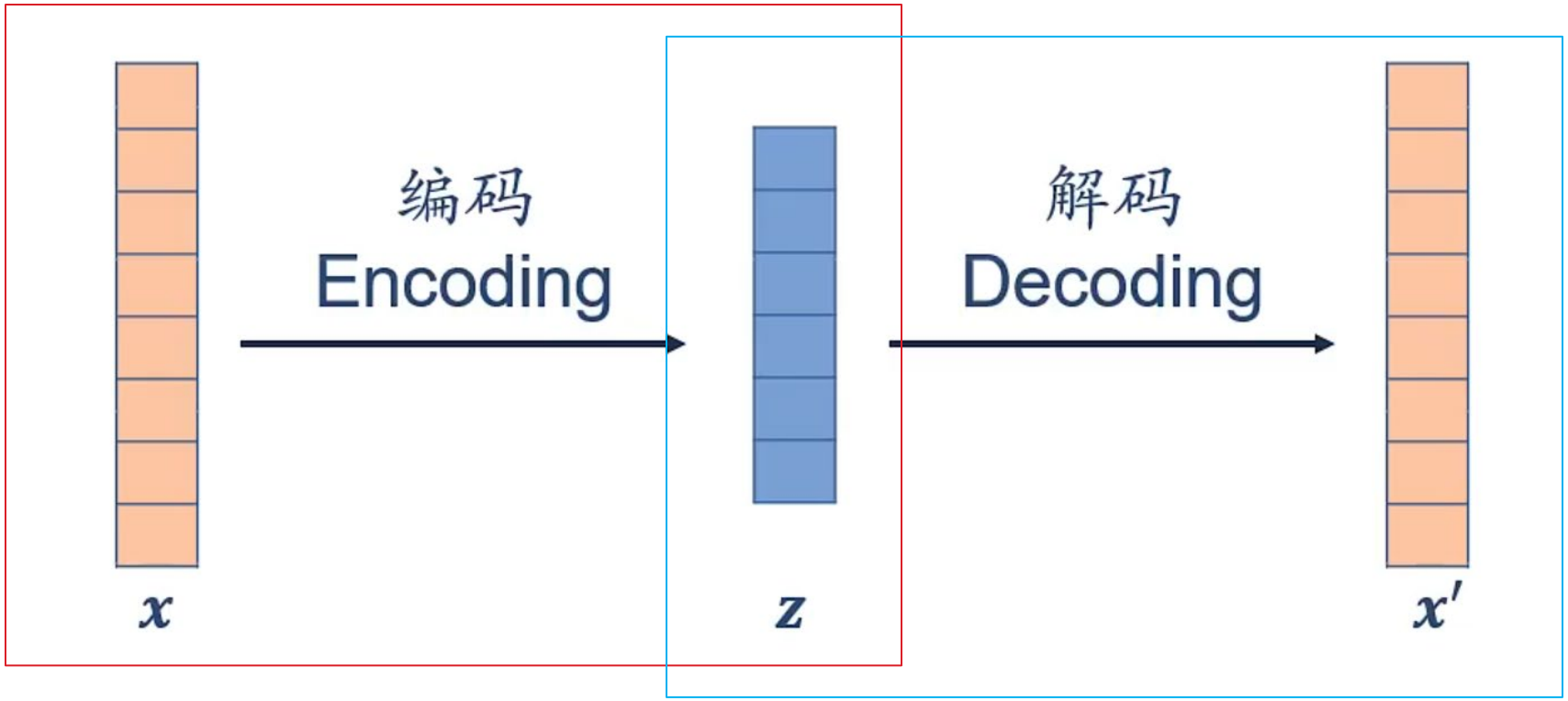
无监督特征学习

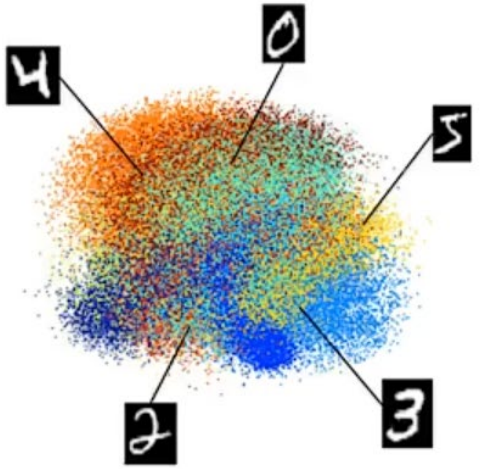
特征学习/编码



从无标签数据中学习有用的特征（表示）

特征学习的应用

- 目的：特征提取、去噪、降维、数据可视化
  - 降维/可视化



降维
数据集可视化

  - 稀疏编码

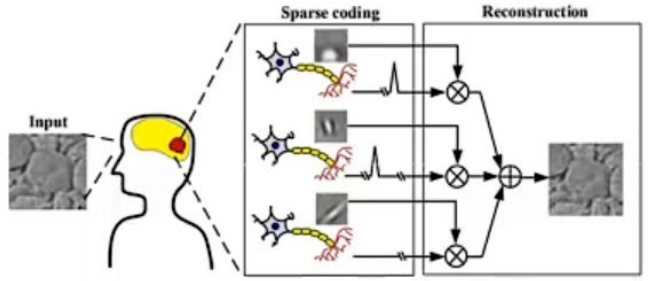


Fig. 1. Sparse coding mimicking sparse neural activities in primary visual cortex. The input is reconstructed by weighted sum of receptive fields of model neurons.

稀疏编码

http://www.jzlab.org/Knag15_SAILnet_Chip.pdf

无监督特征学习
主成分分析

主成分分析（Principal Component Analysis，PCA）

- 数据原始特征的问题：
  - 高维→维度灾难、过拟合
  - 冗余性→学习效果差
  - 解决方法：降维
- 一种最常用的数据降维方法
  - 线性投影

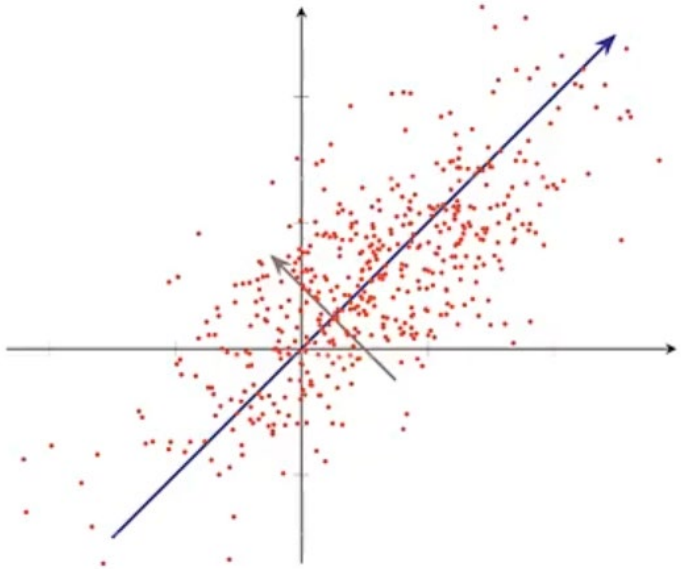
$$\mathbf{z} = \mathbf{W}^{\mathrm{T}}\mathbf{x}$$

