

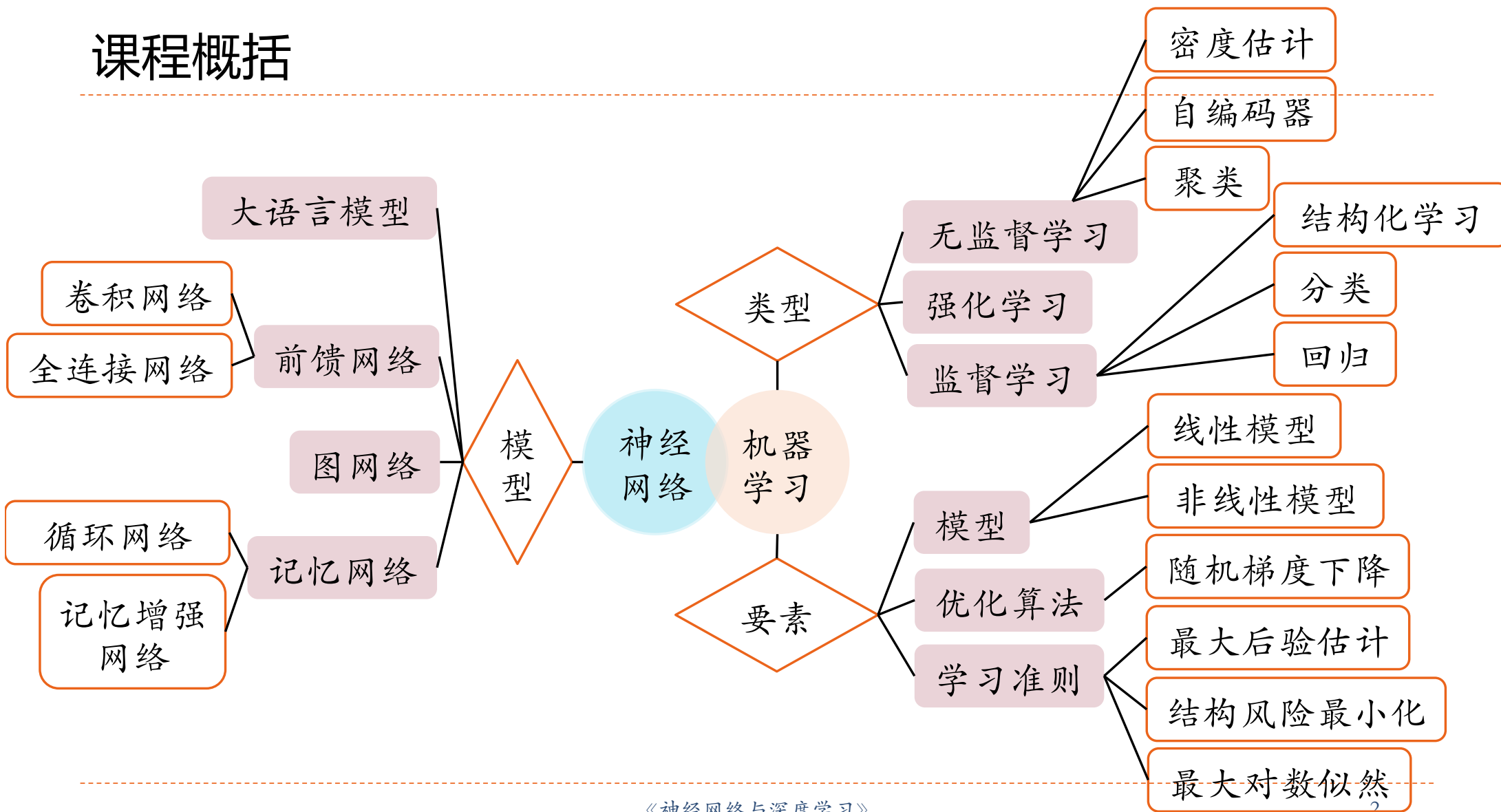
# 《神经网络与深度学习》



## 无监督学习

<https://nndl.github.io/>

# 课程概括



# 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

### ▶ 概率密度估计

- ▶ 参数密度估计
- ▶ 非参数密度估计
