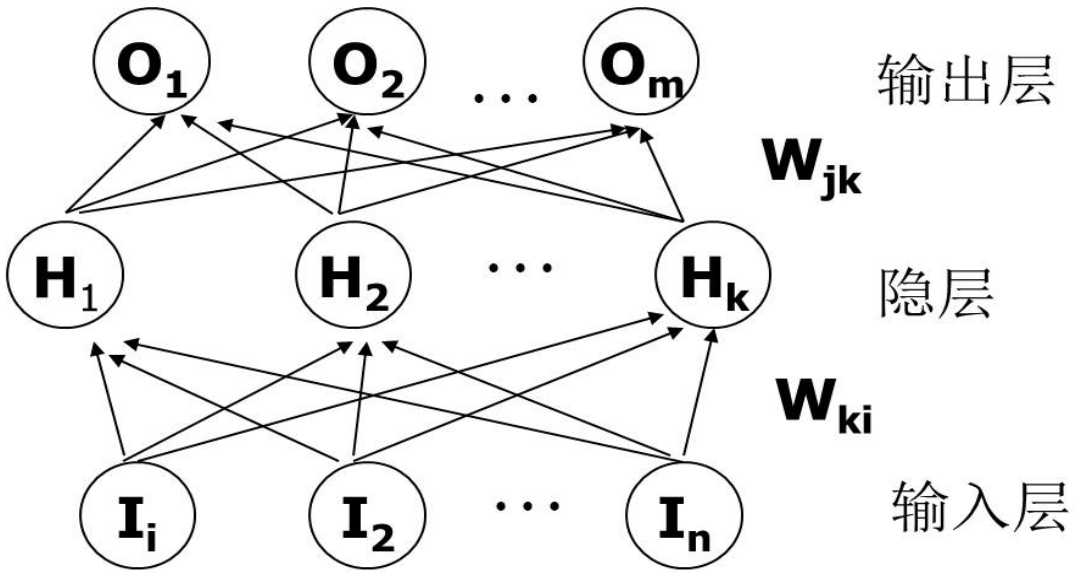
$$g(x) = \frac{2}{\pi}\arctan(ax)$$

- ReLU

$$g(x) = \max\{x, \theta\}$$

多层感知器网络 (MLP)

- 从下到上，底层的输出是高层的输入。
- 多层感知器至少要有三层：输入层、输出层与隐层。



函数的参数化框架

- 一些记号
  1. o_j^μ: 样本 μ 对应于输出单元 j 的输入。
  2. O_j^μ: 样本 μ 对应于输出单元 j 的输出。
  3. h_k^μ: 样本 μ 对应于隐单元 k 的输入。
  4. H_k^μ: 样本 μ 对应于隐单元 k 的输出。

$$
\begin{cases}
h_k^\mu = \sum_i w_{ki} I_i^\mu \\
H_k^\mu = g(\sum_i w_{ki} I_i^\mu) \\
o_j^\mu = \sum_k w_{jk} H_k^\mu = \sum_k w_{jk} g(\sum_i w_{ki} I_i^\mu) \\
O_j^\mu = g(\sum_k w_{jk} H_k^\mu) = g(\sum_k w_{jk} g(\sum_i w_{ki} I_i^\mu))
\end{cases}
$$

有监督学习

- 理想输出 (Teacher)

$$\{T_j^\mu, j = 1, \cdots, m; \mu = 1, \cdots, n\}$$

- 实际输出

$$O_j^\mu = g(\sum_k w_{jk} H_k^\mu) = g(\sum_k w_{jk} g(\sum_i w_{ki} I_i^\mu))$$

- 理想输出与实际输出的误差函数;

$$\begin{cases} E(w) = \sum_{j,\mu} (T_j^\mu - O_j^\mu)^2 \\ O_j^\mu = f(I^\mu; \omega) \end{cases}$$

参数学习

- 梯度下降方法以修改权值。

$$\begin{cases} E(w) = \sum_{j,\mu} (T_j^\mu - O_j^\mu)^2 \\ \Delta w_{jk} = -\eta \dfrac{\partial E(w)}{\partial w_{jk}} \\ \Delta w_{ki} = -\eta \dfrac{\partial E(w)}{\partial w_{ki}} \end{cases}$$

- 误差学习

$$\begin{cases} \Delta w_{jk} = \eta \sum_{\mu} (T_j^\mu - O_j^\mu) g'(o_j^\mu) H_k^\mu \\ \qquad = \eta \sum_{\mu} \delta_j^\mu H_k^\mu \\ \Delta w_{ki} = \eta \sum_{\mu} \sum_{j} (T_j^\mu - O_j^\mu) g'(o_j^\mu) w_{jk} g'(h_k^\mu) I_i^\mu \\ \qquad = \eta \sum_{\mu} \left(g'(h_k^\mu) \sum_{j} \delta_j^\mu w_{jk} \right) I_i^\mu \\ \qquad = \eta \sum_{\mu} \delta_k^\mu I_i^\mu \end{cases}$$

误差后传算法

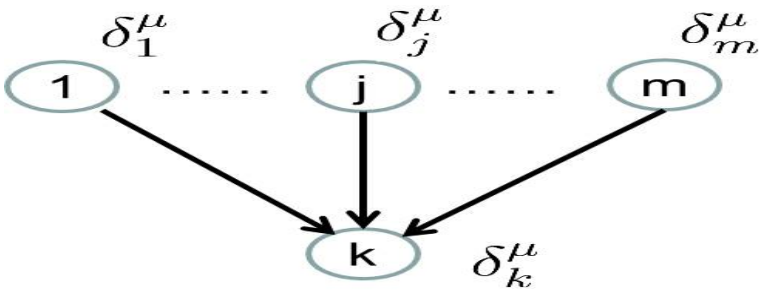
- 输出单元上的误差

$$\delta_j^\mu = g'(o_j^\mu)(T_j^\mu - O_j^\mu)$$

- 隐单元上的误差

$$\delta_k^\mu = g'(h_k^\mu)\sum_j \delta_j^\mu w_{jk}$$

- 隐单元上的误差实际上是输出单元上误差的线性组合 (向后传播)，这就是误差后传算法的由来。



神经网络通用逼近定理

- 只要给予三层 MLP 网络足够数量的隐层神经元，便可以实现以足够高精度来逼近任意一个在 R^n 的紧子集 (Compact subset) 上的连续函数。
- 通用逼近定理的数学描述：令 ϕ 为一单调递增、有界的非常数连续函数。记 m 维单元超立方体 $[0,1]^m$ 为 I_m，并记在 I_m 上的连续函数全体为 $C(I_m)$。则对任意实数 $\epsilon > 0$ 与函数 $f \in C(I_m)$, 存在整数 N、常数 $v_i, b_i \in R$ 与向量 $w_i \in \mathcal{R}^m$, $(i = 1, \cdots, n)$, 由此定义三层 MLP 网络函数：$F(x) = \sum_{i=1}^{N} v_i \phi(w_i^T x + b_i)$, 使得