  - 并满足

$$\mathbf{W}^{\mathrm{T}}\mathbf{W} = I$$

（保证尽可能小的冗余性）

- 优化准则
  - 最大投影方差：使得转换后的空间中数据方差最大，尽可能多地保留原始数据信息
  - 最小重构误差



主成分分析（Principal Component Analysis，PCA）

- 目的：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

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

  - 所有样本投影后的方差为

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

  - 目标函数

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

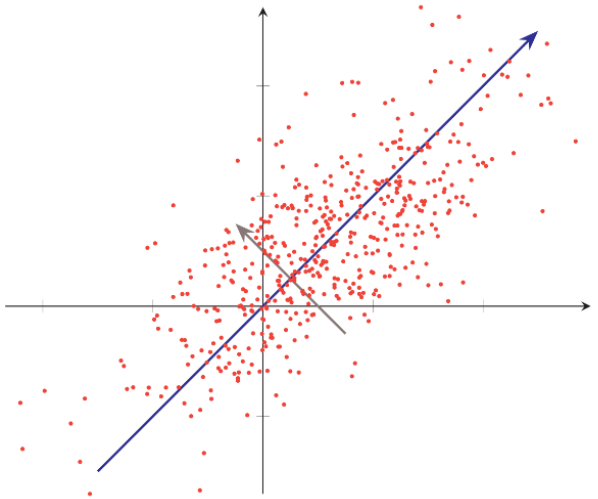
约束：内积=1

（拉格朗日乘子）

- 对目标函数求导并令导数等于0，可得

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

（λ：协方差矩阵的特征值）



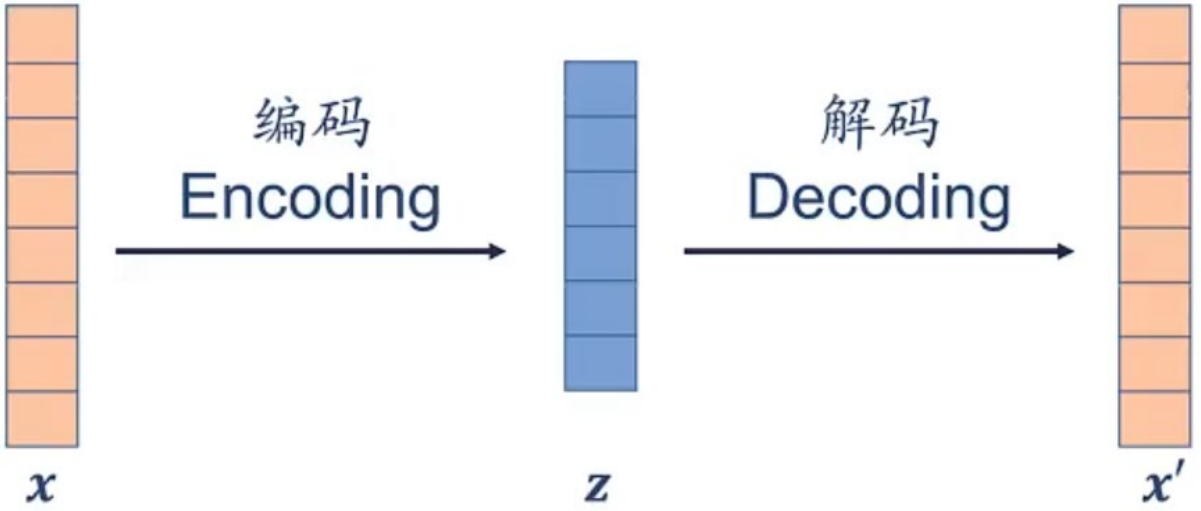
无监督特征学习
编码与稀疏编码

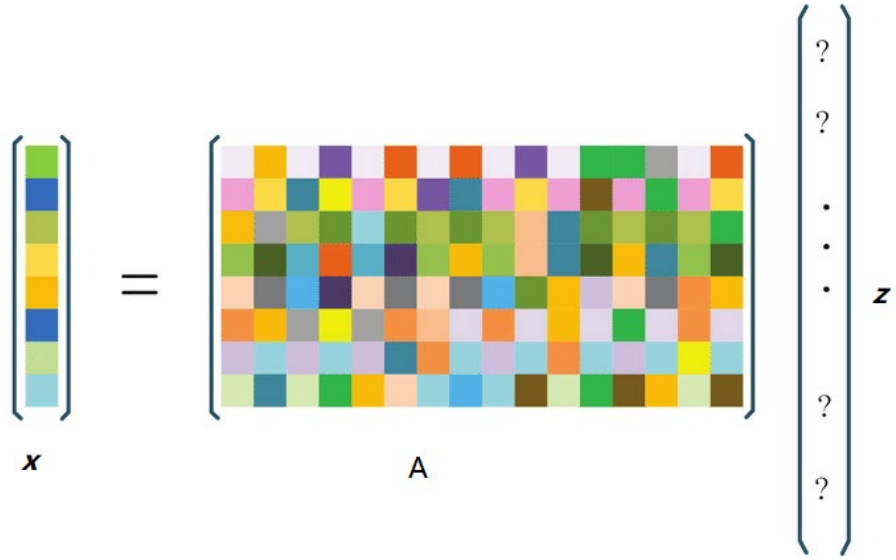
（线性）编码

- 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

$$\boldsymbol{x} = \sum_{m=1}^{M} z_m \boldsymbol{a}_m$$

$$= \boldsymbol{A}\boldsymbol{z},$$





字典（dictionary）

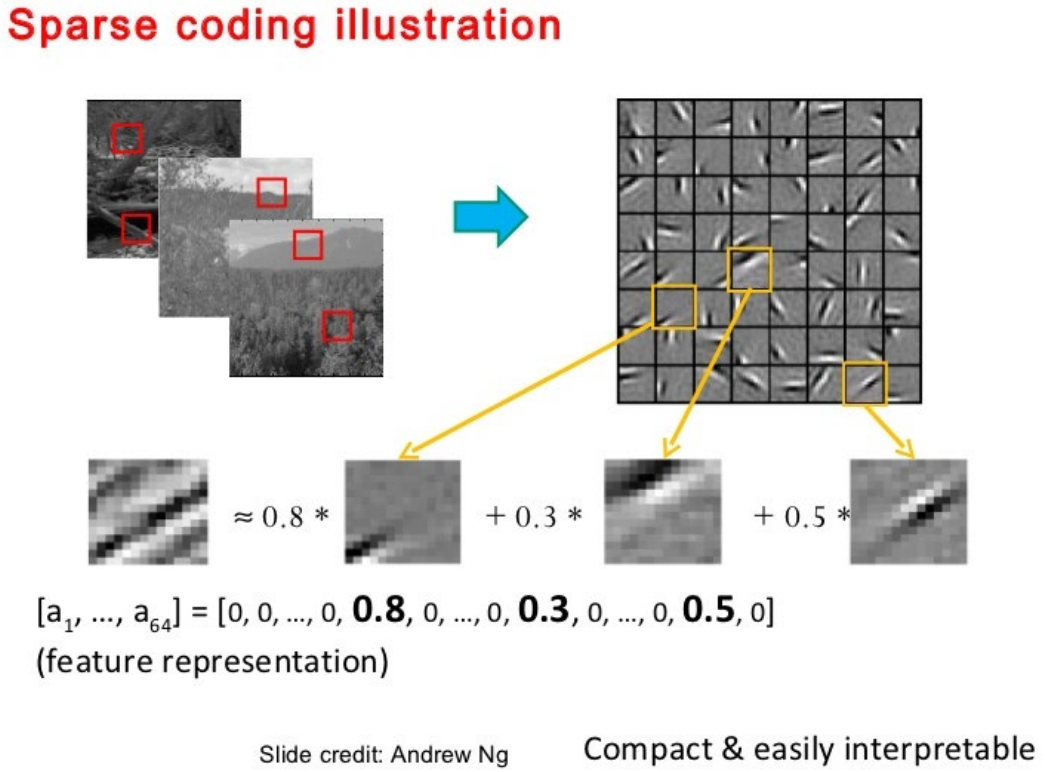
稀疏编码（Sparse Coding）

过完备

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。

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

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

- $\rho(\cdot)$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。

（不是连续可导）

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

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

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

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

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

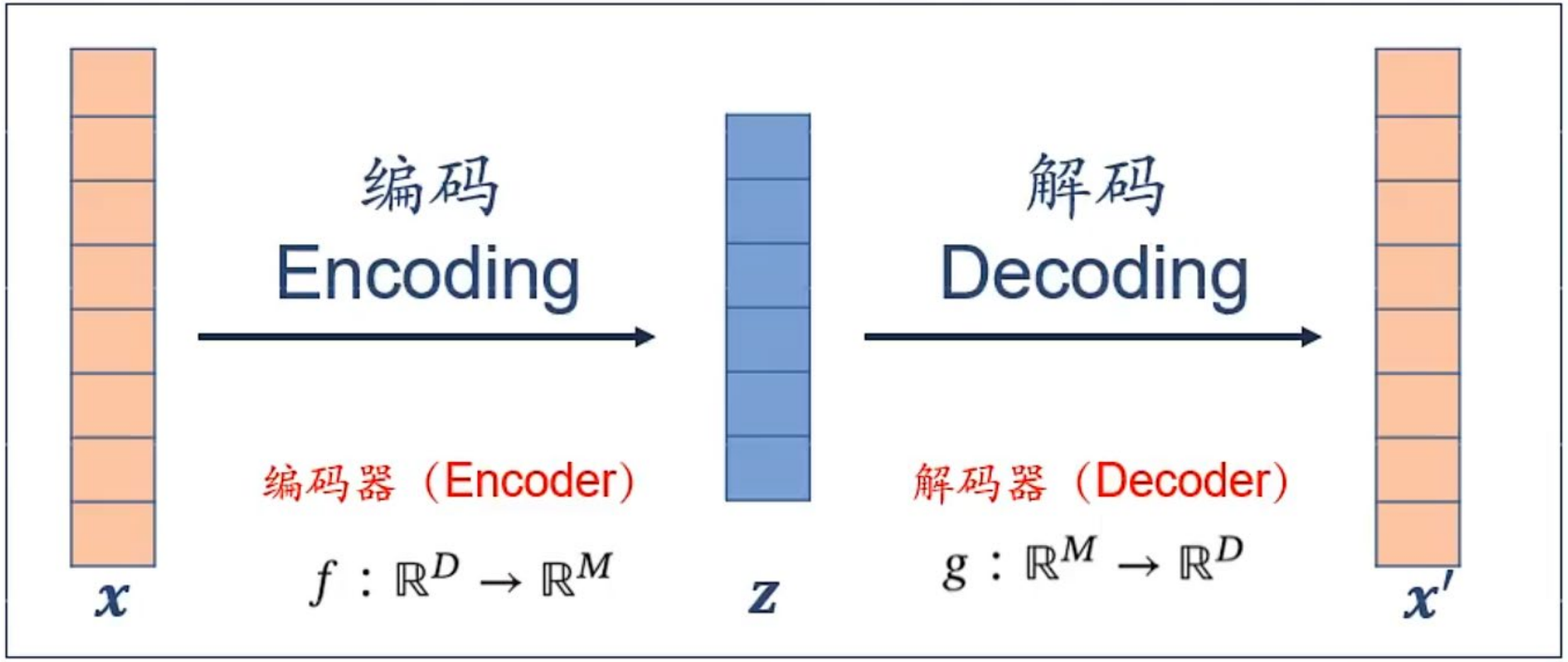
$$\min_{\boldsymbol{A}} \sum_{n=1}^{N} \left(\left\|\boldsymbol{x}^{(n)} - \boldsymbol{A}\boldsymbol{z}^{(n)}\right\|^2\right) + \lambda\frac{1}{2}\|\boldsymbol{A}\|^2,$$

稀疏编码的优点

- 降低后续计算量
  - 稀疏性带来的最大好处就是可以极大地降低计算量。
- 可解释性强
  - 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。
- 实现隐式的特征选择
  - 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。

无监督特征学习
自编码器

自编码器（Auto-Encoder）

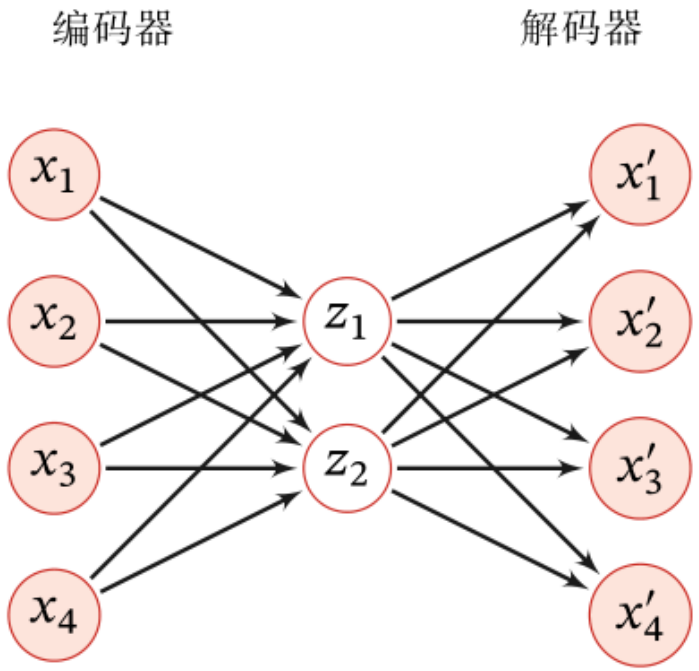


目标函数：重构错误

$$\mathcal{L} = \sum_{n=1}^{N} \|\boldsymbol{x}^{(n)} - g(f(\boldsymbol{x}^{(n)}))\|^2$$

$$= \sum_{n=1}^{N} \|\boldsymbol{x}^{(n)} - f \circ g(\boldsymbol{x}^{(n)})\|^2.$$

自编码器（Auto-Encoder）



目标函数：重构错误

$$\mathcal{L} = \sum_{n=1}^{N} \|\boldsymbol{x}^{(n)} - g(f(\boldsymbol{x}^{(n)}))\|^2$$

$$= \sum_{n=1}^{N} \|\boldsymbol{x}^{(n)} - f \circ g(\boldsymbol{x}^{(n)})\|^2.$$