  - 核方法
  - K近邻方法

## ▶ 生成式模型vs判别式模型

---

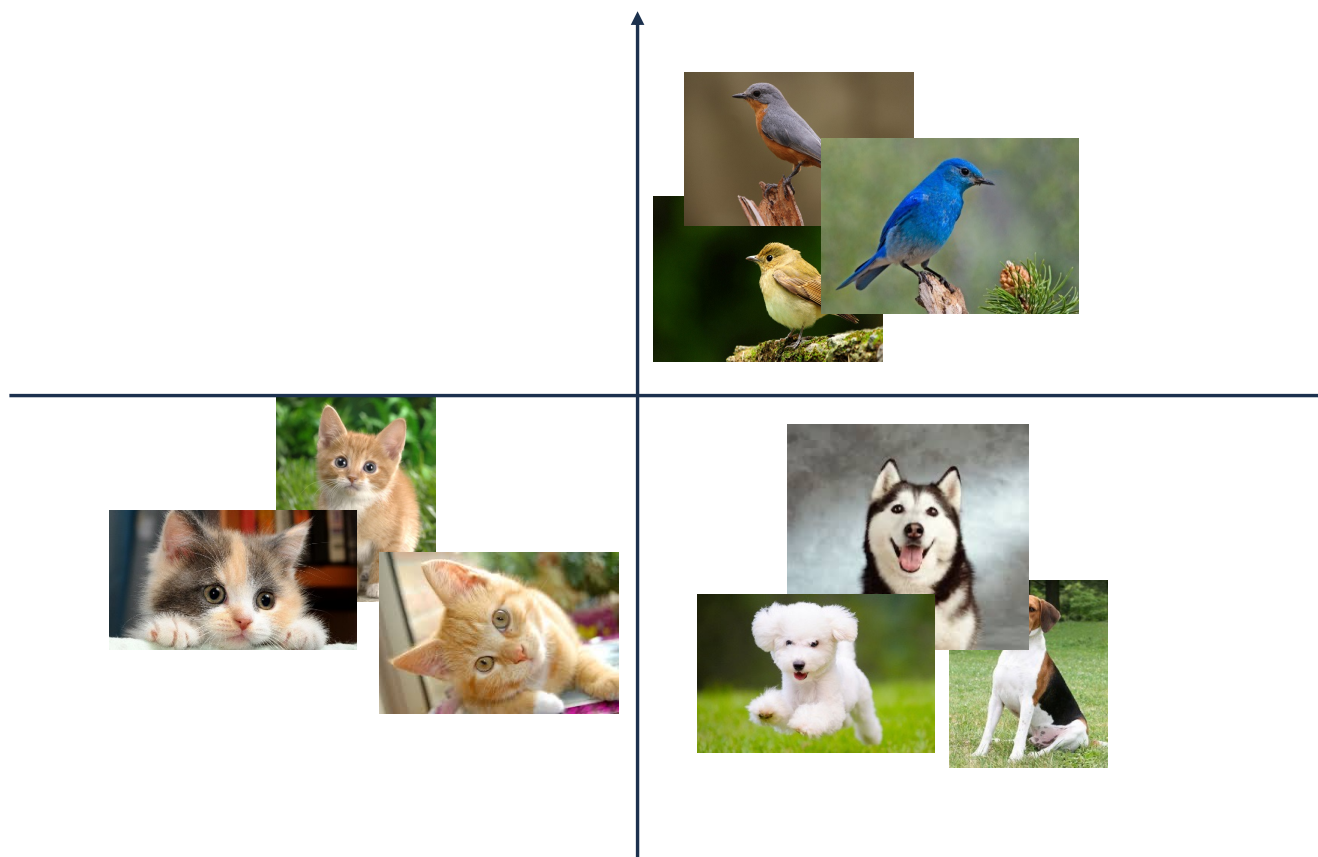
# 芒果机器学习的无监督学习

---

- ▶从市场上随机选取的芒果样本（训练数据），列出每个芒果的所有特征：
  - ▶如颜色，果子大小，果核大小，形状，产地，品牌， ...
- ▶设计一个学习算法来学习芒果的有价值的特征