$$|F(x) - f(x)| < \epsilon, \forall x \in I_m.$$

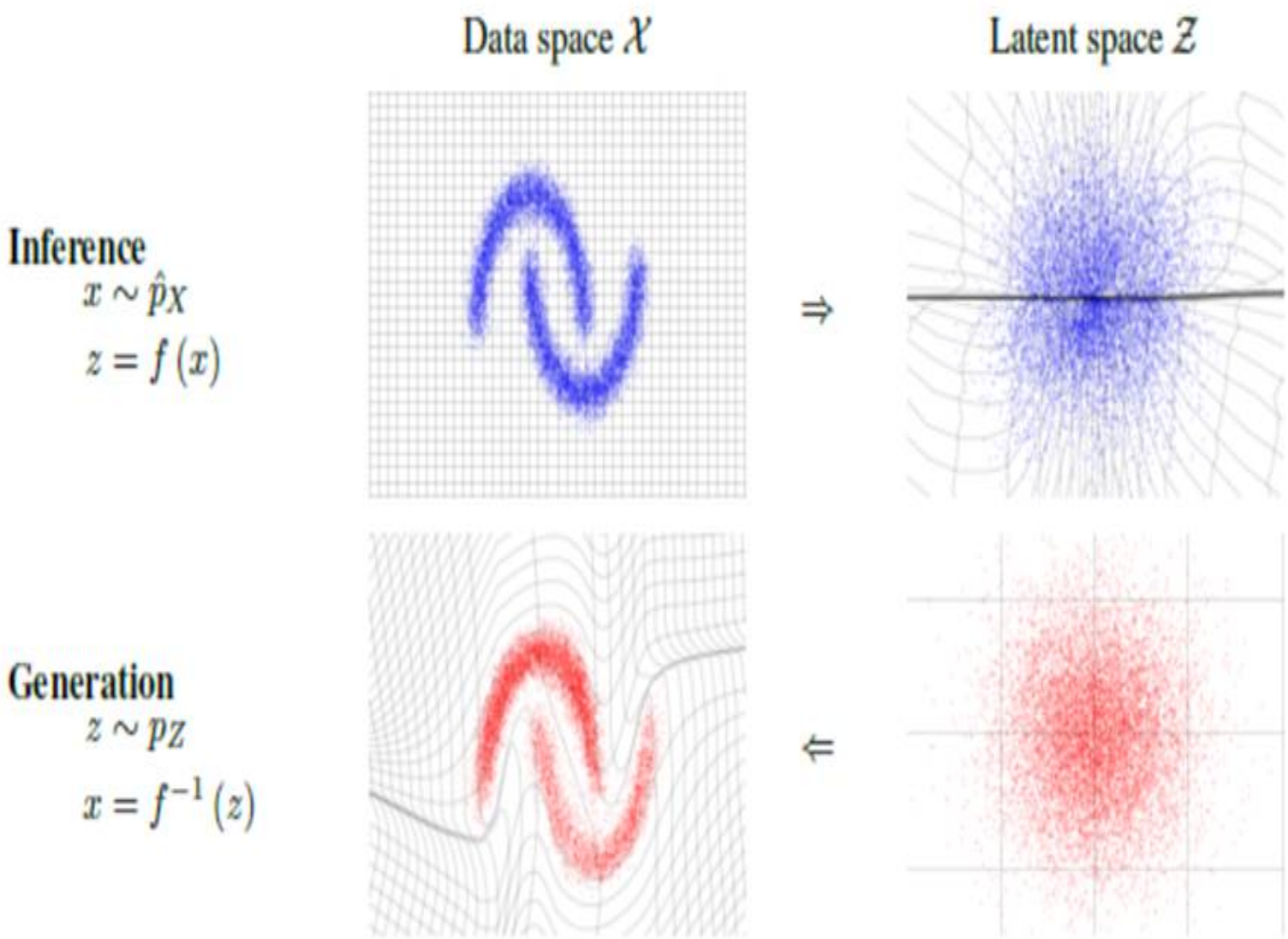
- 即三层 MLP 网络函数 $F(x)$ 在 $C(I_m)$ 里是稠密的。
- 替换上述 I_m 为 $\mathcal{R}^m$ 的任意紧子集，结论依然成立。

https://zhuanlan.zhihu.com/p/39030338

通用逼近定理相关文献

- George Cybenko. Approximation by Superpositions of a Sigmoidal Function. Math. Control Signals System 2: 304-312, 1989.
- Kurt Hornik, Maxwell Stinchcombe, Halbert White. Multilayer feedforward networks are universal approximators. Neural Networks, 2(5): 359–366, 1989.
- Andrew R. Barron, Universal Approximation Bounds for Superpositions of a Sigmoidal Function. IEEE Transaction on Information Theory 39(3), 1993.
- Moshe Leshno, Vladimir Ya. Lin, Allan Pinkus, Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. 6(6): 861–867, 1993.
- Kurt Hornik. Approximation capabilities of multilayer feedforward networks. Neural Networks, 4(2): 251–257, 1991.
- Jonathan W.Siegel, Jinchao Xu. Approximation rates for neural networks with general activation functions, Neural Networks 128: 313-321, 2020.
- Weinan E, Chao Ma, Lei Wu and Stephan Wojtowytsch. Towards a mathematical understanding of neural network-based machine learning: What we know and what we don’t. CSIAM Trans. Appl. Math. 1(4) 561-615, 2020.

基于流的生成式模型


Data space $\mathcal{X}$
Latent space $\mathcal{Z}$
Inference
$x \sim \hat{p}_X$
$z = f(x)$
Generation
$z \sim p_Z$
$x = f^{-1}(z)$

流方法

- 流方法其实就是空间变换，将样本空间 X 映射到隐变量空间 Z

$$x \overset{f_1}{\leftrightarrow} h_1 \overset{f_2}{\leftrightarrow} h_2 \cdots \overset{f_K}{\leftrightarrow} z$$

- 相应的对数概率似然函数变成

$$\log p_\theta(x) = \log p_\theta(z) + \log \left| det(dz/dx) \right|$$

$$\log p_\theta(z) + \sum_{i=1}^{K} \log \left| \det(dh_i/dh_{i-1}) \right|$$

$$\log \left| det(dh_i/dh_{i-1}) \right| = sum(\log \left| (diag(dh_i/dh_{i-1})) \right|)$$

- 最后一个等式需要变换有一定的特殊性 (比如是上三角的)
- 推荐博文: https://lilianweng.github.io/lil-log/2018/10/13/flow-based-deep-generative-models.html

NICE

- NICE: 非线性独立成分估计 (Non-linear Independent Component Estimation)
- 将 X 进行分块，构造上三角式的变换

$$\begin{cases} y_1 = x_1, \\ y_2 = x_2 + m(x_1). \end{cases}$$

- 这样变化的 Jacobi 为 1, 是保体积的变换

$$J = \begin{bmatrix} I_d & 0 \\ \frac{\partial m(x_1)}{\partial x_1} & I_{D-d} \end{bmatrix}$$

RealNVP

- RealNVP: 实非保体积变换 (Real Non-volume Preserve), 顾名思义就是为了去掉 NICE 保体积的限制。
- 坐标变换

$$\begin{cases} y_{1:d} = x_{1:d} \\ y_{d+1:D} = x_{d+1:D} \odot \exp(s(x_{1:d})) + t(x_{1:d}) \end{cases}$$

- 逆变换

$$\begin{cases} x_{1:d} = y_{1:d} \\ x_{d+1:D} = (y_{d+1:D} - t(x_{1:d})) \odot \exp(-s(x_{1:d}) \end{cases}$$

- 变换的 Jacobi 为,

$$J = \frac{\partial y}{\partial x^T} = \begin{bmatrix} I_d & 0 \\ \frac{\partial y_{d+1:D}}{\partial x_{1:d}^T} & \operatorname{diag}(\exp[s(x_{1:d})]) \end{bmatrix}$$