稀疏自编码器

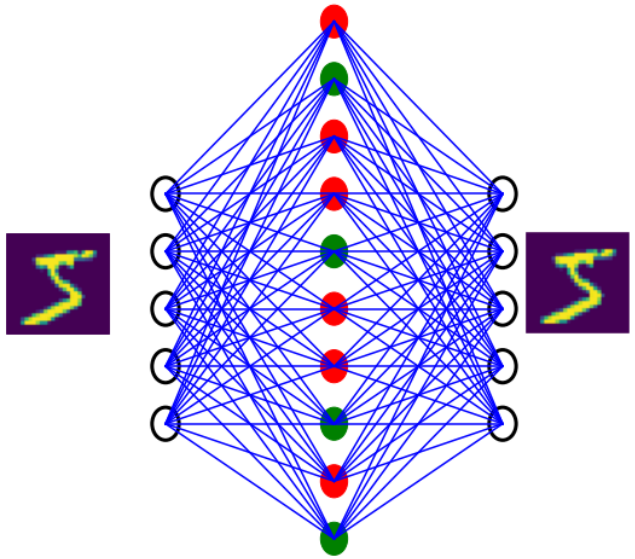
- 通过给自编码器中隐藏层单元z加上稀疏性限制，自编码器可以学习到数据中一些有用的结构。

