《机器学习基础》

无监督学习

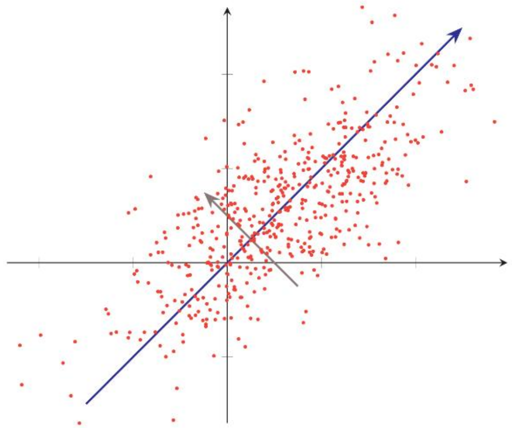
内容

- 无监督学习
  - 无监督特征学习
    - 主成分分析
    - 稀疏编码
    - 自编码器
    - 稀疏自编码器
    - 降噪自编码器

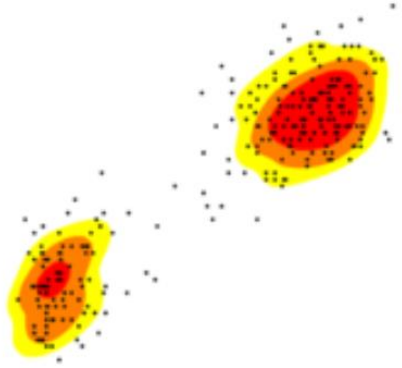
无监督学习（Unsupervised Learning）

- 监督学习
  - 建立映射关系 $f: x \to y$
- 无监督学习
  - 指从无标签的数据中学习出一些有用的模式。
  - 聚类：建立映射关系 $f: x \to y$
    - 不借助于任何人工给出标签或者反馈等指导信息
  - 特征学习

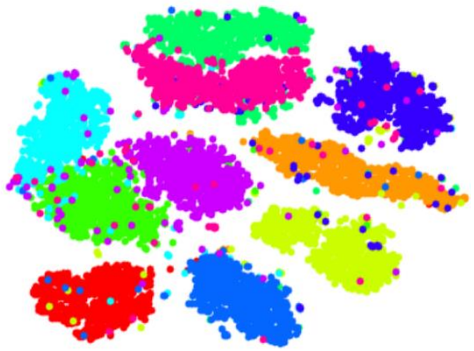
典型的无监督学习问题



无监督特征学习



密度估计



聚类

为什么要无监督学习?

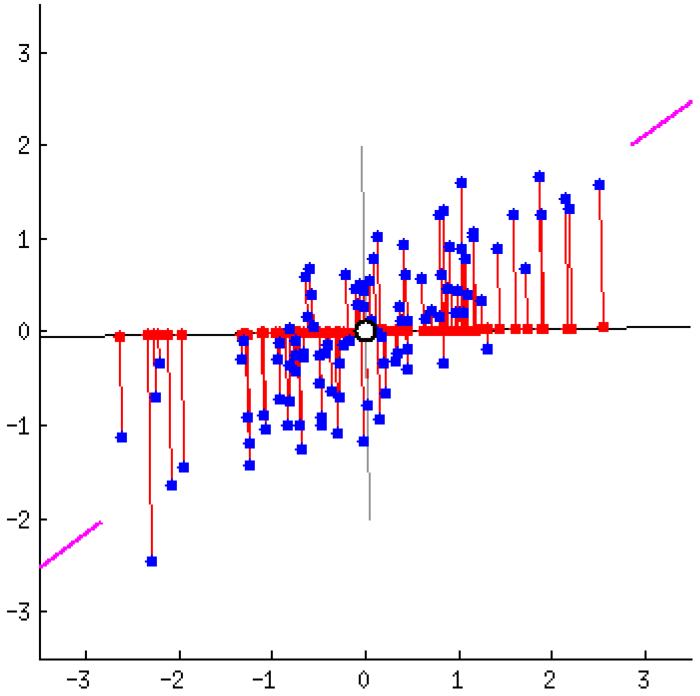
大脑有大约10^{14}个突触，我们只能活大约10^9秒。所以我们有比数据更多的参数。这启发了我们必须进行大量无监督学习的想法，因为感知输入（包括本体感受）是我们可以获得每秒10^5维约束的唯一途径。

