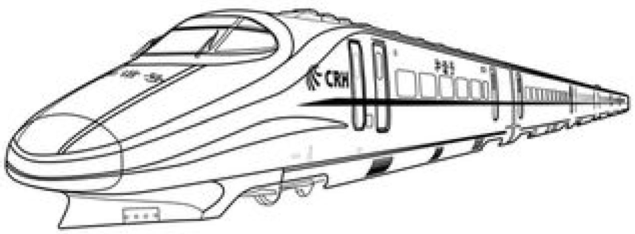
北京交通大學
BEIJING JIAOTONG UNIVERSITY

智能信息技术教育中心
The Education Center of Intelligence Information Technologies

人工智能基础

机器学习

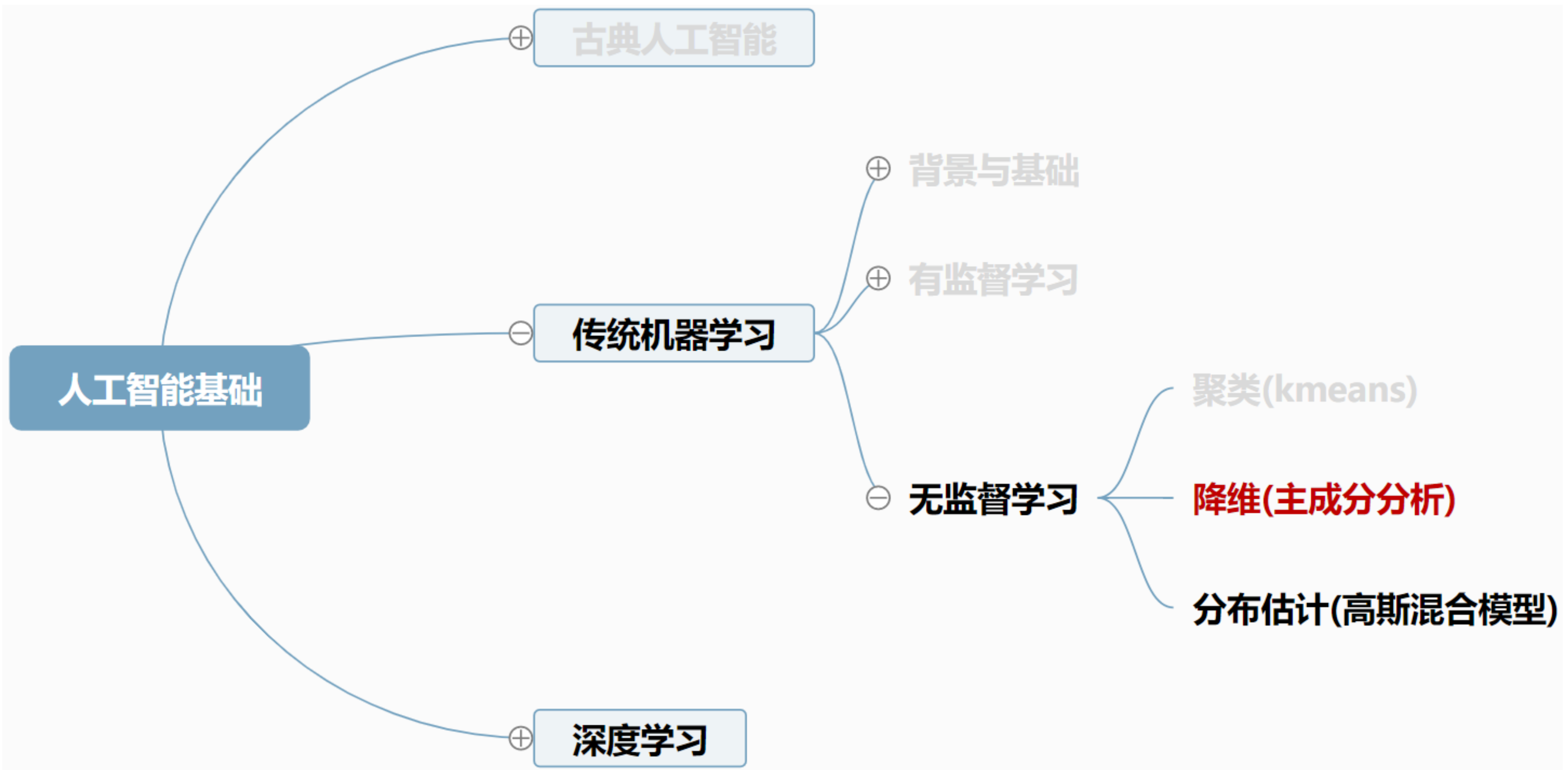
耿阳李敖
2022年3月



内容概要

- **内容回顾**
- 带参分布估计
- 非参分布估计
- 总结

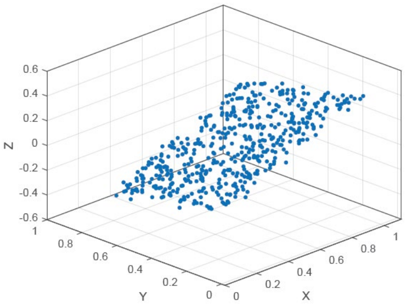
内容回顾

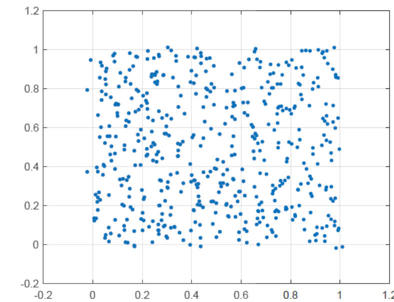

古典人工智能
背景与基础
有监督学习
传统机器学习
人工智能基础
聚类(kmeans)
无监督学习
降维(主成分分析)
分布估计(高斯混合模型)
深度学习

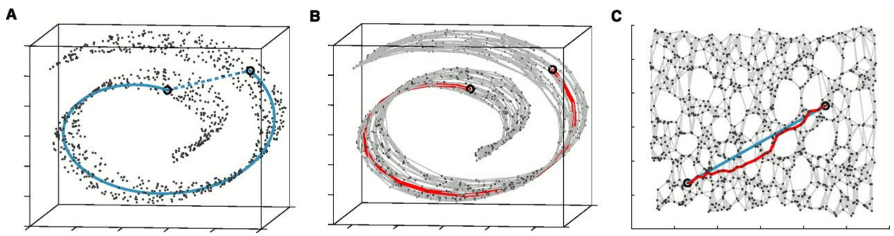
降维算法

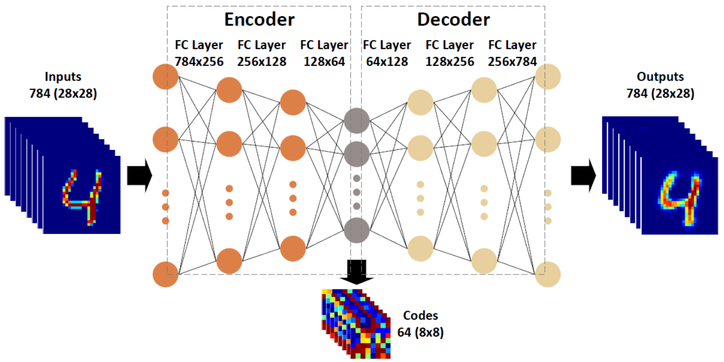
动机

- 发现高维数据内在的低维嵌入，刻画数据的本质属性，减小数据存储
- 线性降维
- 非线性降维
- 神经网络方法