VAE

- VAE 是 variational autoencoder 的简写。

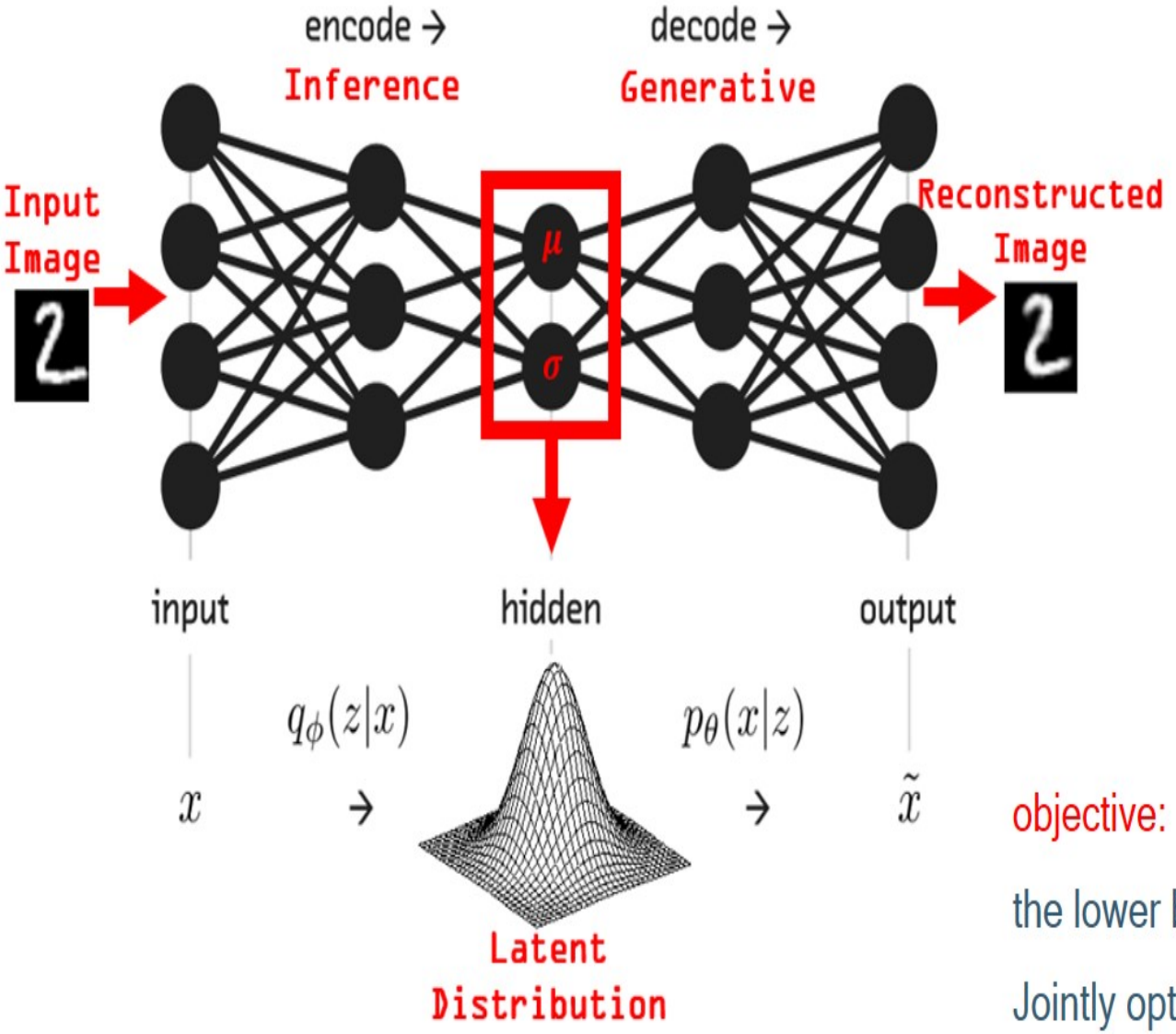
- 文献：Diederik P Kingma, Max Welling. Auto-Encoding Variational Bayes. arXiv:1312.6114 [stat.ML], 1 May 2014.
- 博文介绍：https://zhuanlan.zhihu.com/p/34998569

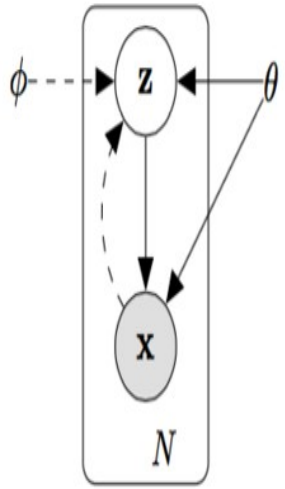
基本假设

- 观察 X 是高维向量，但 X 的分布模型 $p(x)$ 未知（不事先做任何的模型假设）；
- 假设：数据实际上处在一个低维流形上;

- 目标：希望生成 $p(x)$ 的样本。
- 思路：通过数据的低维表示 Z 上的随机抽样而给出 X 的随机样本，它们之间的变换可以用神经网络去逼近；

VAE 简介





objective:

the lower bound of $logP(x^{(i)})$.

Jointly optimized w.r.t. ϕ and θ

VAE 原理 (I)

- 由于后验概率 $p(z|x)$ 太复杂，用一个简单的分布 Q 来代替。如何找一个最好的 Q 呢？
- KL 散度

$$KL(Q(Z)||p(Z|X)) = \int Q(Z) \log \frac{Q(Z)}{p(Z|X)} dZ$$
$$= \int Q(Z) \log \frac{Q(Z)}{p(Z,X)} dZ + \log p(X)$$

- 得到对数似然 $p(x)$ 的下界 $L(Q)$, 称之为变分自由能

$$L(Q) = -\int Q(Z) \log \frac{Q(Z)}{p(Z,X)}$$

VAE 原理 (2)

- 利用 $p(X,Z) = p(Z) * p(X|Z)$ 得到 L(Q) 的表示

$$L(Q) = -KL(Q(Z)||p(Z)) + \int Q(Z)p(X|Z)$$

- 文章假设 $p(Z)$ 标准正态分布，而且假设 Q 也被是神经网络所参数化的正态分布

$$Q(Z|x^{(i)}) \sim N(\mu(x^{(i)}), \sigma^2(x^{(i)}))$$

- 因此有

$$\begin{aligned} &- KL(Q(Z|x^{(i)})||p(z)) \\ =&\frac{1}{2}\sum_{j=1}^{J}\left(1 + \log(\sigma(x^{(i)})^2) - \mu(x^{(i)})^2 - \sigma(x^{(i)})^2\right) \end{aligned}$$

VAE 原理 (3)

- 这里用到了正态分布的 KL 散度的计算

$$
\begin{aligned}
&KL[N(\mu_0, \Sigma_0)||N(\mu_1, \Sigma_1)] \\
=&\frac{1}{2}\left(tr(\Sigma_1^{-1}\Sigma_0) + (\mu_1 - \mu_0)^T\Sigma_1^{-1}(\mu_1 - \mu_0)\right. \\
&\left. -D + \log(\frac{det(\Sigma_1)}{\det(\Sigma_0)})\right)
\end{aligned}
$$

- 特别地

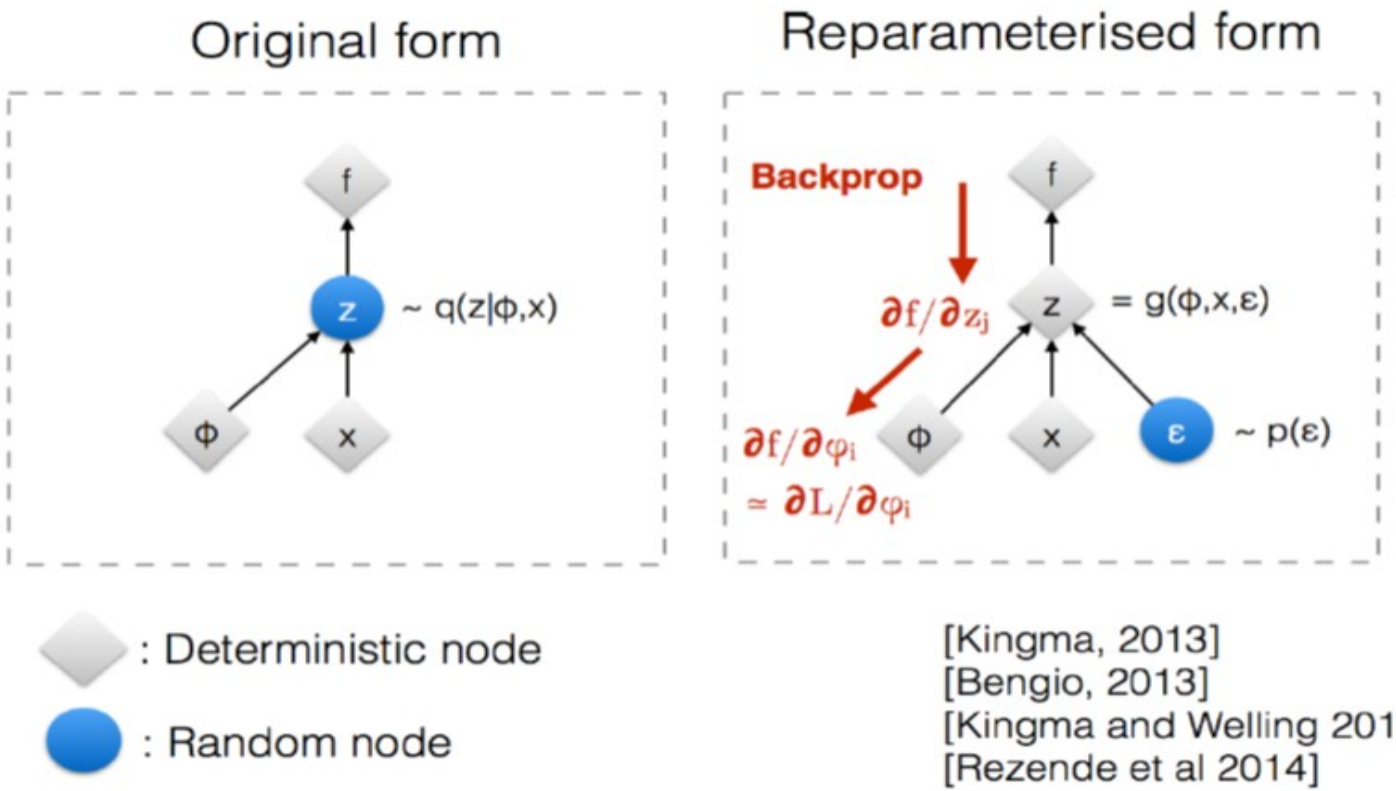
$$
\begin{aligned}
&KL[N(\mu(x), \Sigma_0(x))||N(0, I)] \\
=&\frac{1}{2}\left[tr(\Sigma(x)) + (\mu(x)^T\mu(x)) - D - \log(\det(\Sigma(x))\right]
\end{aligned}
$$