-- Geoffrey Hinton, 2014 AMA on Reddit

无监督特征学习

PCA优化目标

- 一种最常用的数据降维方法，使得在转换后的空间中数据的方差最大。



主成份分析（Principal Component Analysis，PCA）

- 一种最常用的数据降维方法，使得在转换后的空间中数据的方差最大。
  - 样本点 $\boldsymbol{x}^{(n)}$ 投影之后的表示为

$$z^{(n)} = \boldsymbol{w}^{\top}\boldsymbol{x}^{(n)}$$

  - 所有样本投影后的方差为

$$\begin{aligned}\sigma(\boldsymbol{X};\boldsymbol{w}) &= \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{w}^{\top}\boldsymbol{x}^{(n)} - \boldsymbol{w}^{\top}\bar{\boldsymbol{x}})^2 \\ &= \frac{1}{N}(\boldsymbol{w}^{\top}\boldsymbol{X} - \boldsymbol{w}^{\top}\bar{\boldsymbol{X}})(\boldsymbol{w}^{\top}\boldsymbol{X} - \boldsymbol{w}^{\top}\bar{\boldsymbol{X}})^{\top} \\ &= \boldsymbol{w}^{\top}\boldsymbol{\Sigma}\boldsymbol{w},\end{aligned}$$

  - 目标函数

$$\max_{\boldsymbol{w}} \boldsymbol{w}^{\top}\boldsymbol{\Sigma}\boldsymbol{w} + \lambda(1 - \boldsymbol{w}^{\top}\boldsymbol{w})$$

- 对目标函数求导并令导数等于0，可得

$$\boldsymbol{\Sigma}\boldsymbol{w} = \lambda\boldsymbol{w}$$

主成分分析

- 定义一个n×m的矩阵,X^T为去平均值（以平均值为中心移动至原点）的数据，其行为数据样本，列为数据类别。则X的奇异值分解为
  - $X = W\Sigma V^T$
- 据此,

$$
\begin{aligned}
\boldsymbol{Y}^\top &= \boldsymbol{X}^\top \boldsymbol{W} \\
&= \boldsymbol{V}\boldsymbol{\Sigma}^\top \boldsymbol{W}^\top \boldsymbol{W} \\
&= \boldsymbol{V}\boldsymbol{\Sigma}^\top
\end{aligned}
$$

- 我们可以利用$\mathbf{W}_L$把 X 映射到一个只应用前面L个向量的低维空间中去:

$$\mathbf{Y} = \mathbf{W_L}^\top \mathbf{X} = \mathbf{\Sigma_L} \mathbf{V}^\top$$

由主成分重建原始数据

$$
\begin{aligned}
\boldsymbol{Y}^\top &= \boldsymbol{X}^\top \boldsymbol{W} \\
&= \boldsymbol{V}\boldsymbol{\Sigma}^\top \boldsymbol{W}^\top \boldsymbol{W} \\
&= \boldsymbol{V}\boldsymbol{\Sigma}^\top
\end{aligned}
$$

PCA优化目标

- PCA推导有两种主要思路：
  1. 最大化数据投影后的的方差（让数据更分散）
  2. 最小化投影造成的损失

- 采用第一种思路完成推导过程，下图中旋转的是新坐标轴，每个数据点在改坐标轴上垂直投影，最佳的坐标轴为数据投影后的数据之间距离最大。

选择降维后的维度K(主成分的个数)

- average squared projection error

$$\frac{1}{m}\sum_{i=1}^{m}\left\|x^{(i)}-x_{approx}^{(i)}\right\|^2$$

- total variation in the data

$$\frac{1}{m}\sum_{i=1}^{m}\left\|x^{(i)}\right\|^2$$

选择降维后的维度K(主成分的个数)

- 选择不同的K值，然后用下面的式子不断计算，选取能够满足下列式子条件的最小K值即可。

$$\frac{\frac{1}{m}\sum_{i=1}^{m}\left\|x^{(i)}-x_{approx}^{(i)}\right\|^2}{\frac{1}{m}\sum_{i=1}^{m}\left\|x^{(i)}\right\|^2}\leq t$$