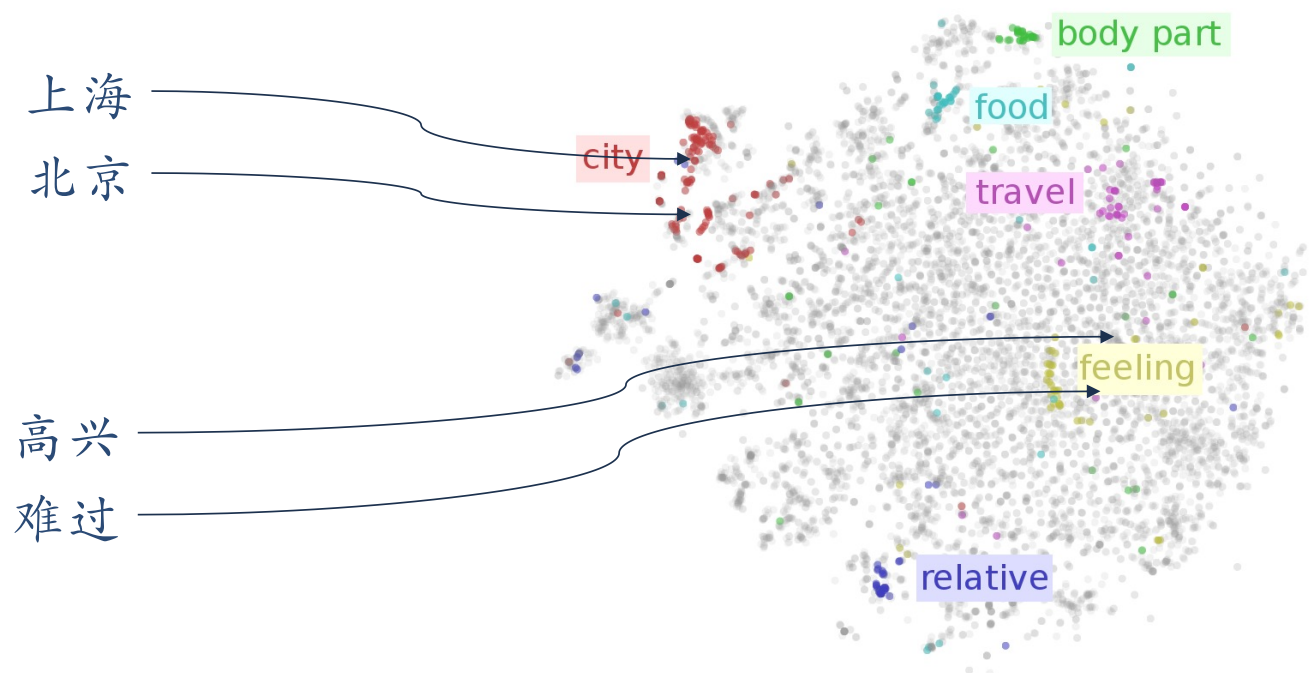
# 计算机视觉中的无监督学习

---



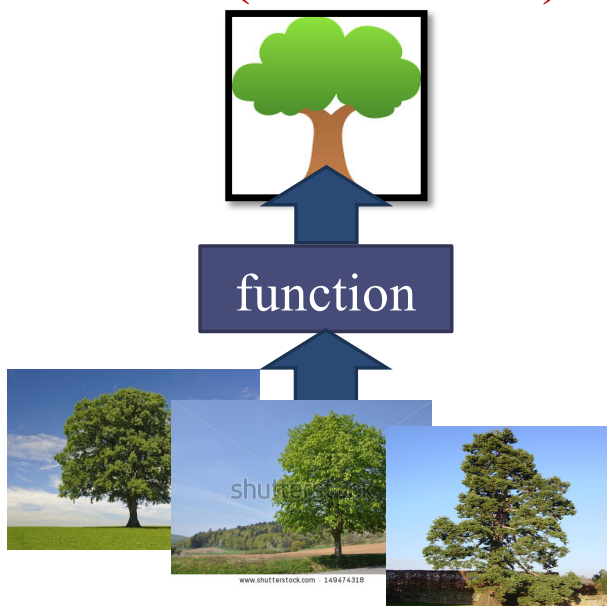
# 自然语言处理中的无监督学习

---

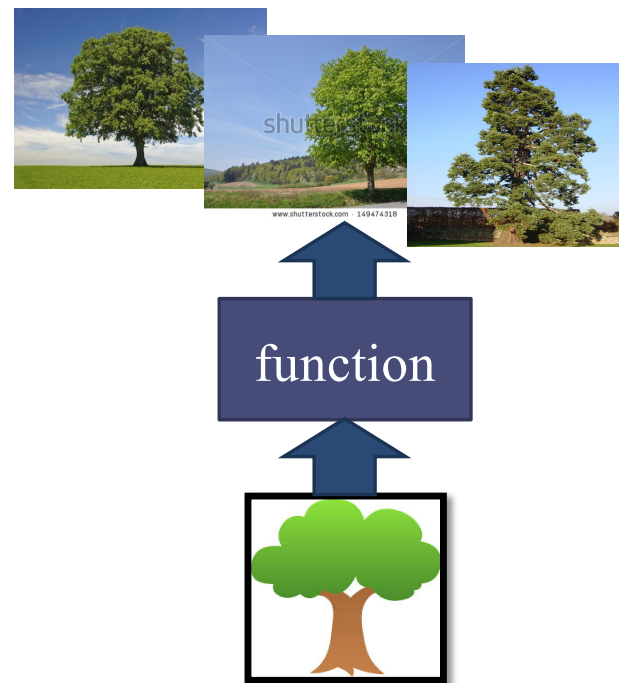


# 深度学习中的无监督学习

## ► Dimension Reduction (化繁为简)



## ► Generation (无中生有)



# 无监督学习 ( Unsupervised Learning )

---

## ▶ 监督学习

- ▶ 建立映射关系  $f: x \rightarrow y$

## ▶ 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  $f: x \rightarrow y$ 
  - ▶ 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
- ▶ 密度估计  $p(x)$