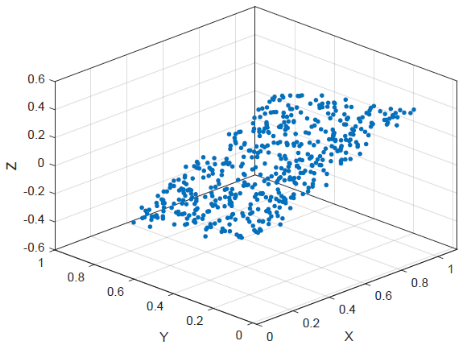


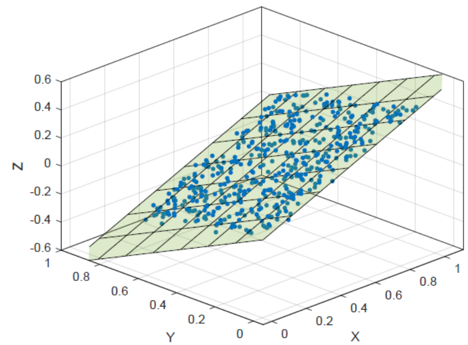


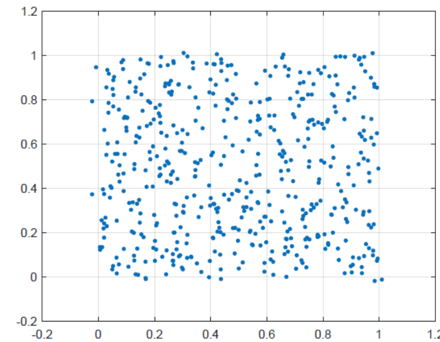


主成分分析（PCA）

主要思想







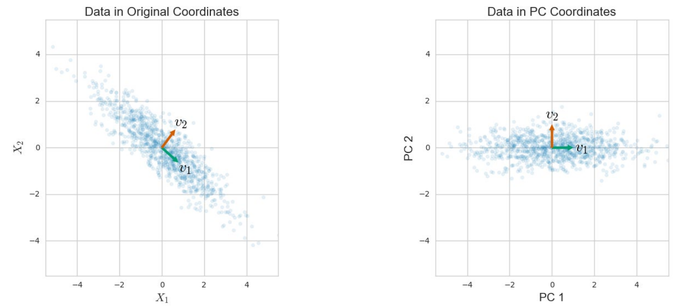
算法流程

- 输入：原始数据点 $\{\mathbf{x}_i \in \mathbb{R}^D\}_{i=1}^N$，目标维度 $d\ (d \ll D)$
- 求样本均值$\boldsymbol{\mu}^\star = \frac{1}{N}\sum_{i=1}^N \mathbf{x}_i$，与样本协方差矩阵$\boldsymbol{\Sigma} = \sum_{i=1}^N (\mathrm{x}_i - \mu^\star)(\mathrm{x}_i - \mu^\star)^T$，计算$\boldsymbol{\Sigma}$ 前 d 大特征值对应的特征向量构成的矩阵$\mathbf{P}^\star$
- 数据降维：$\mathbf{z}_i = \mathbf{P}^{\star T}(\mathbf{x}_i - \boldsymbol{\mu}^\star)$