VAE 原理 (4): 隐变量抽样

- 再参数化策略 (Reparametrization trick)

$$z^{(i,l)} \sim N(\mu^{(i)}, \sigma^{2(i)})$$

$$z^{(i,l)} = \mu^{(i)} + \sigma^{(i)} \odot \epsilon_i, \epsilon_i \sim N(0,1)$$



VAE 原理 (5)

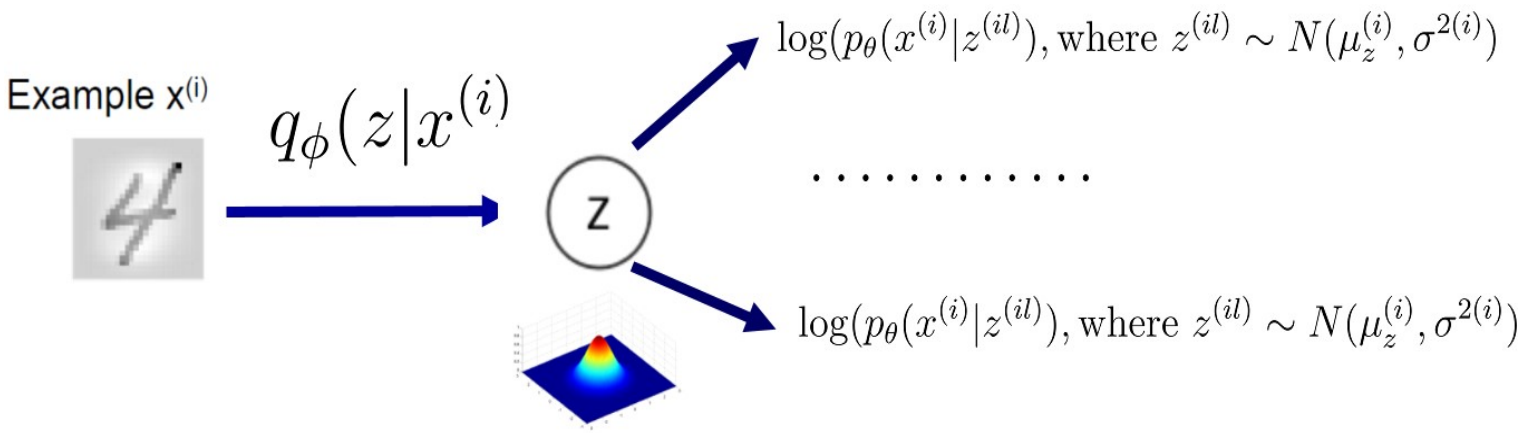
- 通过抽样来计算

$$\int Q(Z)p(x|z)$$

- 抽取隐变量

$$Z^{(i,l)}, l = 1, 2, \cdots, L; Z^{(i,l)} \sim Q(Z|x^{(i)})$$

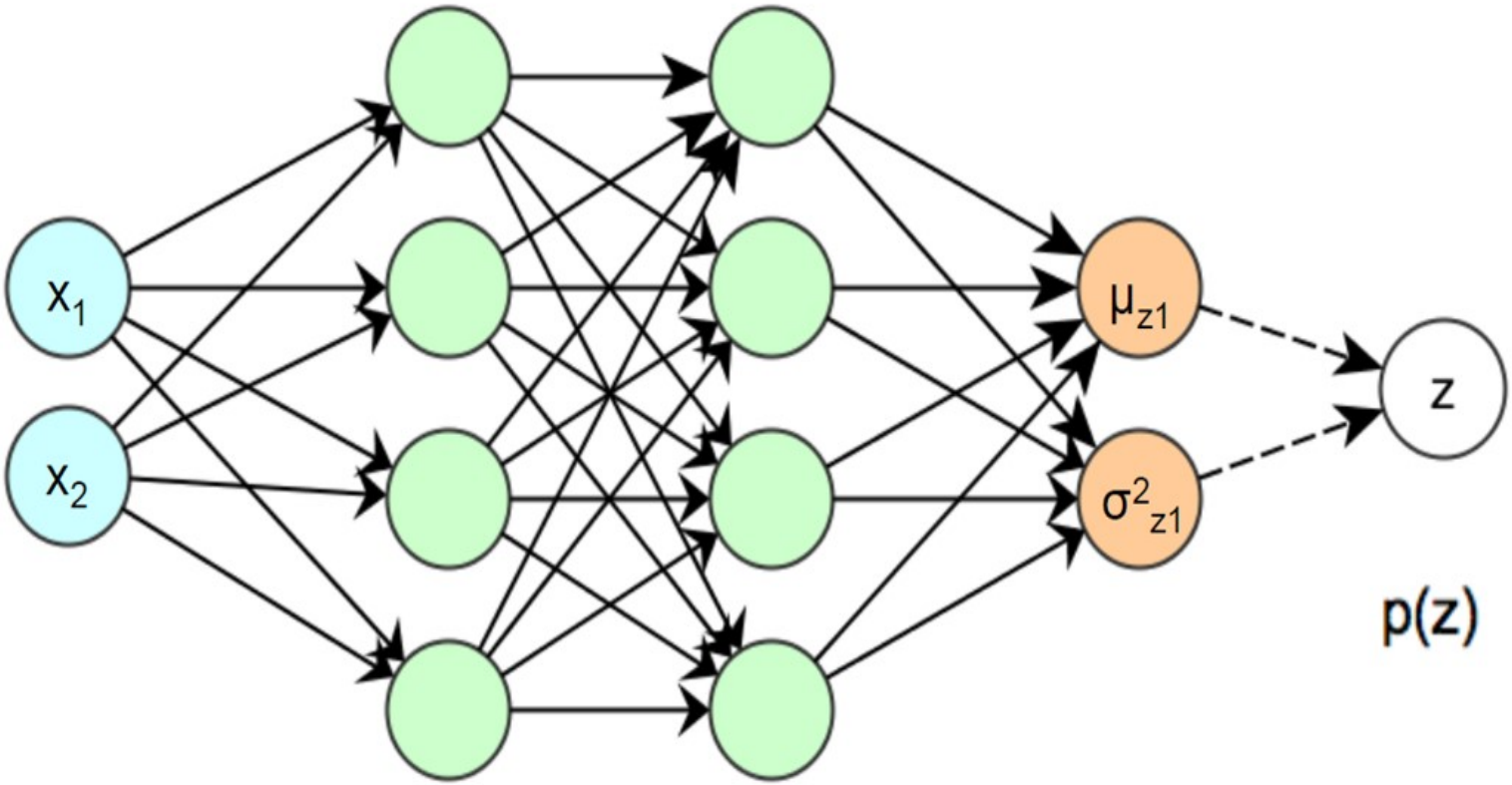
$$L(Q) \approx -KL(Q(Z|x^{(i)}||P(Z)) + \frac{1}{L}\sum_{l=1}^{L} \log P(x^{(i)}|z^{(i,l)})$$