# 无监督学习同样有三个基本要素

---

## ▶ 模型

- ▶ 自编码器、概率模型

## ▶ 学习准则

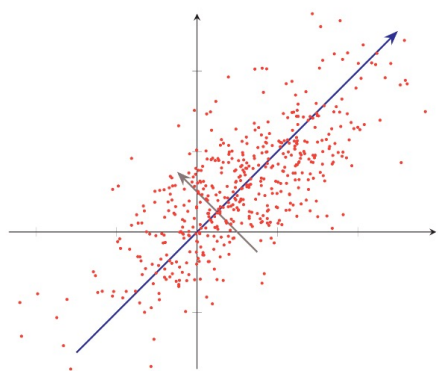
- ▶ 最大似然估计、最小重构错误

## ▶ 优化

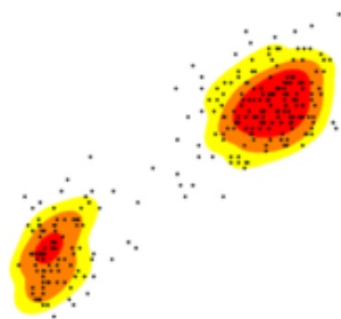
- ▶ 梯度下降

# 典型的无监督学习问题

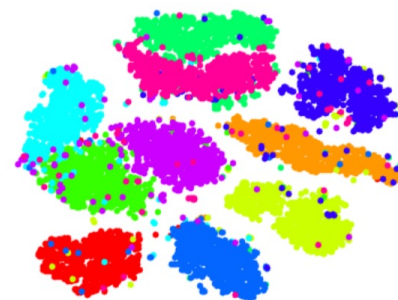
---



无监督特征学习



密度估计



聚类



## 无监督特征学习

# 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  $\mathbf{x}^{(n)} \in R^D$  投影之后的表示为

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

▶ 所有样本投影后的方差为

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

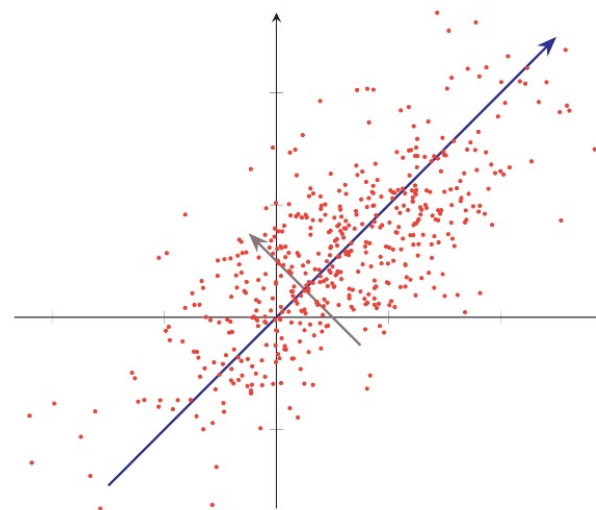
$\Sigma = \frac{1}{N}(\mathbf{X} - \bar{\mathbf{X}})(\mathbf{X} - \bar{\mathbf{X}})^\top$  是原始样本的协方差矩阵

▶ 目标函数

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

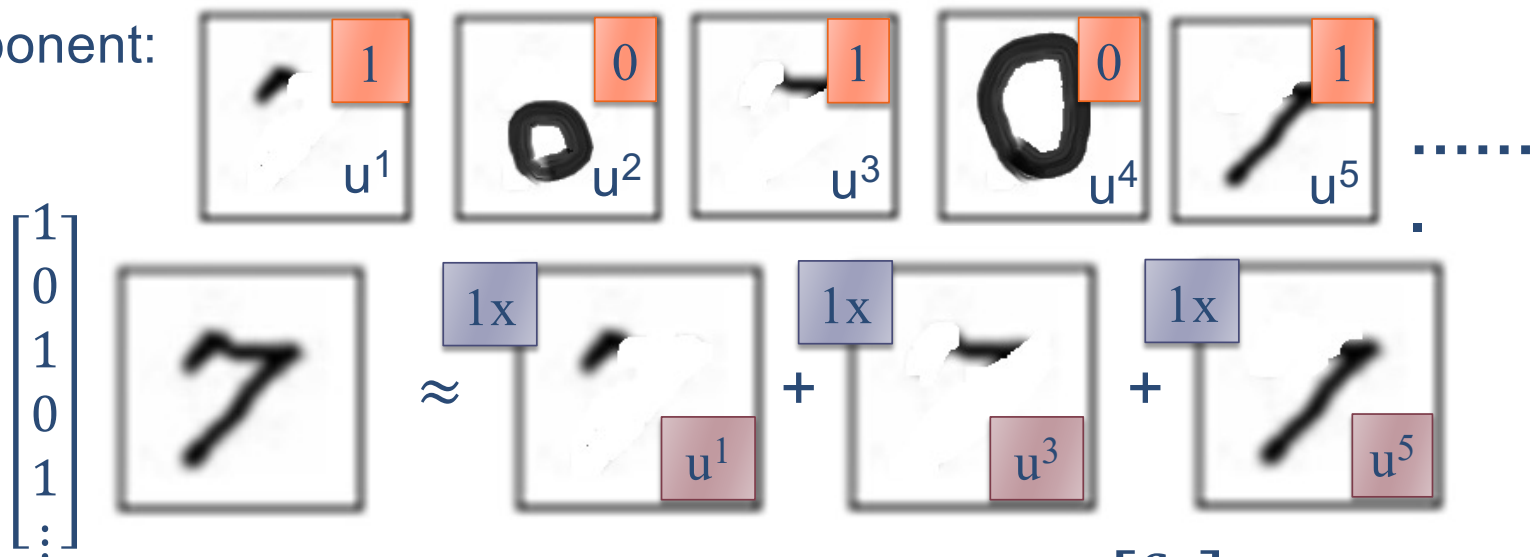
▶ 对目标函数求导并令导数等于 0，可得

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



# PCA的另一种视角

Basic Component:




$$x \approx c_1 u^1 + c_2 u^2 + \dots + c_K u^K + \bar{x}$$

Pixels in a digit image

component

$$\begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_K \end{bmatrix} \text{ Represent a digit image}$$