- 还可以用SVD分解时产生的S矩阵来表示

$$1-\frac{\sum_{i=1}^{k}S_{ii}}{\sum_{i=1}^{n}S_{ii}}\leq t$$

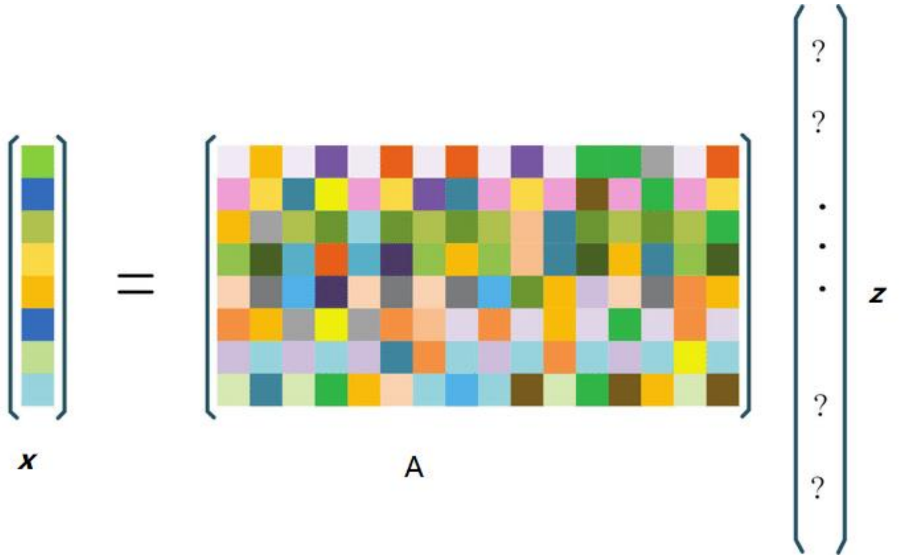
（线性）编码

- 给定一组基向量$A=[\boldsymbol{a}_1,\cdots,\boldsymbol{a}_M]$，将输入样本$\boldsymbol{x}$表示为这些基向量的线性组合

$$
\begin{aligned}
\boldsymbol{x} &= \sum_{m=1}^{M} z_m \boldsymbol{a}_m \\
&= \boldsymbol{A}\boldsymbol{z},
\end{aligned}
$$

字典（dictionary）

编码（encoding）



完备性

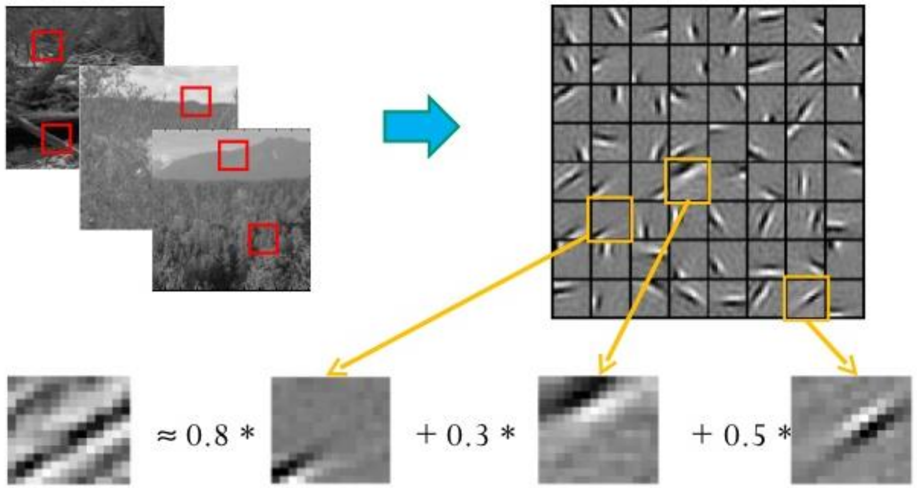
数学小知识 | 完备性

如果 M 个基向量刚好可以支撑 M 维的欧氏空间，则这 M 个基向量是完备的. 如果 M 个基向量可以支撑 D 维的欧氏空间，并且 $M > D$，则这 M 个基向量是过完备的（overcomplete）、冗余的.