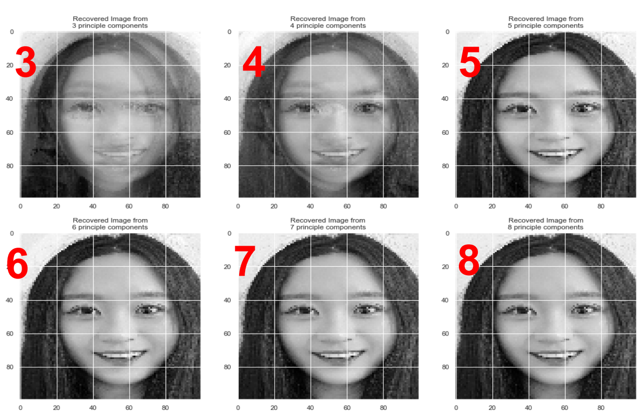
主成分分析（PCA）

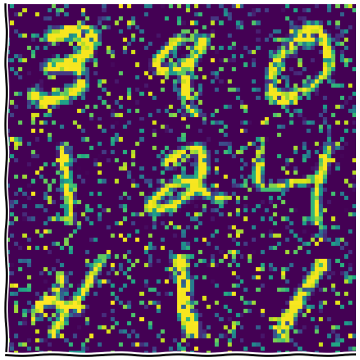
PCA的统计特性

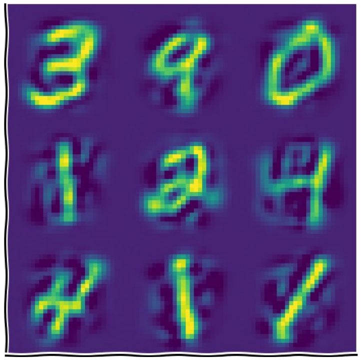
经过PCA变换后，所获得数据的均值为 0，不同维度之间彼此线性无关，且在正交变换意义下保留了原始数据“最多”的方差信息



PCA的应用







主成分分析（PCA）

PCA的改进

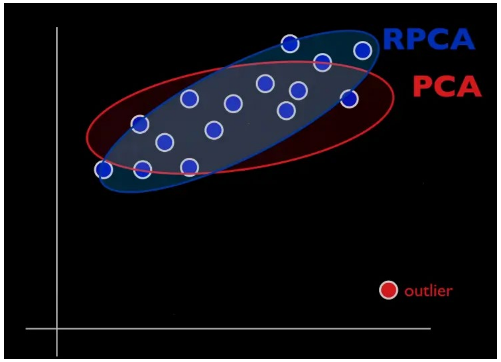
- 应对异常数据：鲁棒PCA（RPCA）

$$\min_{\mathbf{L}\in\mathbb{R}^{D\times D}} \quad \|\mathbf{\Sigma}-\mathbf{L}\|_2^2$$
$$\text{s.t.} \quad \text{rank}(\mathbf{L}) \le d$$

低秩视角PCA

$$\min_{\mathbf{L}\in\mathbb{R}^{D\times D}} \quad \|\mathbf{\Sigma}-\mathbf{L}\|_1$$
$$\text{s.t.} \quad \text{rank}(\mathbf{L}) \le d$$

Robust PCA



- 应对非线性情况：核PCA（Kernel PCA）

$$\phi(\mathbf{X}) = [\phi(\mathbf{x}_1), \phi(\mathbf{x}_2), \dots, \phi(\mathbf{x}_N)] \in \mathbb{R}^{\infty\times N}$$