## PCA的另一种视角

$$x - \bar{x} \approx c_1 u^1 + c_2 u^2 + \cdots + c_K u^K = \hat{x}$$


Reconstruction error:

$$\| (x - \bar{x}) - \hat{x} \|_2$$

Find  $\{u^1, \dots, u^K\}$  minimizing the error

$$L = \min_{\{u^1, \dots, u^K\}} \sum \left\| (x - \bar{x}) - \underbrace{\left( \sum_{k=1}^K c_k u^k \right)}_{\hat{x}} \right\|_2$$

结论:

$\{w^1, w^2, \dots, w^K\}$  (from PCA) is the component  
 $\{u^1, u^2, \dots, u^K\}$  minimizing L

Proof in [Bishop, Chapter 12.1.2]

## PCA的另一种视角

$$x - \bar{x} \approx c_1 u^1 + c_2 u^2 + \dots + c_K u^K = \hat{x}$$

Reconstruction error:

$$\| (x - \bar{x}) - \hat{x} \|_2$$

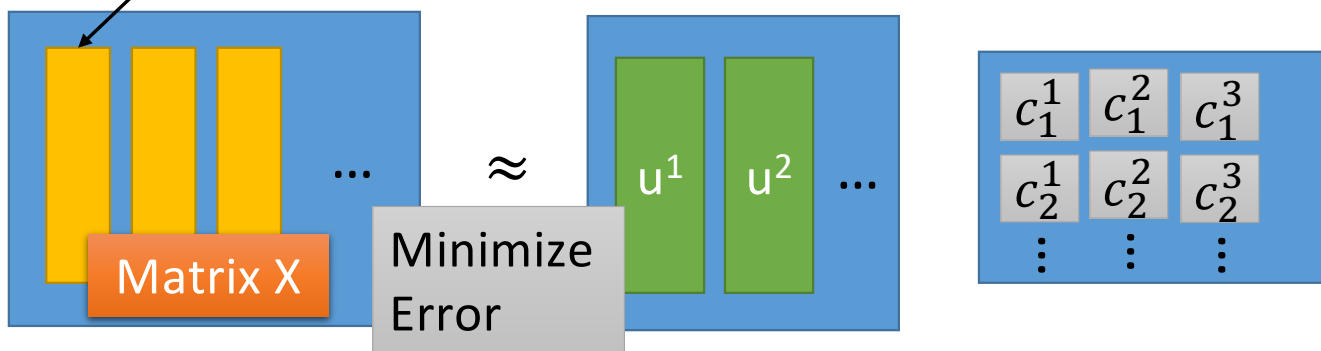
Find  $\{u^1, \dots, u^K\}$  minimizing the error

$$\underline{x^1 - \bar{x}} \approx \underline{c_1^1 u^1} + \underline{c_2^1 u^2} + \dots$$

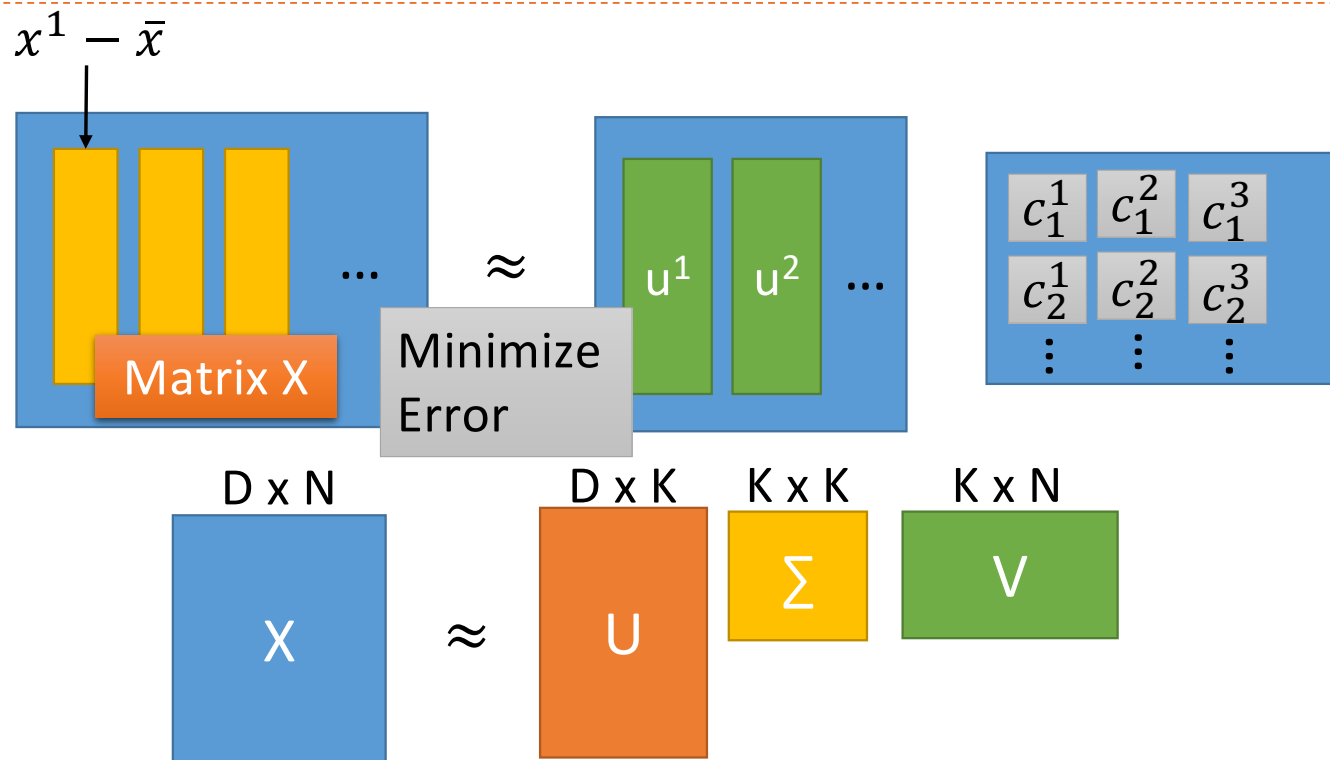
$$x^2 - \bar{x} \approx c_1^2 u^1 + c_2^2 u^2 + \dots$$