- 目标函数

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

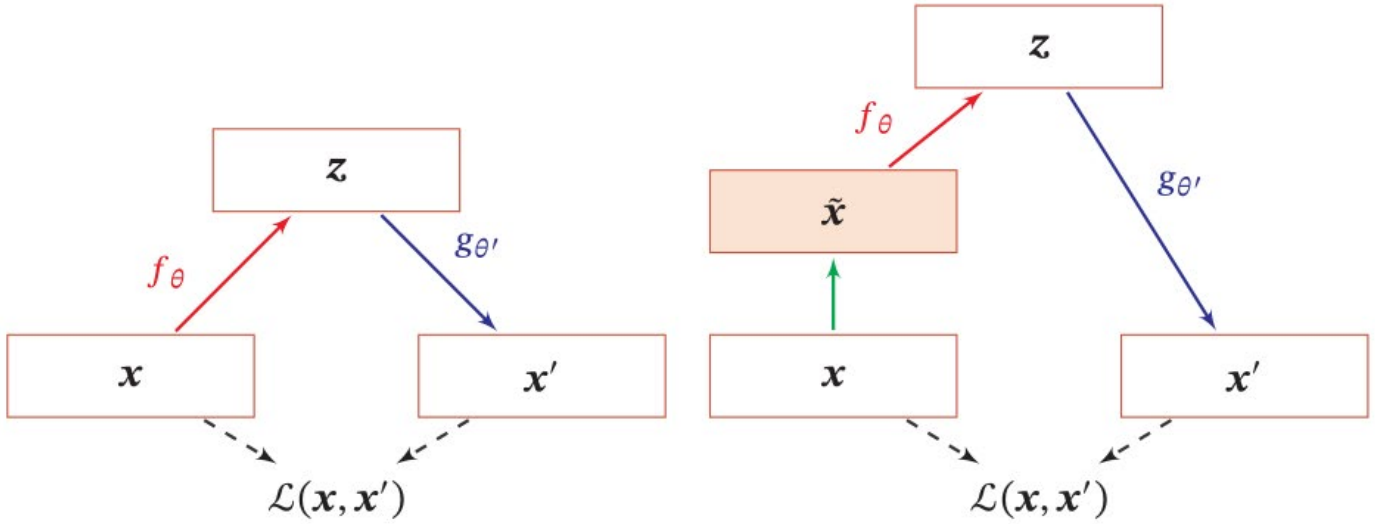
- $\boldsymbol{W}$表示自编码器中的参数

- 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$\boldsymbol{x}$。



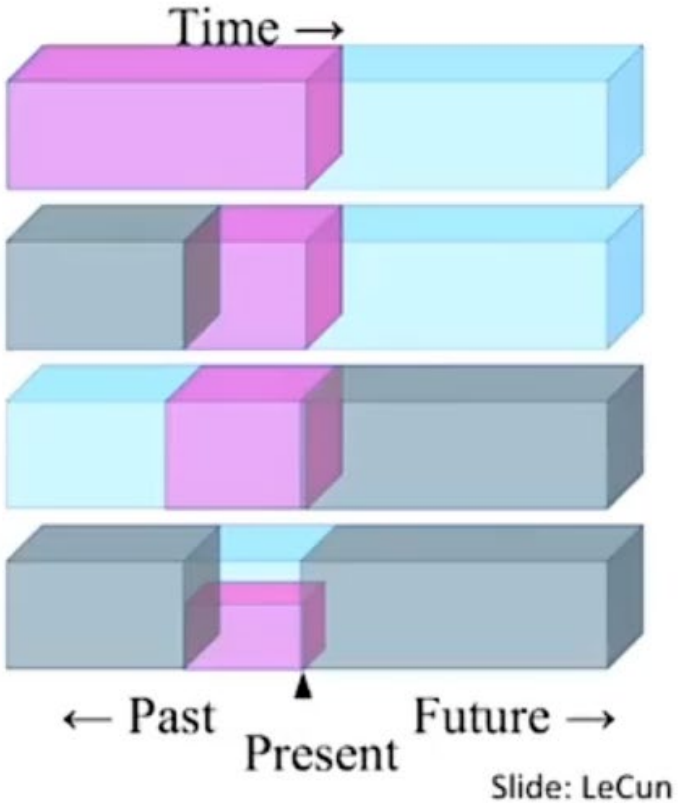
(a) 自编码器　　(b) 降噪自编码器

自监督学习

自监督学习（Self-supervised Learning）

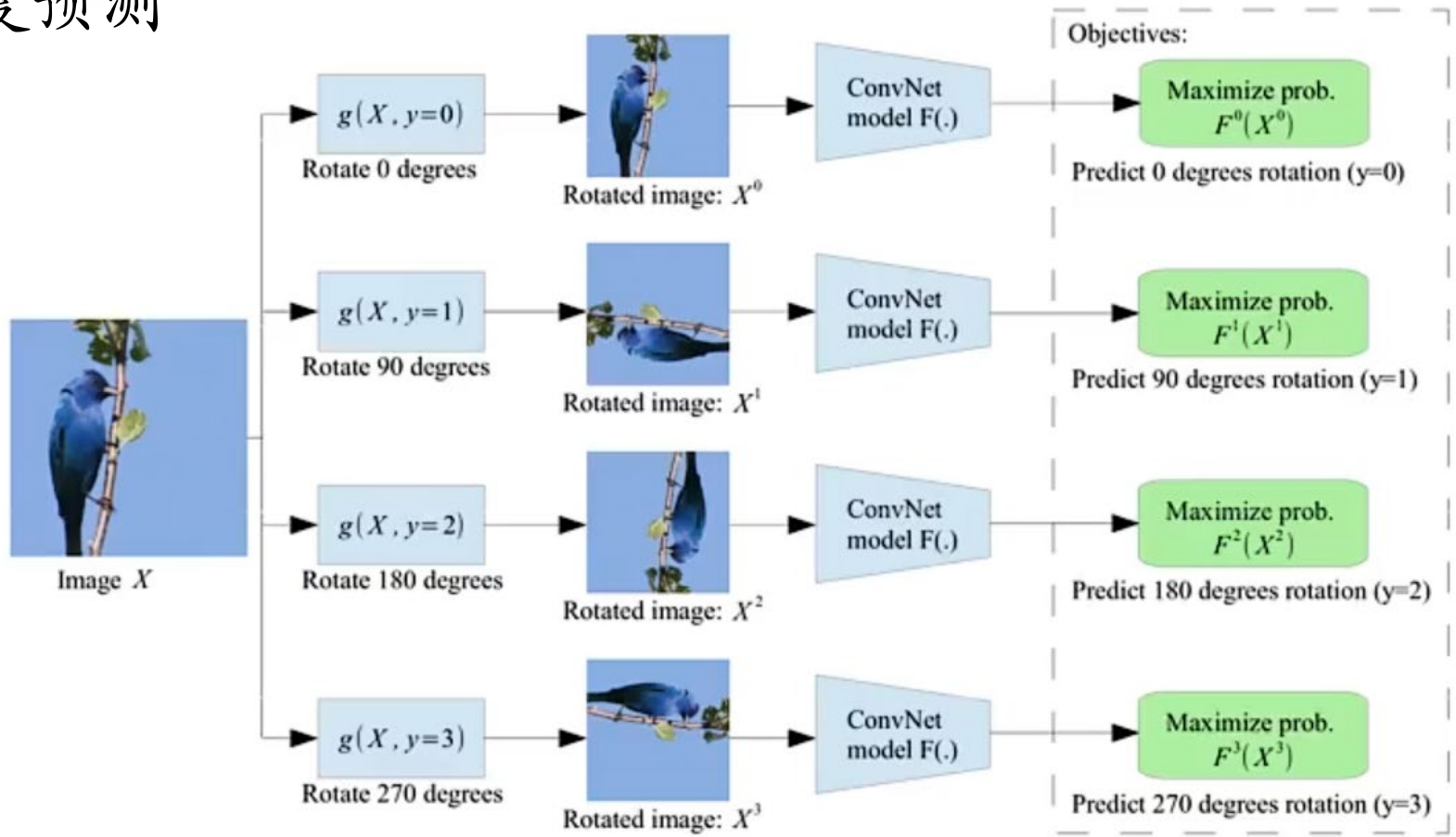
- 不再只是以输入重构作为目标，在无标签样本自身寻找更多样的“目标”

- Predict any part of the input from any other part.
- Predict the future from the past.