"过完备"基向量是指基向量个数远远大于其支撑空间维度. 因此这些基向量一般不具备独立、正交等性质.

- 稀疏编码
  - 找到一组"过完备"的基向量（即 $M > D$）来进行编码。

Sparse coding illustration



$[a_1, ..., a_{64}] = [0, 0, ..., 0, \mathbf{0.8}, 0, ..., 0, \mathbf{0.3}, 0, ..., 0, \mathbf{0.5}, 0]$
(feature representation)

Slide credit: Andrew Ng

Compact & easily interpretable

稀疏编码（Sparse Coding）

- 给定一组N 个输入向量$\boldsymbol{x}^{(1)},\ldots,\boldsymbol{x}^{(N)}$，其稀疏编码的目标函数定义为

$$\mathcal{L}(\boldsymbol{A},\boldsymbol{Z})=\sum_{n=1}^{N}\left(\left\|\boldsymbol{x}^{(n)}-A\boldsymbol{z}^{(n)}\right\|^{2}+\eta\rho(\boldsymbol{z}^{(n)})\right)$$

- $\rho(\cdot)$是一个稀疏性衡量函数，η是一个超参数，用来控制稀疏性的强度。

$$\boxed{\rho(\mathbf{z})=\sum_{i=1}^{p}\mathbf{I}(|z_i|>0))}$$

$$\boxed{\rho(\mathbf{z})=\sum_{i=1}^{p}-\exp(-z_i^2)}$$

$$\boxed{\rho(\mathbf{z})=\sum_{i=1}^{p}|z_i|}$$

$$\boxed{\rho(\mathbf{z})=\sum_{i=1}^{p}\log(1+z_i^2)}$$

训练过程

- 稀疏编码的训练过程一般用交替优化的方法进行。

1) 固定基向量$\boldsymbol{A}$,对每个输入$\boldsymbol{x}^{(n)}$,计算其对应的最优编码

$$\min_{\boldsymbol{z}^{(n)}} \left\| \boldsymbol{x}^{(n)} - \boldsymbol{A}\boldsymbol{z}^{(n)} \right\|^2 + \eta\rho(\boldsymbol{z}^{(n)}), \ \forall n \in [1, N].$$

2) 固定上一步得到的编码$\{\boldsymbol{z}^{(n)}\}_{n=1}^{N}$,计算其最优的基向量

$$\min_{\boldsymbol{A}} \sum_{n=1}^{N} \left(\left\| \boldsymbol{x}^{(n)} - \boldsymbol{A}\boldsymbol{z}^{(n)} \right\|^2 \right) + \lambda \frac{1}{2} \|\boldsymbol{A}\|^2,$$

稀疏编码的优点

- 计算量
  - 稀疏性带来的最大好处就是可以极大地降低计算量。
- 可解释性
  - 因为稀疏编码只有少数的非零元素，相当于将一个输入样本表示为少数几个相关的特征。这样我们可以更好地描述其特征，并易于理解。
- 特征选择
  - 稀疏性带来的另外一个好处是可以实现特征的自动选择，只选择和输入样本相关的最少特征，从而可以更好地表示输入样本，降低噪声并减轻过拟合。

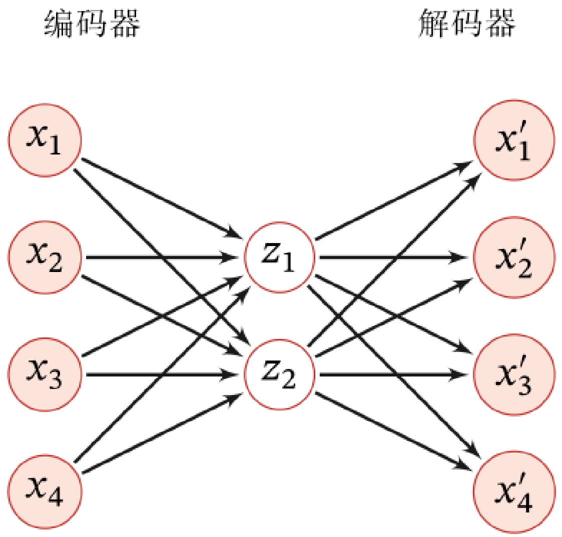
自编码器（Auto-Encoder）

- 编码器（Encoder）$f : \mathbb{R}^D \to \mathbb{R}^M$
- 解码器（Decoder）$g : \mathbb{R}^M \to \mathbb{R}^D$