$$x^3 - \bar{x} \approx c_1^3 u^1 + c_2^3 u^2 + \dots$$

.....



## PCA的另一种视角



$K$  columns of  $U$ : a set of orthonormal eigen vectors corresponding to the  $K$  largest eigenvalues of  $XX^T$

This is the solution of PCA : SVD



## (线性) 编码

- 给定一组基向量  $A = [a_1, \dots, a_M]$ ，将输入样本  $x$  表示为这些基向量的线性组合

$$x = \sum_{m=1}^M z_m a_m$$
$$= Az,$$

字典 (dictionary)

编码 (encoding)

$$x = Az$$

# 完备性

## 数学小知识 | 完备性

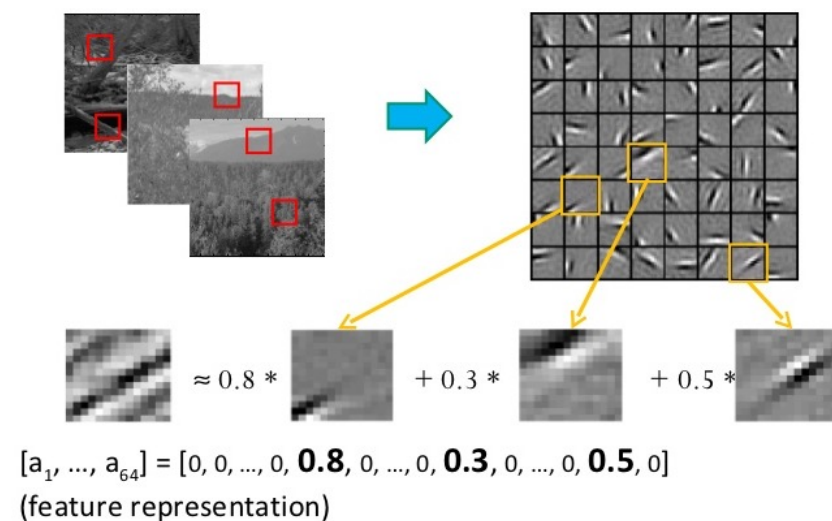
如果  $M$  个基向量刚好可以支撑  $M$  维的欧氏空间, 则这  $M$  个基向量是完备的. 如果  $M$  个基向量可以支撑  $D$  维的欧氏空间, 并且  $M > D$ , 则这  $M$  个基向量是过完备的 (overcomplete)、冗余的.

“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



Slide credit: Andrew Ng

Compact & easily interpretable

## 稀疏编码 ( Sparse Coding )

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

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

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

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

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

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

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

# 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  $\mathbf{A}$ , 对每个输入  $\mathbf{x}^{(n)}$ , 计算其对应的最优编码

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

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

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

# 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 )

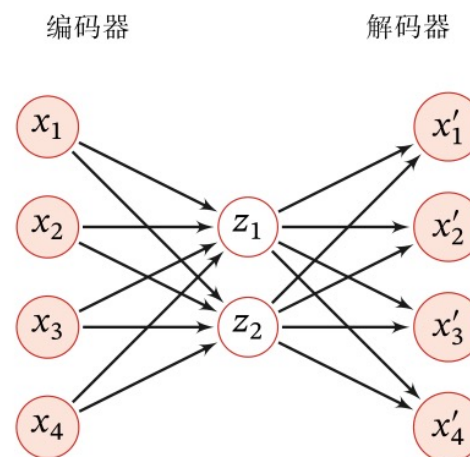
► 编码器 (Encoder)  $f : \mathbb{R}^D \rightarrow \mathbb{R}^M$