编码器

- 前传网络 NN+ 正态分布

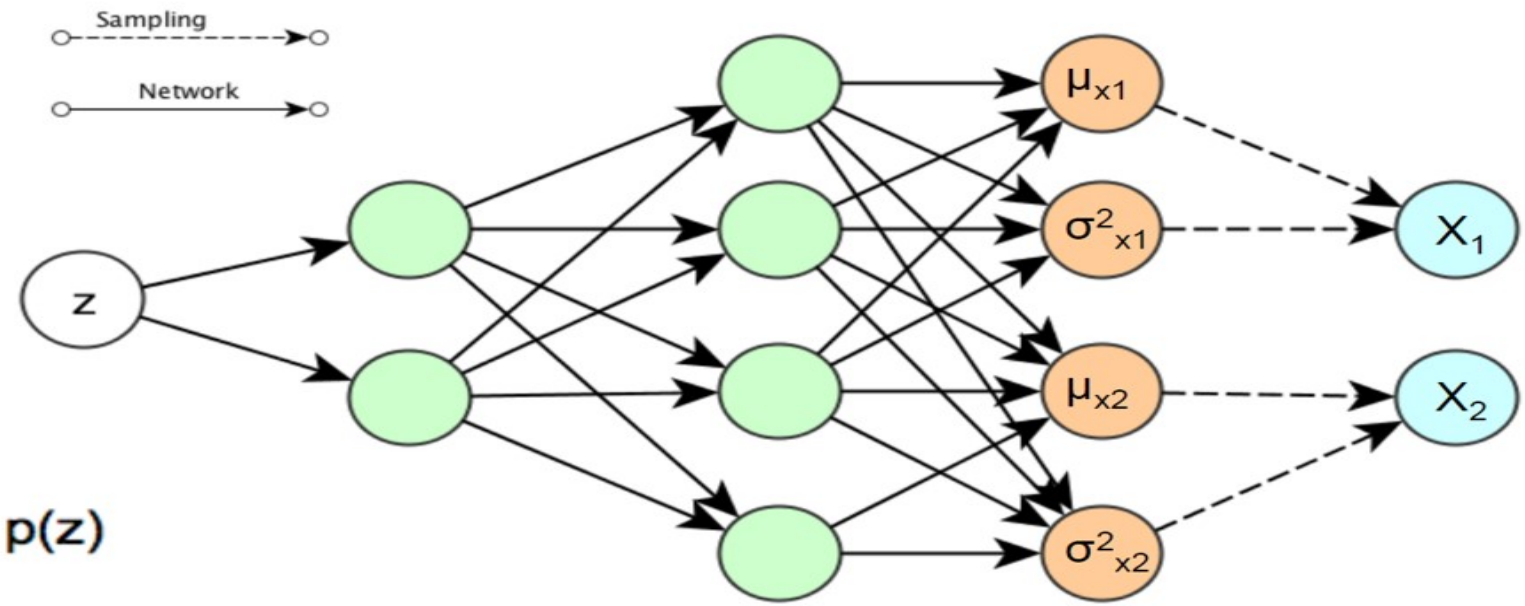
$$Q(Z|X) \sim N(z; \mu(x), \sigma(x)^2)$$



解码器

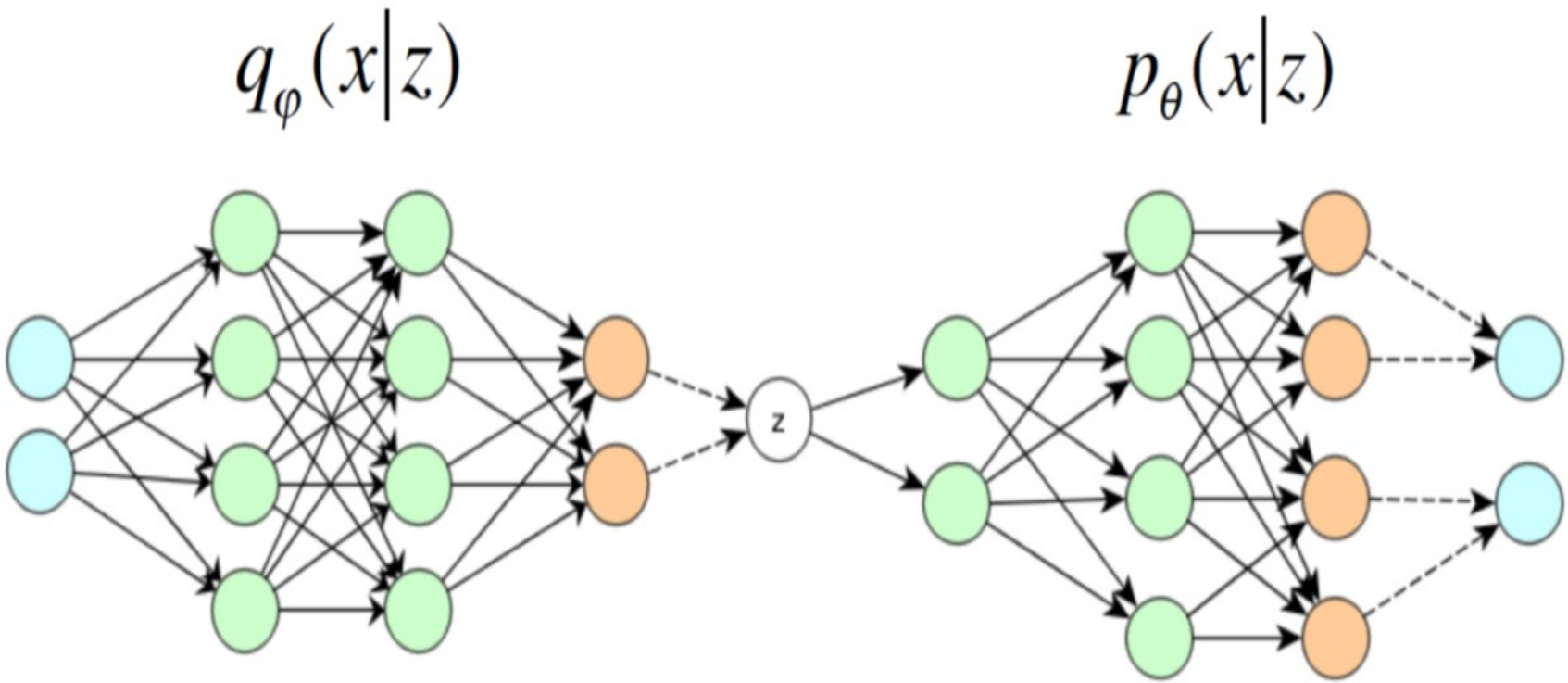
- 前传网络 NN+ 正态分布

$$X|Z \sim N(\mu_x, \sigma_x^2)$$



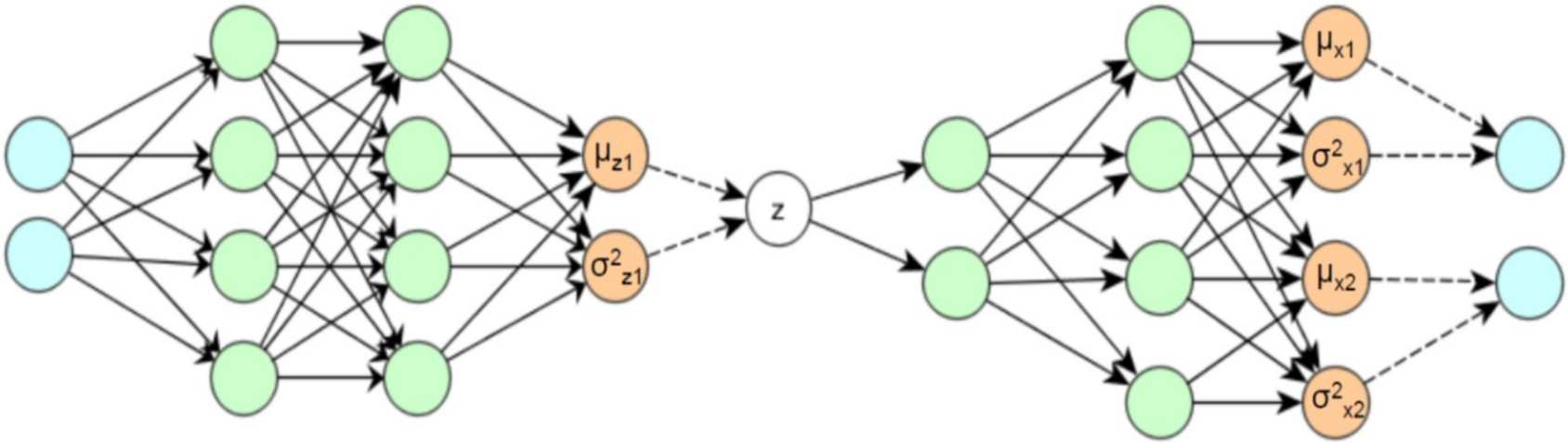
完整 AutoEncoder

- 通过 BP 算法学习参数
- 目标函数：极大化下界 L(Q)