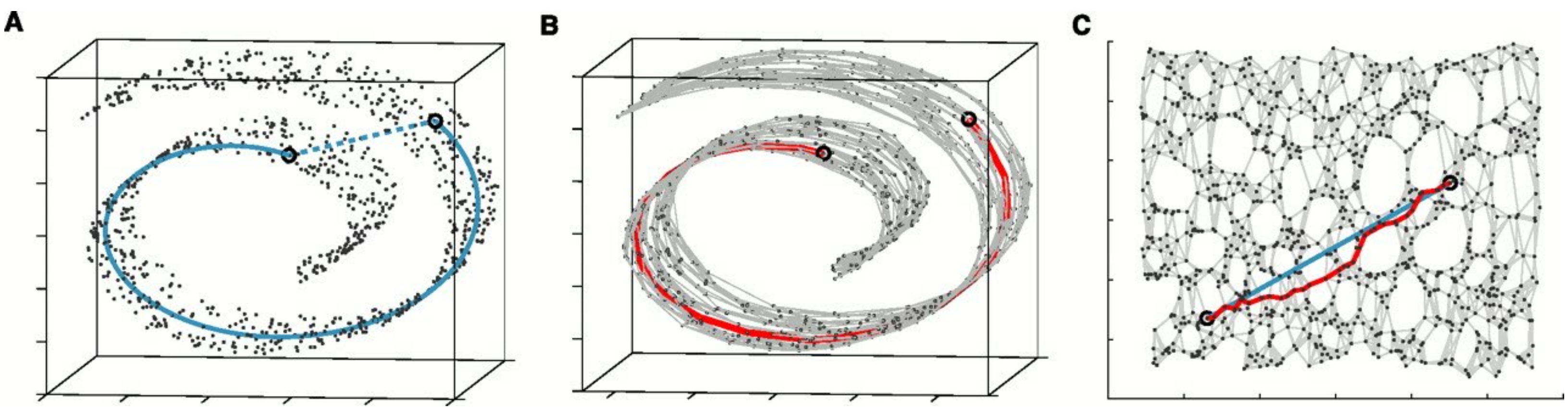
$$\phi(\mathbf{X})^T\phi(\mathbf{X}) = \mathbf{K} \in \mathbb{R}^{N\times N} \quad \mathbf{K}\mathbf{v} = \lambda\mathbf{v}$$

$$\mathbf{z}_i = \mathbf{V}^{\star T}\phi(\mathbf{X})^T\phi(\mathbf{x}_i)$$
$$= \mathbf{V}^{\star T}\mathbf{K}_{\cdot i}$$

非线性降维

等度量映射（Isometric Mapping）

- 算法原理



$$\left\|z_i - z_j\right\|^2 = {z_i}^T z_i + {z_j}^T z_j - 2{z_i}^T z_j$$

距离矩阵可以由内积唯一确定，当样本均值确定时反之亦然

非线性降维

等度量映射（Isometric Mapping）

- 算法步骤

输入： 样本集 $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$;
近邻参数 k;
低维空间维数 d'.

过程：

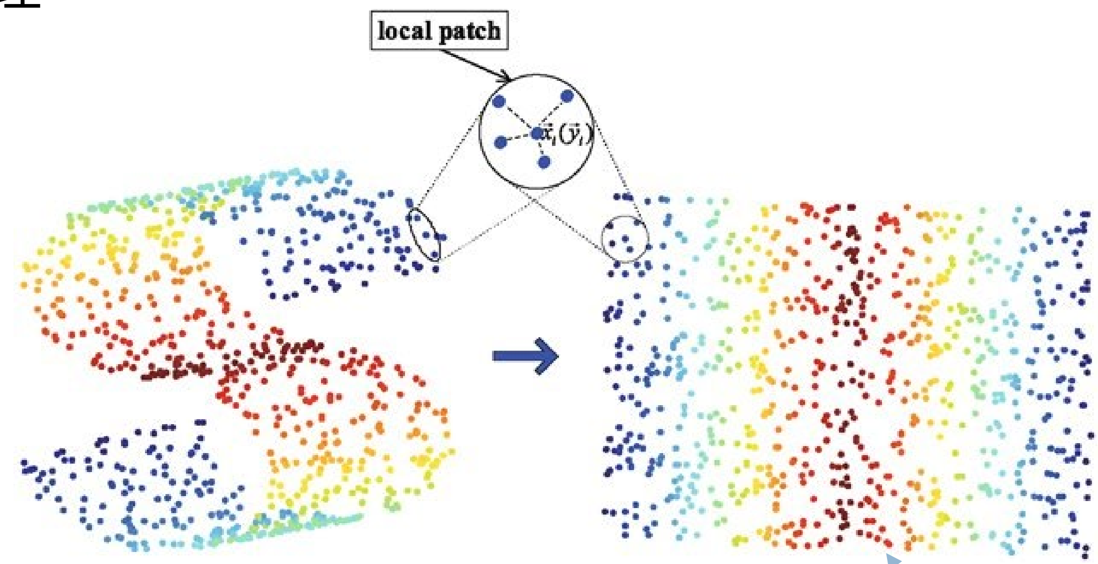
```
for i = 1, 2, ..., m do
    确定 x_i 的 k 近邻;
    x_i 与 k 近邻点之间的距离设置为欧氏距离，与其他点的距离设置为无穷大;
end for
调用最短路径算法计算任意两样本点之间的距离 dist(x_i, x_j);
将 dist(x_i, x_j) 作为 MDS 算法的输入;
return MDS 算法的输出
```

输出： 样本集 D 在低维空间的投影 $Z = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_m\}$.