► 解码器 (Decoder)  $g : \mathbb{R}^M \rightarrow \mathbb{R}^D$

► 目标函数：重构错误

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

► 两层网络结构的自编码器



# 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。

$$\mathbf{z} = f(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}),$$

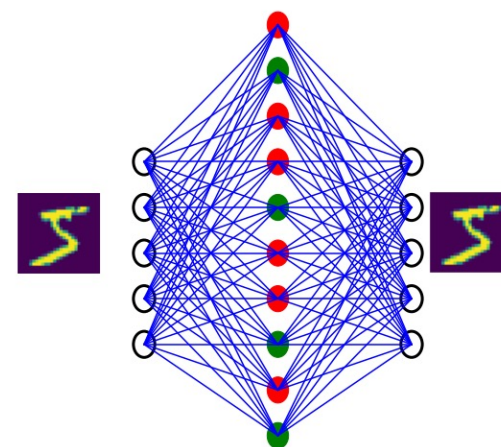
$$\mathbf{x}' = f(\mathbf{W}^{(2)}\mathbf{z} + \mathbf{b}^{(2)}),$$

- ▶ 目标函数

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

- ▶  $\mathbf{W}$ 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。



# From PCA to Auto-Encoder

PCA looks like a neural network with one hidden layer (linear activation function)

Autoencoder

If  $\{w^1, w^2, \dots, w^K\}$  is the component  $\{u^1, u^2, \dots, u^K\}$

$$\hat{x} = \sum_{k=1}^K c_k w^k \longleftrightarrow x - \bar{x}$$

To minimize reconstruction error:

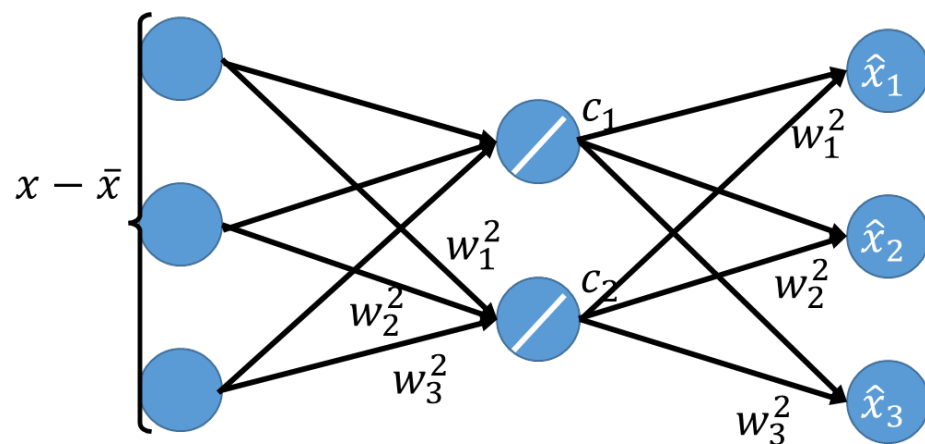
$$c_k = (x - \bar{x}) \cdot w^k$$

$K = 2$ :

It can be deep.



Deep Autoencoder

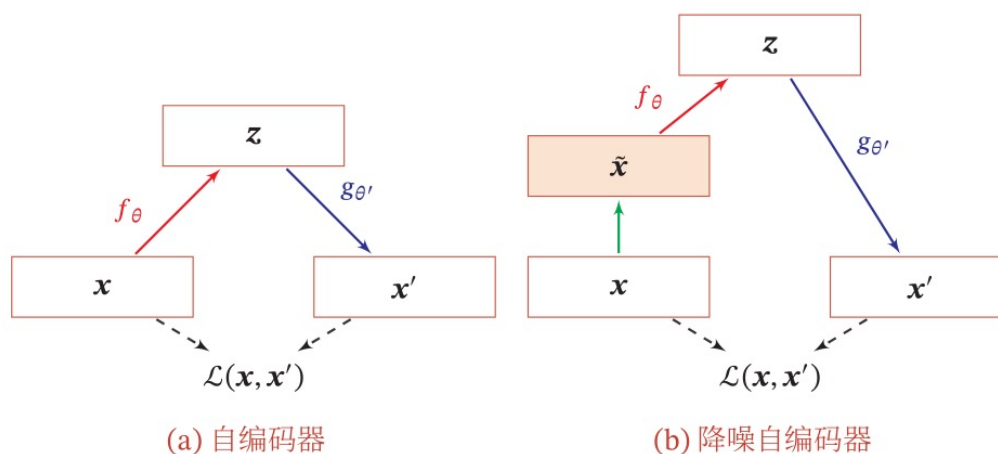


Minimize  
error  
 $\longleftrightarrow x - \bar{x}$



# 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 $\mathbf{x}$ ，我们首先根据一个比例 $\mu$ 随机将 $\mathbf{x}$ 的一些维度的值设置为0，得到一个被损坏的向量 $\tilde{\mathbf{x}}$ 。
- ▶ 然后将被损坏的向量 $\tilde{\mathbf{x}}$ 输入给自编码器得到编码 $\mathbf{z}$ ，并重构出原始的无损输入 $\mathbf{x}$ 。