VAE 略图

Prior p(z) ~ N(0,1) and p, q Gaussian, extension to dim(z) > 1 trivial



Cost: Regularisation

$$-D_{\mathrm{KL}}\left(q(z|x^{(i)})\|p(z)\right) = \frac{1}{2}\sum_{j=1}^{J}\left(1+\log(\sigma_{z_j}^{(i)^2}) - \mu_{z_j}^{(i)^2} - \sigma_{z_j}^{(i)^2}\right)$$

We use mini batch gradient decent to optimize the cost function over all $x^{(i)}$ in the mini batch

Cost: Reproduction

$$-\log\left(p(x^{(i)}|z^{(i)})\right) = \sum_{j=1}^{D}\frac{1}{2}\log(\sigma_{x_j}^2) + \frac{(x_j^{(i)} - \mu_{x_j})^2}{2\sigma_{x_j}^2}$$

Least Square for constant variance

实验效果

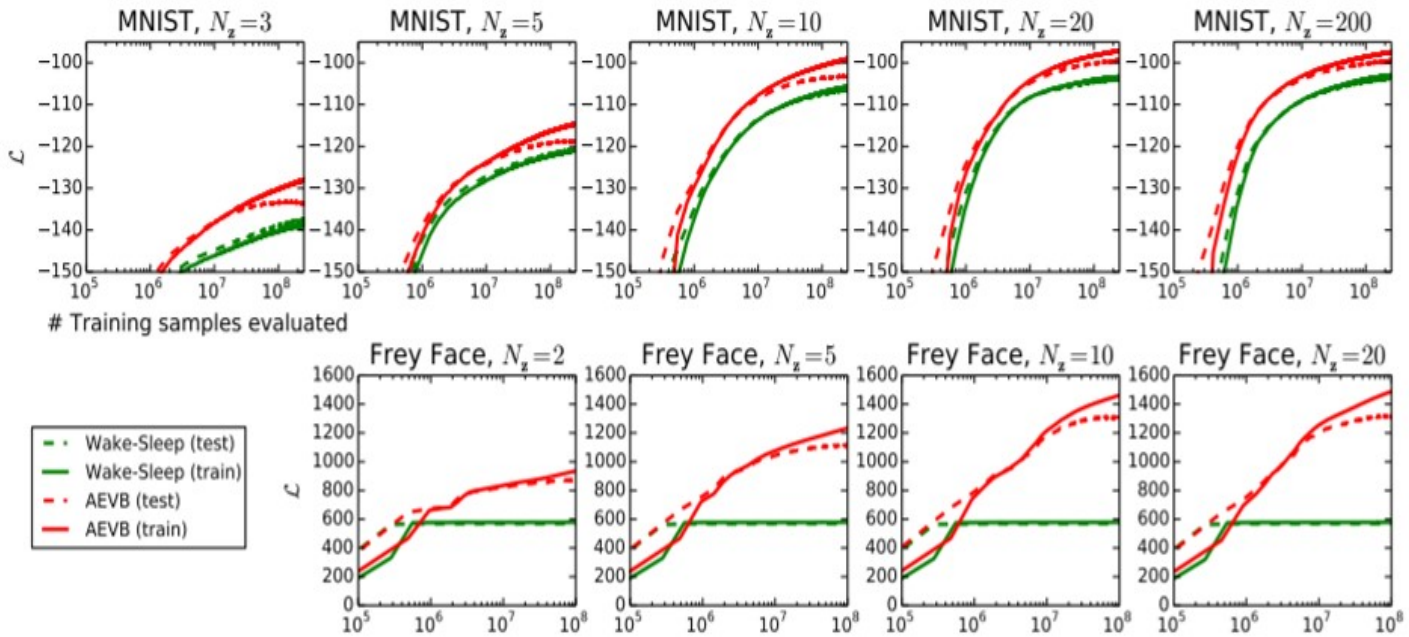


Figure 2: Comparison of our AEVB method to the wake-sleep algorithm, in terms of optimizing the lower bound, for different dimensionality of latent space ($N_{\mathbf{z}}$). Our method converged considerably faster and reached a better solution in all experiments. Interestingly enough, more latent variables does not result in more overfitting, which is explained by the regularizing effect of the lower bound. Vertical axis: the estimated average variational lower bound per datapoint. The estimator variance was small (< 1) and omitted. Horizontal axis: amount of training points evaluated. Computation took around 20-40 minutes per million training samples with a Intel Xeon CPU running at an effective 40 GFLOPS.

图像生成效果

(a) 2-D latent space (b) 5-D latent space (c) 10-D latent space (d) 20-D latent space

Figure 5: Random samples from learned generative models of MNIST for different dimensionalities of latent space.

RoundTrip

- 论文：Qiao Liu, Jiaze Xu, Rui Jiang and Wing Hung Wong. Density estimation using deep generative neural networks. PNAS 118(15): e2101344118, 2021.
- https://doi.org/10.1073/pnas.2101344118.

Significance

Density estimation is among the most fundamental problems in statistics. It is notoriously difficult to estimate the density of high-dimensional data due to the "curse of dimensionality." Here, we introduce a new general-purpose density estimator based on deep generative neural networks. By modeling data normally distributed around a manifold of reduced dimension, we show how the power of bidirectional generative neural networks (e.g., cycleGAN) can be exploited for explicit evaluation of the data density. Simulation and real data experiments suggest that our method is effective in a wide range of problems. This approach should be helpful in many applications where an accurate density estimator is needed.