非线性降维

局部线性嵌入（Locally Linear Embedding）

- 算法原理



$$\min_{\boldsymbol{w}_1,\boldsymbol{w}_2,\ldots,\boldsymbol{w}_m} \sum_{i=1}^{m} \left\| \boldsymbol{x}_i - \sum_{j\in Q_i} w_{ij}\boldsymbol{x}_j \right\|_2^2$$

$$\text{s.t.} \sum_{j\in Q_i} w_{ij} = 1,$$

$$\min_{\boldsymbol{z}_1,\boldsymbol{z}_2,\ldots,\boldsymbol{z}_m} \sum_{i=1}^{m} \left\| \boldsymbol{z}_i - \sum_{j\in Q_i} w_{ij}\boldsymbol{z}_j \right\|_2^2$$

非线性降维

局部线性嵌入（Locally Linear Embedding）

- 算法步骤

$$\min_{z_1,z_2,\ldots,z_m} \sum_{i=1}^{m} \left\| z_i - \sum_{j\in Q_i} w_{ij} z_j \right\|_2^2$$

$$\mathbf{M} = (\mathbf{I}-\mathbf{W})^{\mathrm{T}}(\mathbf{I}-\mathbf{W})$$

$$\min_{\mathbf{Z}} \ \mathrm{tr}(\mathbf{Z}\mathbf{M}\mathbf{Z}^{\mathrm{T}}),$$

$$\text{s.t.} \ \ \mathbf{Z}\mathbf{Z}^{\mathrm{T}} = \mathbf{I}\,.$$

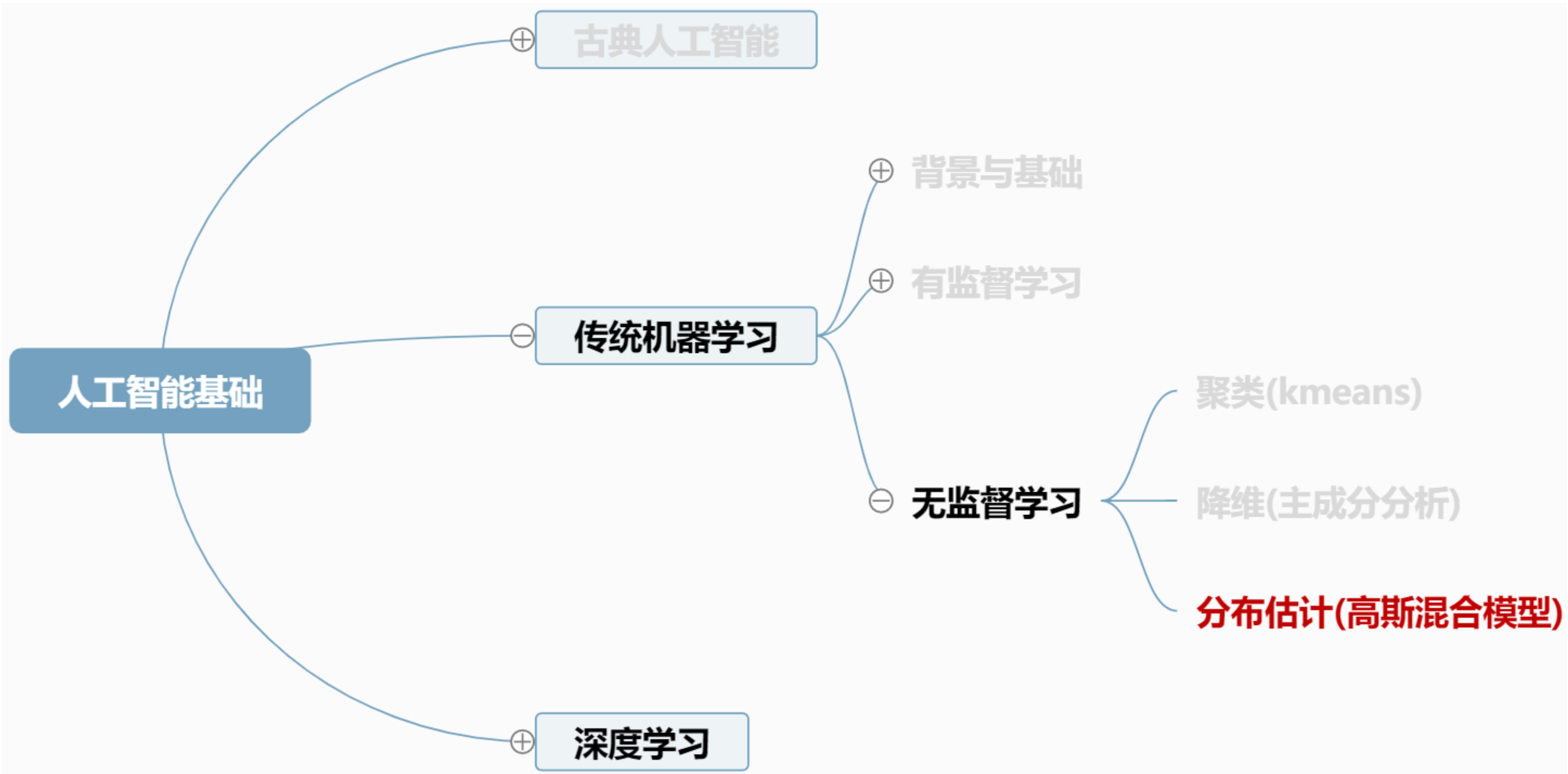
输入： 样本集 $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$;
近邻参数 k;
低维空间维数 d'.