- Predict the future from the recent past.
- Predict the past from the present.
- Predict the top from the bottom.
- Predict the occluded from the visible
- Pretend there is a part of the input you don't know and predict that.



自监督学习（Self-supervised Learning）

- 图像任务中的自监督学习
  - 旋转角度预测



自监督学习（Self-supervised Learning）

- 文本任务中的自监督学习
  - 掩码语言模型（Masked Language Model）

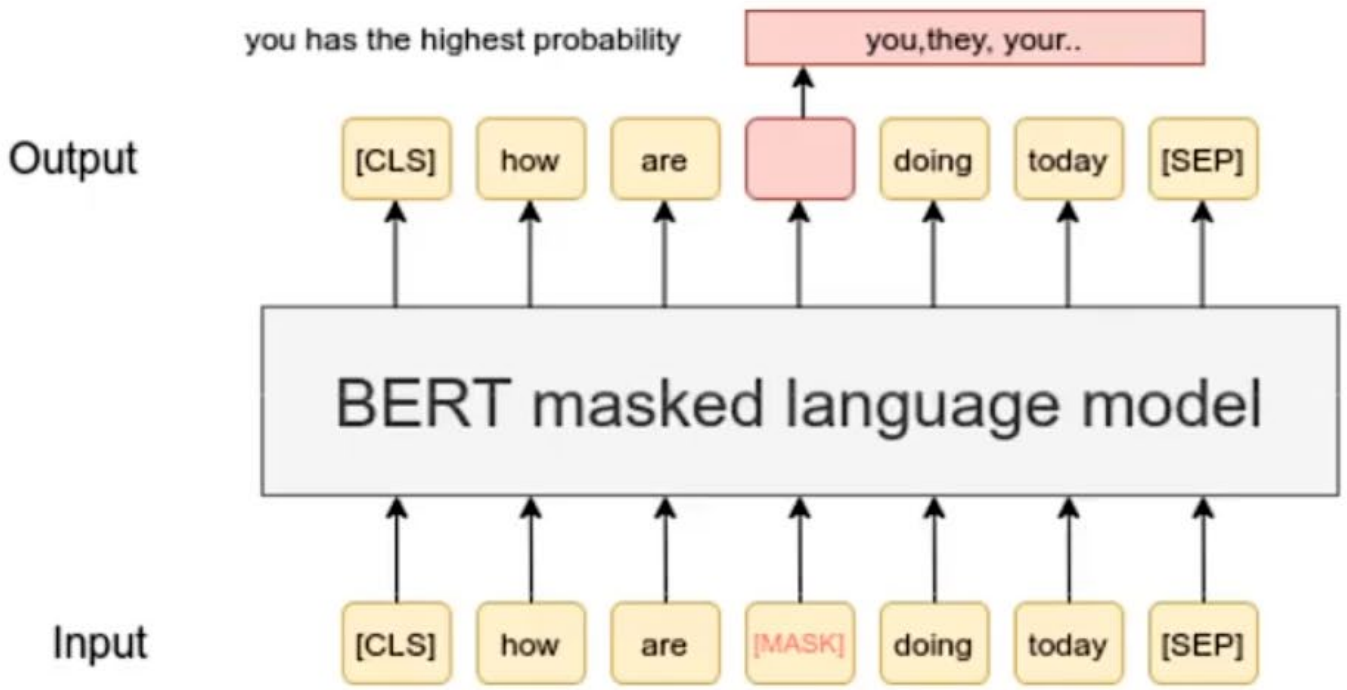
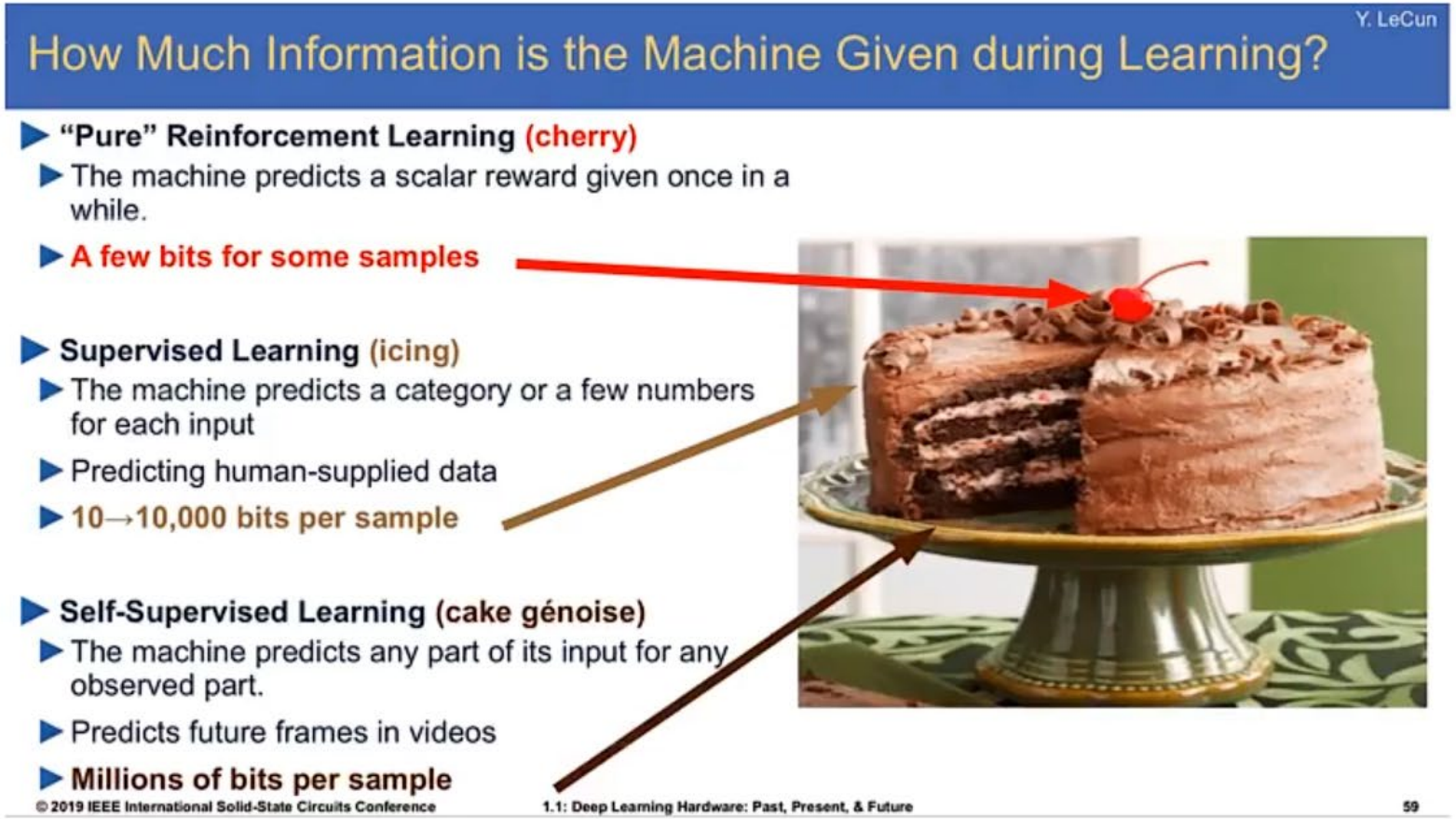