Roundtrip 结构图

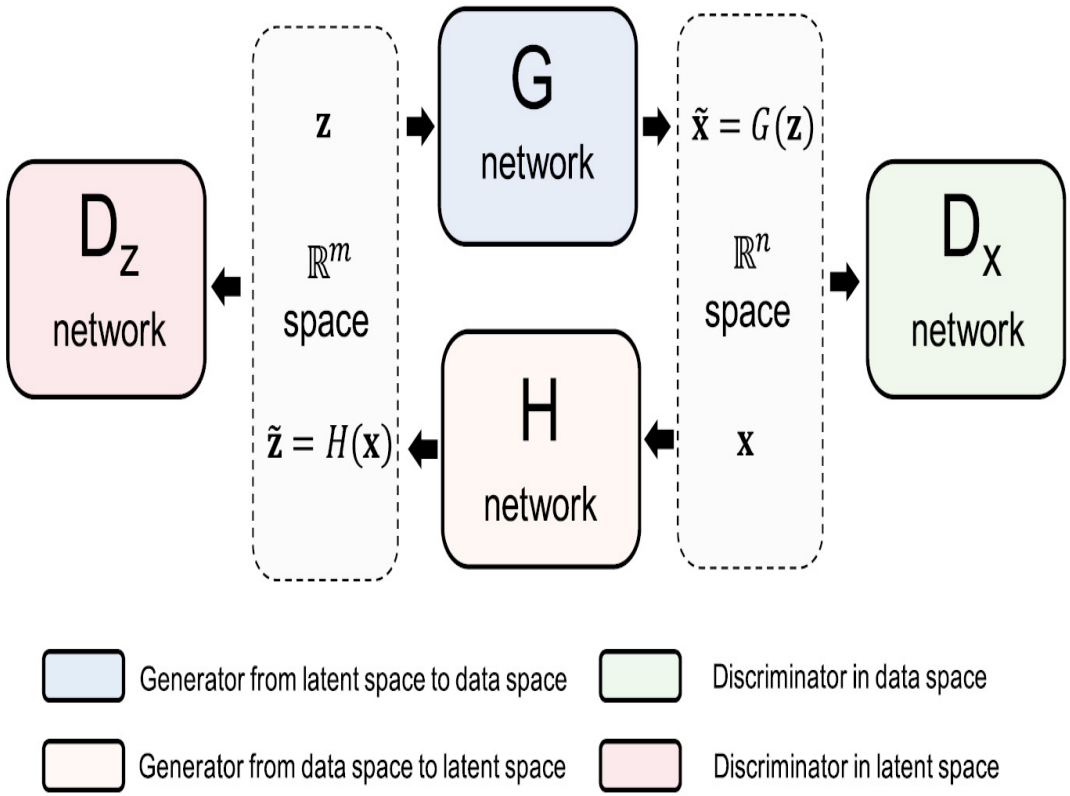


Fig. 1. The overview of Roundtrip framework. In the latent space, latent variable $\mathbf{z}$ follows a standard Gaussian distribution, which is fed to the G network. The H network maps $\mathbf{x}$ from data space to the latent space. The D_x network works as a discriminator for discerning the true data ($\mathbf{x}$) from the generated data ($\tilde{\mathbf{x}}$). The D_z network is another discriminator for distinguishing the generated latent variable ($\tilde{\mathbf{z}}$) from the real latent variable ($\mathbf{z}$).

RoundTrip 模型：模型框架

- 输入：K 个 n 维空间上的简单随机样本 $x_1, \cdots, x_K$;
- 输出：数据服从的密度函数 $p_x(x)$
- 假设：数据落在 R^n 上的低维流形的附近，是低维流形上点的一个随机扰动。低维流形的本征维数为 m, 设流形表示为 $\tilde{x} = G(z)$, 样本到隐空间的投影 $\tilde{z} = H(x)$,

$$x = \tilde{x} + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

- 假设隐空间上的随机变量 $Z \sim N(0, I_m)$, $p_{x|z}(x|z) \sim N(G(z), \sigma^2)$, 于是待求的密度为

$$p_x(x) = \int p_{x|z}(x|z) p_z(z)\, dz = \left(\frac{1}{\sqrt{2\pi}}\right)^{m+n} \sigma^{-n} e^{-\frac{w(x,z)}{2}}\, dz$$

其中 $w(x, z) = ||z||_2^2 + \sigma^{-2}||x - G(z)||_2^2$.

Roundtrip 模型：重要性抽样

- 常规的抽样方法低效

$$z_i \sim p_z(z), i = 1, \cdots, N, \quad p_x(x) \approx \frac{1}{N} \sum_{i=1}^{N} p_{x|z}(x|z_i)$$

- 重要性抽样

$$z_i^q \sim q_z(z), i = 1, \cdots, N, \quad p^{IS}(x) \approx \frac{1}{N} \sum_{i=1}^{N} p_{x|z}(x|z_i) w(z_i^q)$$

其中 $w(z) = p_z(z)/q_z(z)$, $q_z(z)$ 是中心在 $\tilde{z} = H(x)$ 的学生氏 t 分布，原因有二 (1). 给定 x, 条件概率 $p_{x|z}(x|z)$ 在 z 处取最大；(2). t 分布重尾可以减少上述估计的方差。

Roundtrip 模型：Laplace 近似

- 将 $G(z)$ 在 $\tilde{z} = H(x)$ 处得到二阶近似 $v(x, z)$, 得到近似解

$$p^{LP}(x) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \sigma^{-n}\sqrt{det(\Sigma)}e^{-\frac{c(x)}{2}}$$

这里

$$\Sigma = \left(I_m + \sigma^{-2}J_{\tilde{z}}^T J_{\tilde{z}}\right)^{-1}$$

$$c(x) = ||H(x)||_2^2 + \sigma^{-2}||x - G(H(x))||_2^2 - \mu^T\Sigma^{-1}\mu$$

其中 (μ, Σ) 是 Laplace 近似给出的正态分布参数,$J_{\tilde{z}}$ 是 $G(z)$ 在 $\tilde{z}$ 处的 Jacobi 矩阵;

Roundtrip 模型：对抗学习

- 网络 G 学习

$$\begin{cases} L_{GAN}(G) = E_{z\sim p(z)}(D_x(G(z)) - 1)^2 \\ L_{GAN}(D_x) = E_{x\sim p_x}(D(x) - 1)^2 + E_{z\sim p(z)} D_x^2(G(z)) \end{cases}$$

- 网络 H 学习

$$\begin{cases} L_{GAN}(H) = E_{x\sim p(x)}(D_z(H(x)) - 1)^2 \\ L_{GAN}(D_z) = E_{z\sim p_z}(D(x) - 1)^2 + E_{x\sim p(x)} D_z^2(G(z)) \end{cases}$$

Roundtrip 模型：对抗学习

- Roundtrip 损失函数

$$L_{RT} = \alpha||x - G(H(x))||_2^2 + \beta||z - H(G(z))||_2^2$$

- 总体损失的联系训练

$$L(G, H) = L_{GAN}(G) + L_{GAN}(H) + L_{RT}(G, H)$$
$$(D_x, D_z) = L_{GAN}(D_x) + L_{GAN}(G_Z)$$

- 交替更新两个生成模型 $G, H,$ 和两个判别模型 D_z, D_x 中的参数：

$$G^*, H^*, D_x^*, D_z^* = \begin{cases} \underset{G\ H}{\text{Argmin}}\ L(G, H) \\ \underset{D_x\ D_z}{\text{Argmin}}\ L(D_z, D_z) \end{cases}$$