过程：

1: **for** $i = 1, 2, \ldots, m$ **do**
2: 　确定 $\boldsymbol{x}_i$ 的 k 近邻;
3: 　从式(10.27)求得 w_{ij}, $j \in Q_i$;
4: 　对于 $j \notin Q_i$, 令 $w_{ij} = 0$;
5: **end for**
6: 从式(10.30)得到 $\mathbf{M}$;
7: 对 $\mathbf{M}$ 进行特征值分解;
8: **return** $\mathbf{M}$ 的最小 d' 个特征值对应的特征向量

输出： 样本集 D 在低维空间的投影 $Z = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_m\}$.

本节概况



人工智能基础
古典人工智能
传统机器学习
背景与基础
有监督学习
无监督学习
聚类(kmeans)
降维(主成分分析)
分布估计(高斯混合模型)
深度学习

内容概要

- 内容回顾
- **带参分布估计**
- 非参分布估计
- 总结

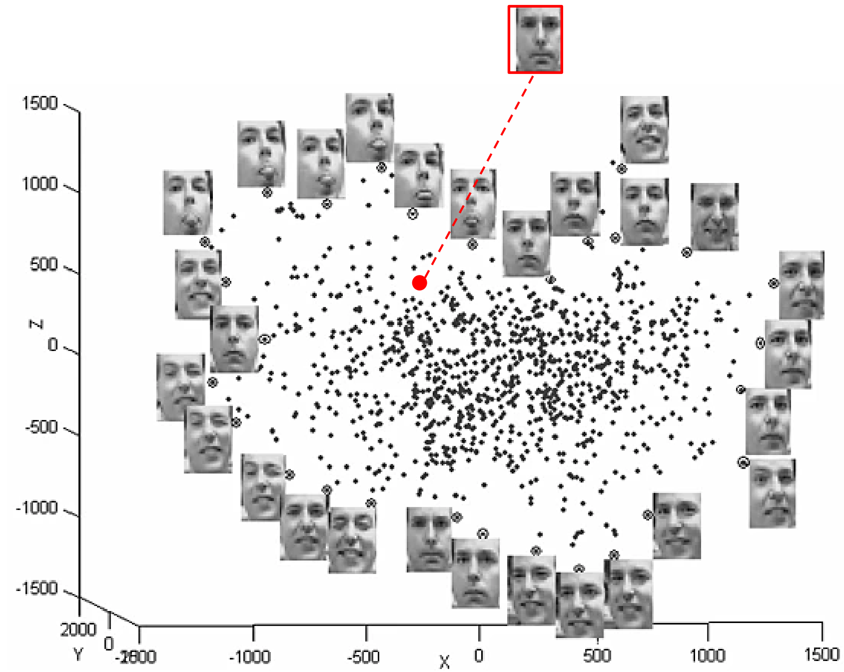
分布估计

生成式学习

- 给定数据集 $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$
- 假设样本点均采样自某未知分布：

$$\mathbf{x}_i \sim P(X)$$

- 利用 $\mathcal{X}$ 去估计 $P(X)$

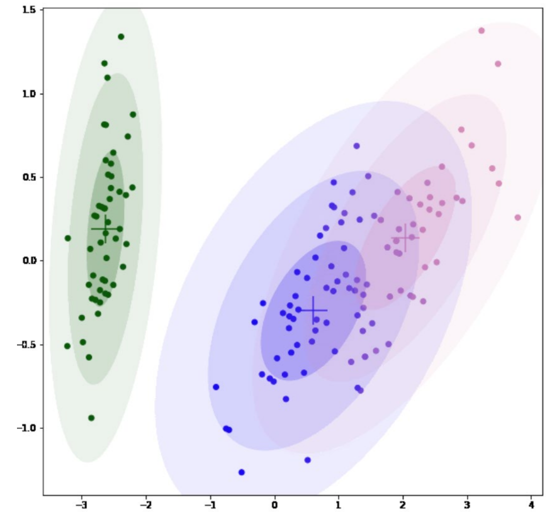


- 利用估计得到的 $P(X)$ 去生成新的样本

分布估计

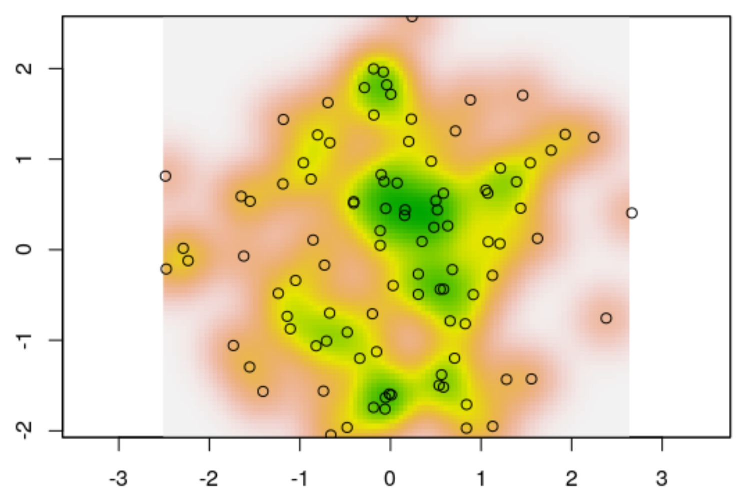
带参估计

- 假设分布 $P(X|\Theta)$ 由某参数集合 Θ 确定
- 利用数据集$\mathcal{X}$对 Θ 进行估计
- 估计完成后 $\mathcal{X}$ 不用保留