Fig. 1. BERT-Original sentence 'how are you doing today'

自监督学习（Self-supervised Learning）

Yann LeCun

SSL is the future!



概率密度估计

概率密度估计

- 参数密度估计（Parametric Density Estimation）
  - 根据先验知识假设随机变量服从某种分布，然后通过训练样本来估计分布的参数.
  - 估计方法：最大似然估计

$$\log p(\mathcal{D};\theta) = \sum_{n=1}^{N} \log p(\boldsymbol{x}^{(n)};\theta).$$

- 非参数密度估计（Nonparametric Density Estimation）
  - 不假设数据服从某种分布，通过将样本空间划分为不同的区域并估计每个区域的概率来近似数据的概率密度函数。

参数密度估计

正态分布

假设样本 $\boldsymbol{x} \in \mathbb{R}^D$ 服从正态分布

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\Big(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\Big),$$

其中 $\boldsymbol{\mu}$ 和 $\boldsymbol{\Sigma}$ 分别为正态分布的均值和方差.

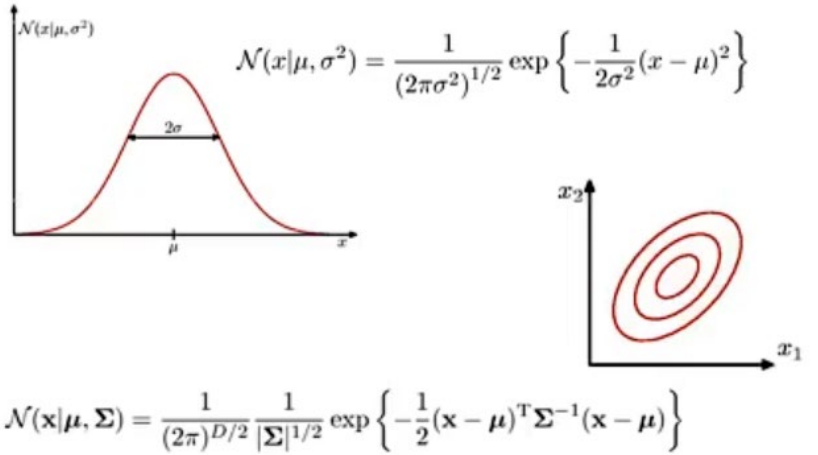
数据集 $\mathcal{D} = \{\boldsymbol{x}^{(n)}\}_{n=1}^N$ 的对数似然函数为

$$\log p(\mathcal{D}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = -\frac{N}{2}\log\Big((2\pi)^D|\boldsymbol{\Sigma}|\Big) - \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}^{(n)}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}^{(n)}-\boldsymbol{\mu}).$$

分别求上式关于 $\boldsymbol{\mu},\boldsymbol{\Sigma}$ 的偏导数,并令其等于0. 可得,

$$\boldsymbol{\mu}^{ML} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}^{(n)},$$

$$\boldsymbol{\Sigma}^{ML} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}^{(n)}-\boldsymbol{\mu}^{ML})(\boldsymbol{x}^{(n)}-\boldsymbol{\mu}^{ML})^{\mathrm{T}}.$$



参数密度估计

多项分布

数据集 $\mathcal{D} = \{\boldsymbol{x}^{(n)}\}_{n=1}^{N}$ 的对数似然函数为

$$\log p(\mathcal{D}|\boldsymbol{\mu}) = \sum_{n=1}^{N}\sum_{k=1}^{K} x_k^{(n)} \log(\mu_k). \tag{9.34}$$

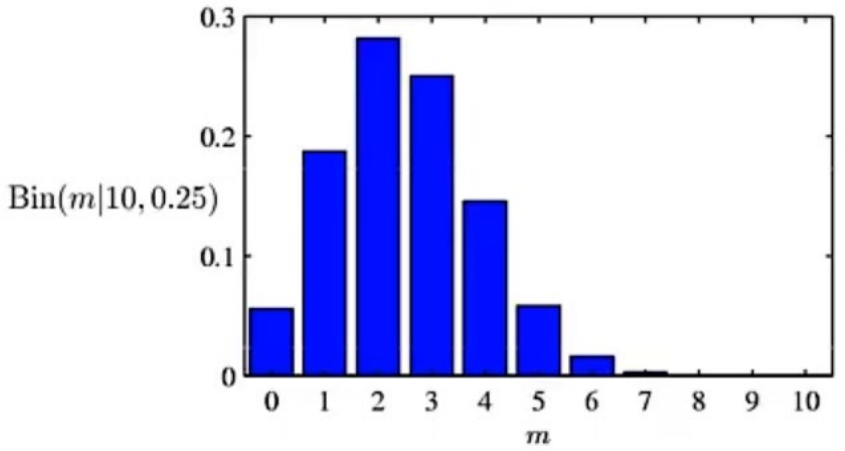
多项分布的参数估计为约束优化问题. 引入拉格朗日乘子 λ, 将原问题转换为无约束优化问题.

$$\max_{\mu,\lambda} \quad \sum_{n=1}^{N}\sum_{k=1}^{K} x_k^{(n)} \log(\mu_k) + \lambda\Big(\sum_{k=1}^{K}\mu_k - 1\Big). \tag{9.35}$$

分别求上式关于 μ_k, λ 的偏导数, 并令其等于 0. 可得,

$$\mu_k^{ML} = \frac{m_k}{N}, \qquad 1 \leq k \leq K \tag{9.36}$$

其中 $m_k = \sum_{n=1}^{N} x_k^{(n)}$ 为数据集中取值为第 k 个状态的样本数量.