## 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

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

## ▶ 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  $\mathbf{x} \in \mathbb{R}^D$  服从正态分布

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right),$$

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

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

$$\log p(\mathcal{D}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{N}{2} \log\left((2\pi)^D |\boldsymbol{\Sigma}|\right) - \frac{1}{2} \sum_{n=1}^N (\mathbf{x}^{(n)} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}^{(n)} - \boldsymbol{\mu}).$$

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

$$\begin{aligned}\boldsymbol{\mu}^{ML} &= \frac{1}{N} \sum_{n=1}^N \mathbf{x}^{(n)}, \\ \boldsymbol{\Sigma}^{ML} &= \frac{1}{N} \sum_{n=1}^N (\mathbf{x}^{(n)} - \boldsymbol{\mu}^{ML})(\mathbf{x}^{(n)} - \boldsymbol{\mu}^{ML})^\top.\end{aligned}$$

# 参数密度估计

## ► 多项分布

假设样本服从  $K$  个状态的多项分布, 令 one-hot 向量  $\mathbf{x} \in \{0, 1\}^K$  来表示第  $k$  个状态, 即  $x_k = 1$ , 其余  $x_{i, i \neq k} = 0$ . 样本  $\mathbf{x}$  的概率密度函数为

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^K \mu_k^{x_k}, \quad (9.33)$$

其中  $\mu_k$  为第  $k$  个状态的概率, 并满足  $\sum_{k=1}^K \mu_k = 1$ .

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

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

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

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

分别求上式关于  $\mu_k, \lambda$  的偏导数, 并令其等于 0. 可得,

$$\mu_k^{ML} = \frac{m_k}{N}, \quad 1 \leq k \leq K \quad (9.36)$$

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

- ▶ 即我们用来训练的样本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 $\mathbf{x}$ ，假设其服从一个未知分布 $p(\mathbf{x})$ ，则 $\mathbf{x}$ 落入空间中的小区域 $\mathcal{R}$ 的概率为

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

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

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

- ▶ 当 $N$ 非常大时，我们可以近似认为

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

- ▶ 假设区域 $\mathcal{R}$ 足够小，其内部的概率密度是相同的，则有

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

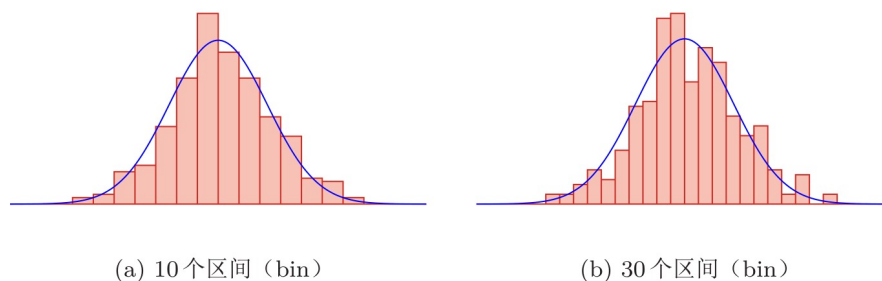
- ▶ 结合上述两个公式，得到

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

## 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  $\Delta_m$ . 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.



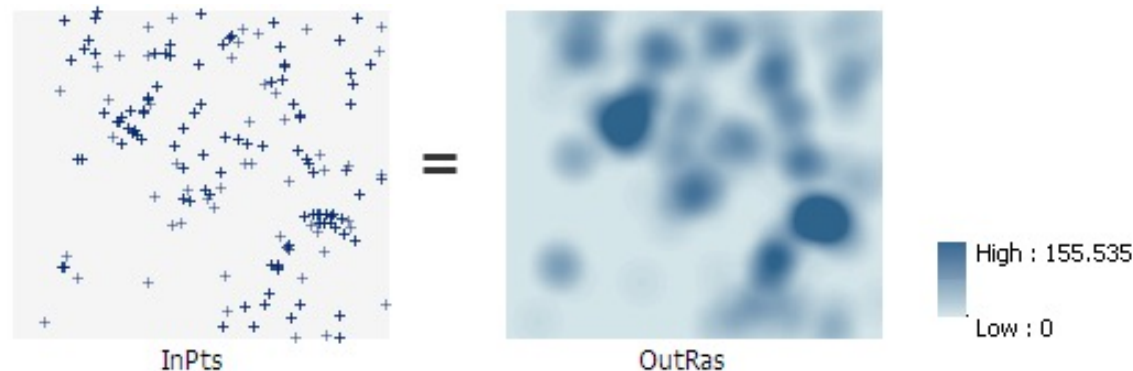
- ▶ 缺点：很难扩展到高维变量



# 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 $\mathbf{x}$ 为中心的“超立方体”，并定义核函数来表示一个样本 $\mathbf{z}$ 是否落入该超立方体中

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

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

## K近邻方法

---

- ▶ 核密度估计方法中的核宽度是固定的，因此同一个宽度可能对高密度的区域过大，而对低密度区域过小。
- ▶ 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。