非参估计

- 分布 $P(X)$ 的形式未知
- 利用 $\mathcal{X}$ 本身进行分布表示
- 需要保留完整数据集 $\mathcal{X}$

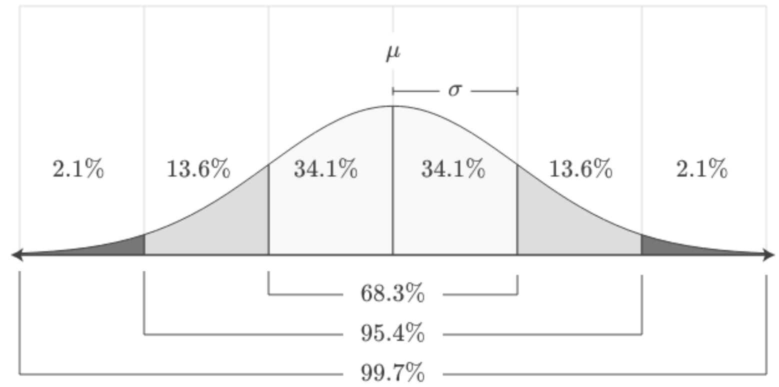


带参估计

一维高斯

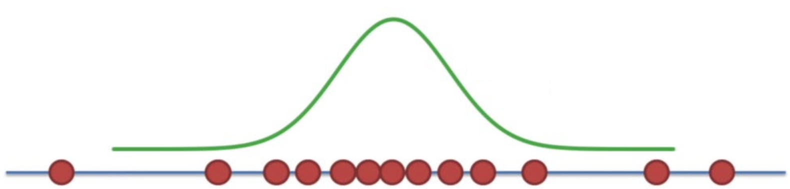
- 参数集合 $\Theta = \{\mu, \sigma^2\} \quad (\sigma^2 > 0)$
- 假设分布 $P(X|\mu, \sigma^2)$ 为

$$P(X = x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$



一维高斯参数估计

- 给定样本集$\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$
- 估计参数 μ 和 σ^2



带参估计

一维高斯参数估计

◆ 在独立采样假设下进行，最大似然法进行估计

$$P(\mathcal{X}|\mu,\sigma^2)=\prod_{i=1}^{N}P(X=x_i|\mu,\sigma^2)\ =\prod_{i=1}^{N}\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\}$$

$$L(\mu,\sigma^2)=\ln P(\mathcal{X}|\mu,\sigma^2)\ =\sum_{i=1}^{N}\ln\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\}$$

$$=\sum_{i=1}^{N}\left(-\frac{\ln 2\pi+\ln\sigma^2}{2}-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

$$\frac{\partial L}{\mu}=\sum_{i=1}^{N}\left(-\frac{2(x_i-\mu)}{2\sigma^2}\right)=0 \quad\Rightarrow\quad \mu=\frac{1}{N}\sum_{i=1}^{N}x_i$$

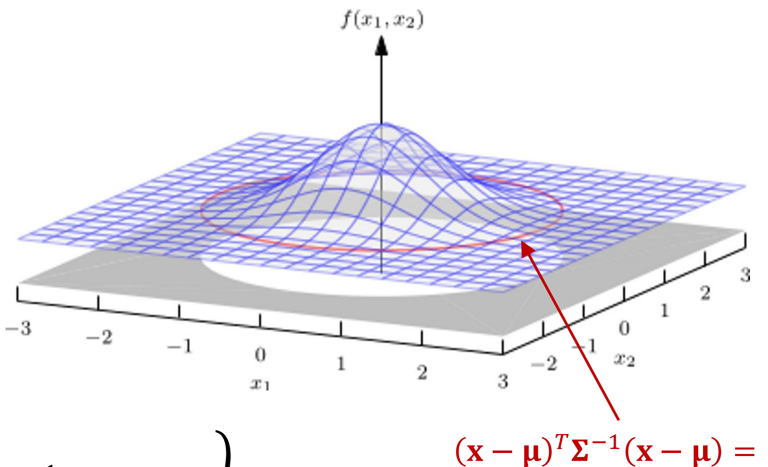
$$\frac{\partial L}{\sigma^2}=\sum_{i=1}^{N}\left(-\frac{1}{2\sigma^2}+\frac{(x_i-\mu)^2}{2\sigma^4}\right)=0 \quad\Rightarrow\quad \sigma^2=\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$$

带参估计

多维高斯

- 参数集合 $\Theta = \{\boldsymbol{\mu} \in \mathbb{R}^D, \boldsymbol{\Sigma} \in \mathbb{R}^{D\times D}\}$