Roundtrip 密度估计效果

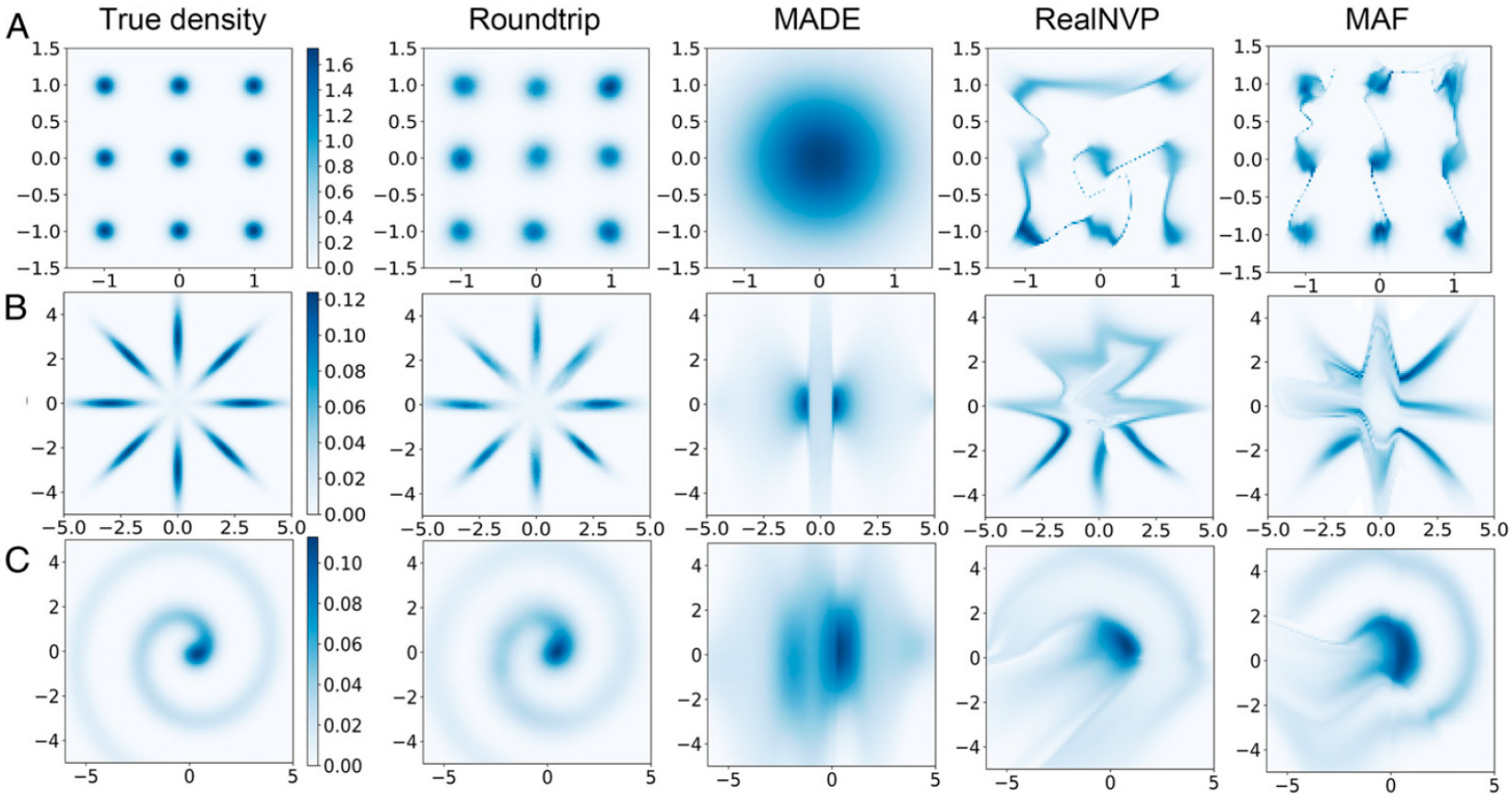


Fig. 2. True density and estimated density by different neural density estimators (Roundtrip, MADE, RealNVP, and MAF) with three simulation datasets. Density plots were shown on a 100 × 100 grid 2D bounded region. True densities were shown in the first column. Each row gives results of a dataset: (*A*) independent Gaussian mixture; (*B*) eight-octagon Gaussian mixture; (*C*) Involute.

Roundtrip 似然函数比较

Table 1. Performance of different methods on five UCI datasets

	AReM	CASP	HEPMASS	BANK	YPMSD
KDE	6.26 ± 0.07	20.47 ± 0.10	-25.46 ± 0.03	15.84 ± 0.12	247.03 ± 0.61
MADE	6.00 ± 0.11	21.82 ± 0.23	-15.15 ± 0.02	14.97 ± 0.53	273.20 ± 0.35
RealNVP	9.52 ± 0.18	26.81 ± 0.15	-18.71 ± 0.02	26.33 ± 0.22	287.74 ± 0.34
MAF	9.49 ± 0.17	27.61 ± 0.13	-17.39 ± 0.02	20.09 ± 0.20	290.76 ± 0.33
Roundtrip	**11.74** ± 0.04	**28.38** ± 0.08	**−4.18** ± 0.02	**35.16** ± 0.14	**297.98** ± 0.52

The average log likelihood and 2 SDs are shown. The best performance among five methods is shown in bold.

Roundtrip 图片生成

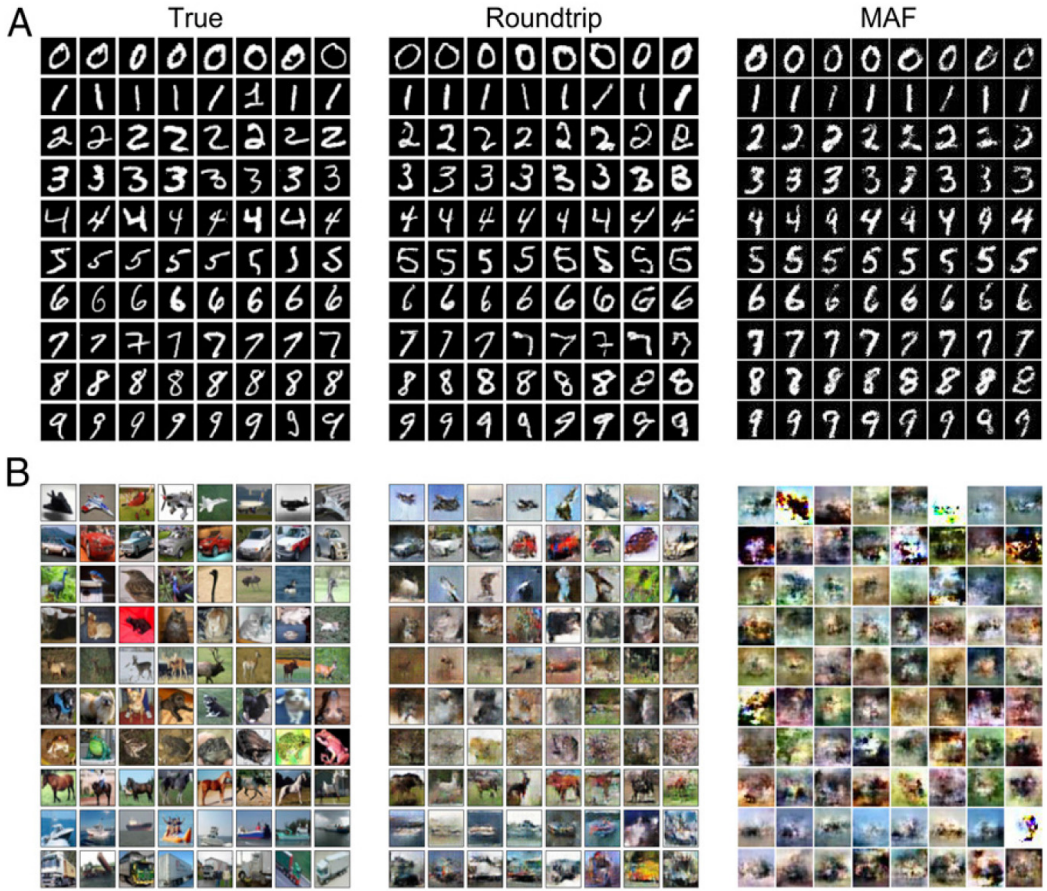


Fig. 3. True images and generated images from Roundtrip and MAF models trained on MNIST and CIFAR-10 databases. Each row denotes the true or generated images of a specific class. (*A*) True and generated images of MNIST. (*B*) True and generated images of CIFAR-10. Images generated by Roundtrip and MAF are sorted by decreased likelihood for each class.

作业

- 考虑正态分布 $f(x) \sim N(\mu, \sigma^2)$ 和核函数 $K(x) \sim N(0,1)$, 证明核估计

$$f_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K(\frac{x - X_i}{h_n})$$

满足如下性质

1. $Ef_n(x) \sim N(\mu, \sigma^2 + h_n^2)$
2. $Var[f_n(x)] \approx \dfrac{f(x)}{2nh_n\sqrt{\pi}}$
3. $f(x) - Ef_n(x) \approx \dfrac{1}{2}\left(\dfrac{h}{\sigma}\right)^2 \left[1 - (\dfrac{x-\mu}{\sigma})^2\right] f(x), \quad h_n \to 0.$

作业

- 考虑 $[0, a]$ 上的均匀分布 $f(x) \sim U(0, a)$, 并设核函数为指数分布

$$K(x) = \begin{cases} e^{-x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

若 $X_1, \cdots, X_n$ 是均匀分布的简单随机样本。求均匀分布 U(0,a) 的核估计 $f_n(x)$ 的期望。

作业

- 对于直方图密度估计，有如下定理：
- **定理：** 假设密度函数 $f(x)$ 满足：导函数 $f'(x)$ 绝对连续，且 $\int [f'(u)]^2 du < +\infty$, 那么

$$MISE(f_n(x)) = \frac{h^2}{12} \int [f'(u)]^2 du + \frac{1}{nh} + o(h^2) + o(\frac{1}{nh})$$

于是，

$$h_{opt} = \frac{1}{n^{1/3}} \left\{ \frac{6}{\int [f'(u)]^2 du} \right\}^{1/3}$$

在此带宽下，

$$MISE(f_n(x)) \approx \frac{C}{n^{2/3}}, \qquad C = (\frac{3}{4})^{2/3} \left\{ \int [f'(u)]^2 du \right\}^{1/3}$$