参数密度估计一般存在以下问题

- 模型选择问题
  - 如何选择数据分布的密度函数?
  - 实际数据的分布往往是非常复杂的，而不是简单的正态分布或多项分布。
- 不可观测变量问题
  - 样本可能只包含部分的可观测变量，还有一些非常关键的变量是无法观测的，这导致我们很难准确估计数据的真实分布。
- 维度灾难问题
  - 高维数据的参数估计十分困难
  - 随着维度的增加，估计参数所需要的样本数量指数增加。在样本不足时会出现过拟合。

非参数密度估计

- 对于高维空间中的一个随机向量$\boldsymbol{x}$，假设其服从一个未知分布$p(\boldsymbol{x})$，则$\boldsymbol{x}$落入空间中的小区域$\mathcal{R}$的概率为

$$P = \int_{\mathcal{R}} p(\boldsymbol{x})d\boldsymbol{x}.$$

- 给定N个训练样本$D = \{\boldsymbol{x}^{(n)}\}_{n=1}^{N}$，落入区域R的样本数量K服从二项分布

$$P_K = \binom{N}{K} P^K (1-P)^{1-K},$$

- 当N非常大时，我们可以近似认为

$$P \approx \frac{K}{N}$$

- 假设区域R足够小，其内部的概率密度是相同的，则有

$$P \approx p(\boldsymbol{x})V$$

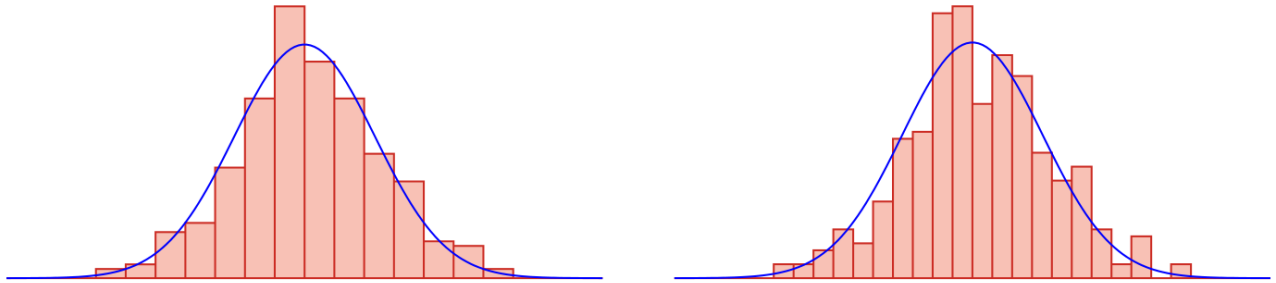
- 结合上述两个公式，得到

$$p(\boldsymbol{x}) \approx \frac{K}{NV}$$

直方图方法（Histogram Method）

- 一种非常直观的估计连续变量密度函数的方法，可以表示为一种柱状图。

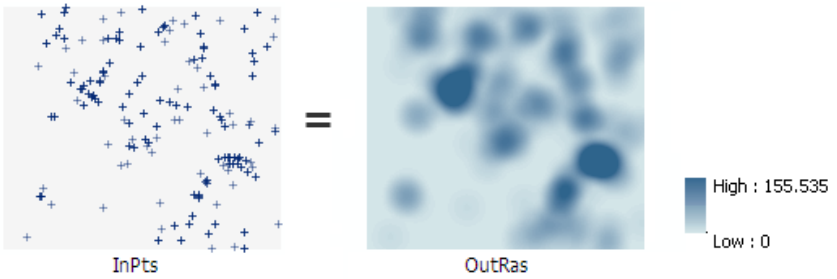
以一维随机变量为例，首先将其取值范围分成M个连续的、不重叠的区间（bin），每个区间的宽度为Δ_m. 给定N个训练样本$\mathcal{D}=\{x^{(n)}\}_{n=1}^{N}$，我们统计这些样本落入每个区间的数量$K_m$，然后将它们归一化为密度函数.



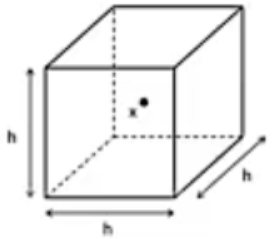
(a) 10个区间（bin）　　(b) 30个区间（bin）

核密度估计（Kernel Density Estimation）

- 核密度估计是一种直方图方法的改进。
  - 也叫Parzen窗方法



- 假设$\mathcal{R}$为d维空间中的一个以点$\boldsymbol{x}$为中心的“超立方体”，并定义核函数来表示一个样本$\boldsymbol{z}$是否落入该超立方体中

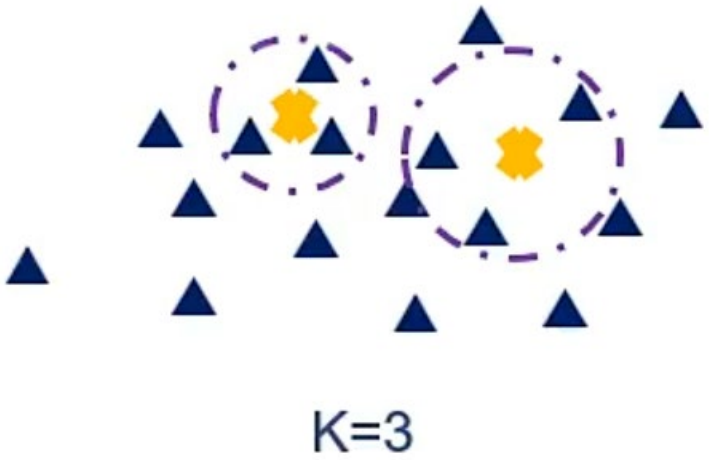


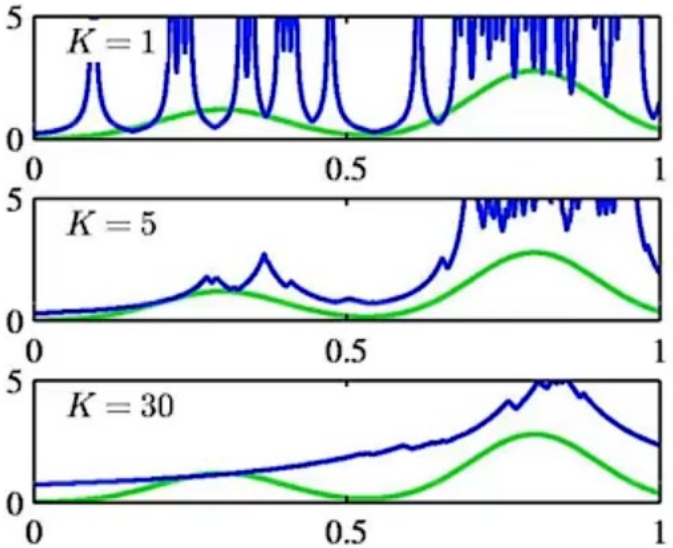
$$\phi\left(\frac{\boldsymbol{z}-\boldsymbol{x}}{h}\right)=\begin{cases}1 & \text{if } |z_i-x_i|<\frac{H}{2}, 1\le i\le D\\ 0 & \text{else}\end{cases}$$