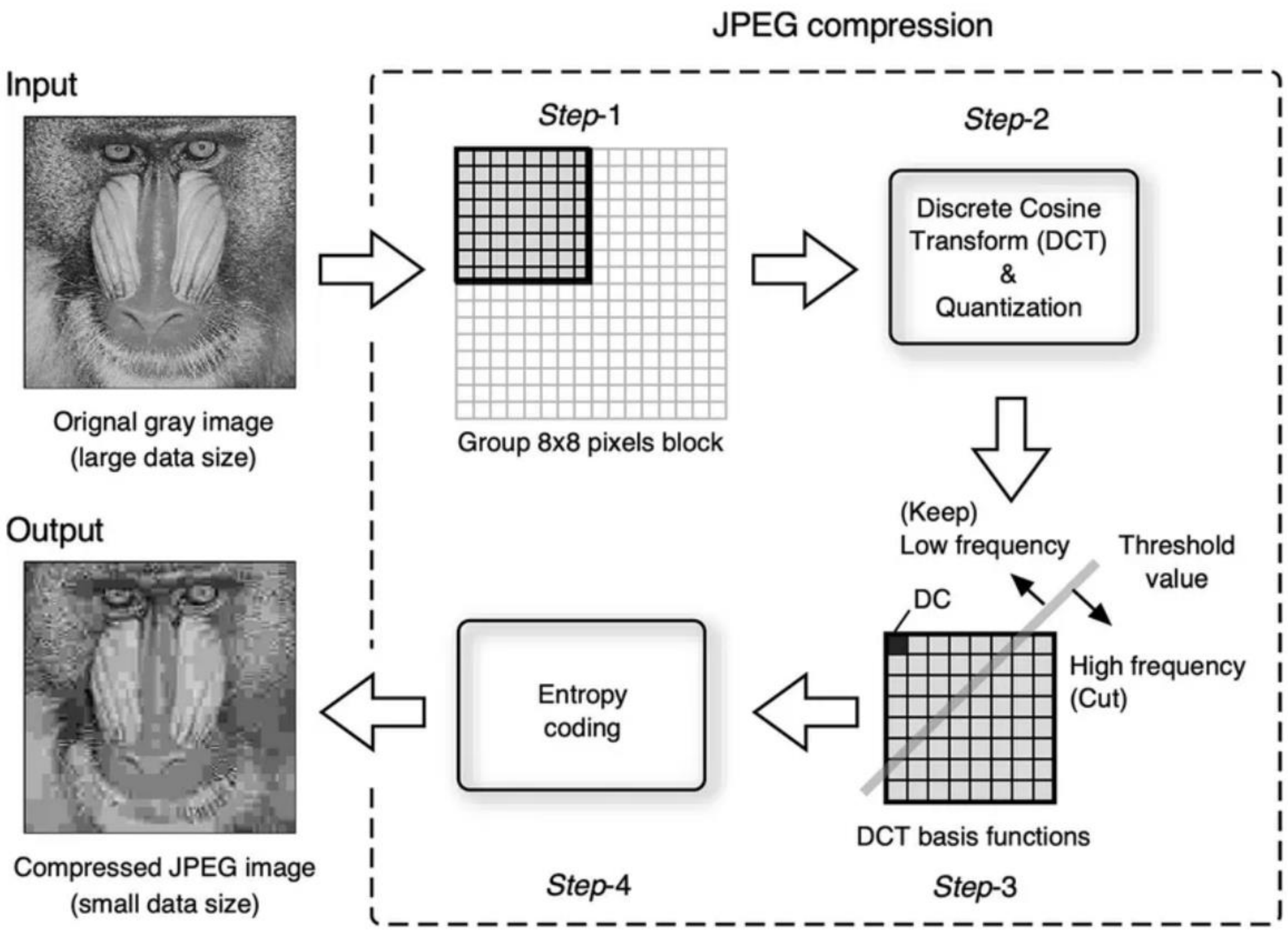
- 目标函数：重构错误

$$
\begin{aligned}
\mathcal{L} &= \sum_{n=1}^{N} \|\boldsymbol{x}^{(n)} - g(f(\boldsymbol{x}^{(n)}))\|^2 \\
&= \sum_{n=1}^{N} \|\boldsymbol{x}^{(n)} - f \circ g(\boldsymbol{x}^{(n)})\|^2.
\end{aligned}
$$