- 点$\boldsymbol{x}$的密度估计为 $p(\boldsymbol{x})=\frac{K}{NH^D}=\frac{1}{NH^D}\sum_{n=1}^{N}\phi\left(\frac{\boldsymbol{x}^{(n)}-\boldsymbol{x}}{H}\right)$

K近邻方法

- 一种更灵活的方式是设置一种可变宽度的区域，并使得落入每个区域中样本数量为固定的K。
- 要估计点x的密度，首先找到一个以x为中心的球体，使得落入球体的样本数量为K，就可以计算出点x的密度。





K acts as a smoother.

非参数密度估计

- 非参数密度估计(直方图方法除外)需要保留整个训练集。
- 而参数密度估计不需要保留整个训练集，因此在存储和计算上更加高效。

思考题

- 估计以下三种密度函数的区别： p(x), p(y|x) and p(x|y)?
  - 生成模型Generative model
  - 判别式模型Discriminative model

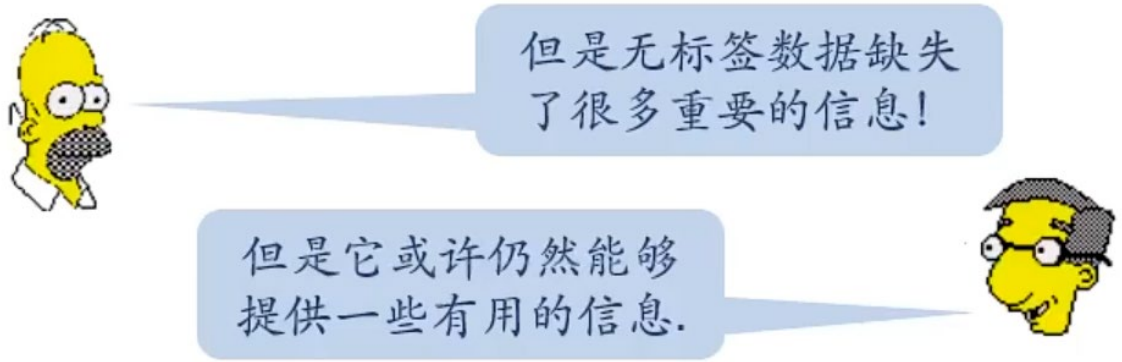
- 最大似然函数的本质是什么？
  - 真实数据分布 $p_r(x)$
  - 模型数据分布 $p_\theta(x)$

半监督学习

半监督学习（Semi-supervised Learning）

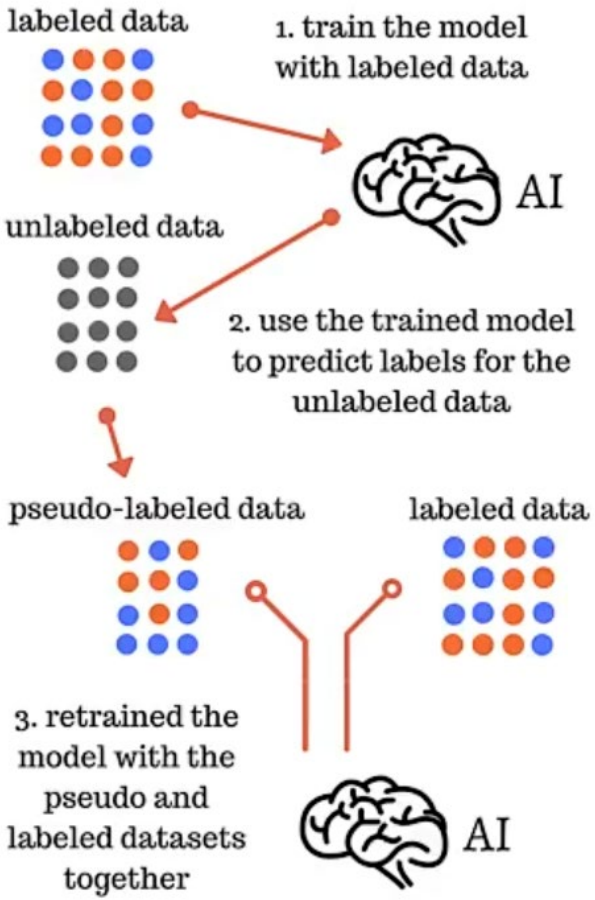
- 监督学习
  - 提供任务相关的标签
  - 打标签费时费力，数量有限
- 无监督学习
  - 不用打标签，数量充足
  - 仅限特定任务

同时利用少量有标签数据和大量无标签数据?

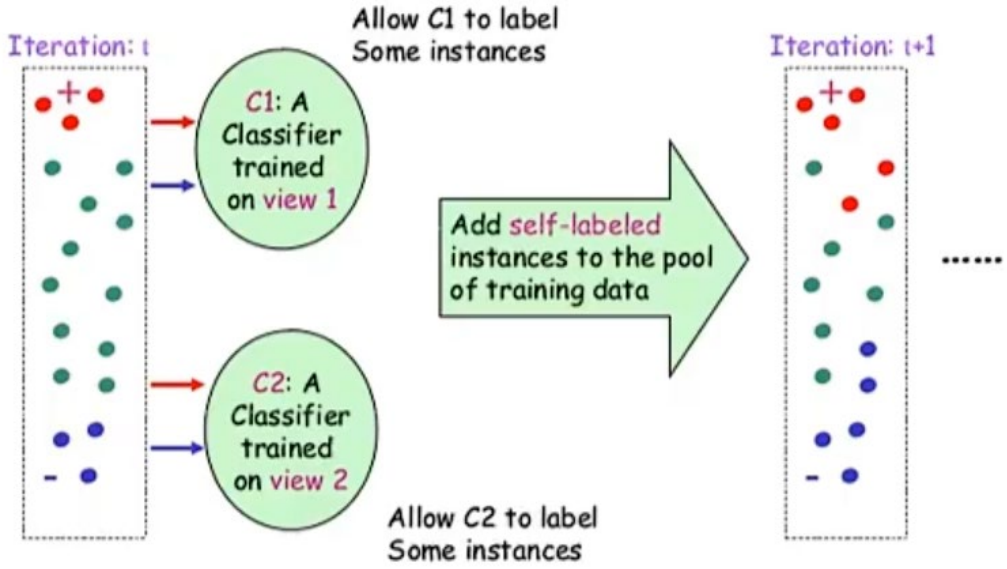


半监督学习（Semi-supervised Learning）

自训练（Self-training）



协同训练（Co-training）



谢 谢