- 两层网络结构的自编码器



自编码器（Auto-Encoder）



JPEG pipeline (Figure is taken from[3])

稀疏自编码器

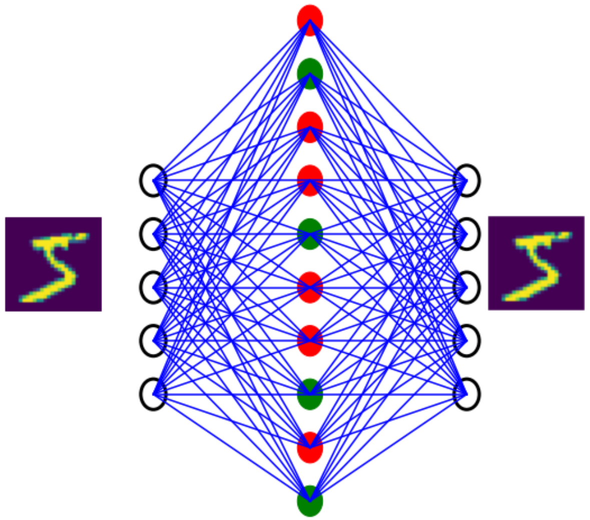
- 通过给自编码器中隐藏层单元z加上稀疏性限制，自编码器可以学习到数据中一些有用的结构。

- 目标函数

$$\mathcal{L} = \sum_{n=1}^{N} ||\boldsymbol{x}^{(n)} - \boldsymbol{x}'^{(n)})||^2 + \eta\rho(\boldsymbol{Z}) + \lambda||\boldsymbol{W}||^2$$

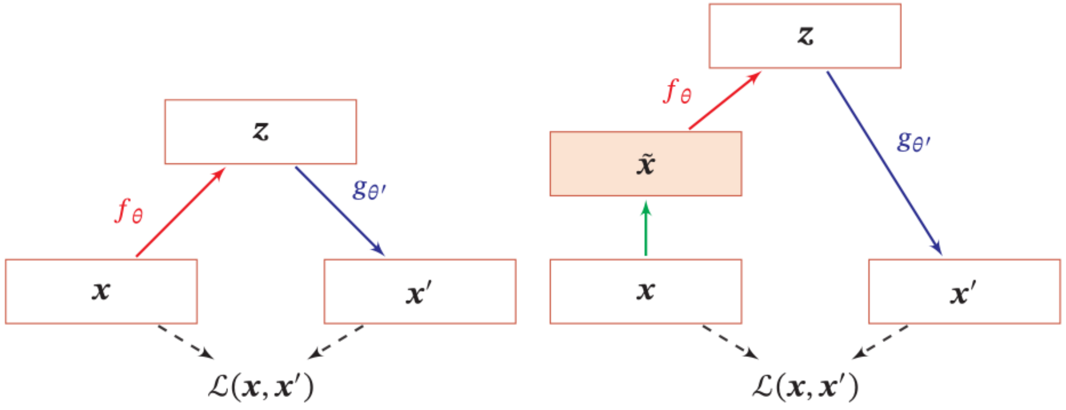
- $\boldsymbol{W}$表示自编码器中的参数

- 和稀疏编码一样，稀疏自编码器的优点是有很高的可解释性，并同时进行了隐式的特征选择.



降噪自编码器

- 通过引入噪声来增加编码鲁棒性的自编码器
  - 对于一个向量$\boldsymbol{x}$，我们首先根据一个比例μ随机将$\boldsymbol{x}$的一些维度的值设置为0，得到一个被损坏的向量$\tilde{\boldsymbol{x}}$。
  - 然后将被损坏的向量$\tilde{\boldsymbol{x}}$输入给自编码器得到编码$\boldsymbol{z}$，并重构出原始的无损输入$\boldsymbol{x}$ 。



(a) 自编码器　　(b) 降噪自编码器