  $\boldsymbol{\Sigma}$ 为对称正定矩阵（所有特征值大于零）



- 假设分布 $P(X|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 为

$$P(X = \mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}\sqrt{|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

$$\frac{x_1^2}{d_1^2} + \frac{x_2^2}{d_2^2} + \cdots + \frac{x_N^2}{d_N^2} = 1$$

$$\mathbf{x}^T\mathbf{D}^{-1}\mathbf{x} = 1$$

$\mathbf{D} = \mathbf{diag}\left(d_1^2, d_2^2, \ldots, d_N^2\right)$

旋转

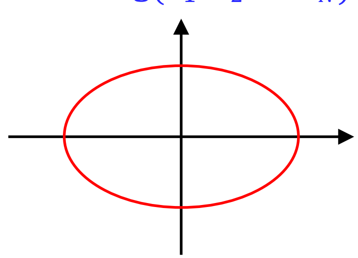
$$\mathbf{x}^T\mathbf{P}\mathbf{D}^{-1}\mathbf{P}^T\mathbf{x} = 1$$

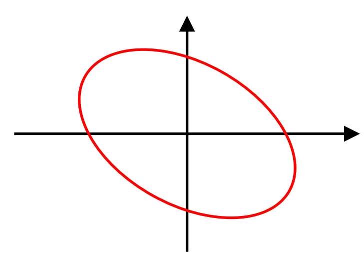
$$\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} = 1$$

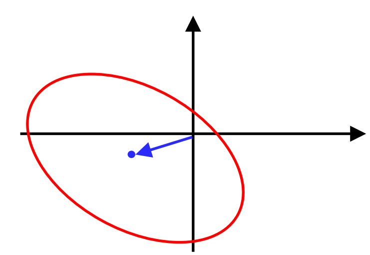
$\boldsymbol{\Sigma}^{-1} = \mathbf{P}\mathbf{D}^{-1}\mathbf{P}^T$

平移

$$(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = 1$$

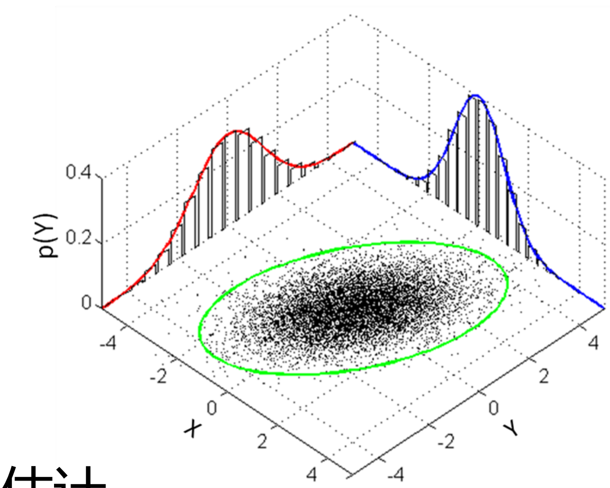






带参估计

多维高斯参数估计



- 给定样本集$\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$
- 估计参数 $\boldsymbol{\mu}$ 和 $\boldsymbol{\Sigma}$
- 在独立采样假设下进行，最大似然法进行估计

$$P(\mathcal{X}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \prod_{i=1}^{N} P(X = \mathbf{x}_i|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \prod_{i=1}^{N} \frac{1}{(2\pi)^{D/2}\sqrt{|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right\}$$

$$L(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \ln P(\mathcal{X}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{i=1}^{N} \ln \frac{1}{(2\pi)^{D/2}\sqrt{|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right\}$$

$$= \sum_{i=1}^{N} \left(-\frac{D\ln 2\pi + \ln|\boldsymbol{\Sigma}|}{2} - \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right)$$

带参估计

多高斯参数估计

- 在独立采样假设下进行，最大似然法进行估计

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{i=1}^{N}\left(-\frac{D\ln 2\pi + \ln|\boldsymbol{\Sigma}|}{2} - \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right)$$

$$\nabla_{\boldsymbol{\mu}} L = \sum_{i=1}^{N}(\ ?\ ) = \mathbf{0}$$

对任意对称矩阵 $\mathbf{A}$，有 $\nabla_{\mathbf{x}}\mathbf{x}^T\mathbf{A}\mathbf{x} = 2\mathbf{A}\mathbf{x}$

A	B	C	D
$-\boldsymbol{\Sigma}^{-1}\mathbf{x}_i$	$\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})$	$-\frac{1}{2}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})$	$\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})$

带参估计

多维高斯参数估计

- 在独立采样假设下进行，最大似然法进行估计

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{i=1}^{N} \left(-\frac{D \ln 2\pi + \ln|\boldsymbol{\Sigma}|}{2} - \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right)$$

$$\nabla_{\boldsymbol{\mu}} L = \sum_{i=1}^{N} \left(0 - \frac{1}{2} \nabla_{\boldsymbol{\mu}} (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right)$$

对任意对称矩阵 $\mathbf{A}$，有 $\nabla_{\mathbf{x}} \mathbf{x}^T \mathbf{A} \mathbf{x} = 2\mathbf{A}\mathbf{x}$

$$= \sum_{i=1}^{N} \left(-\frac{1}{2} \times 2\boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \times -1 \right)$$

$$= \sum_{i=1}^{N} \left(\boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right) = \mathbf{0} \Rightarrow \sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu}) = \mathbf{0} \Rightarrow \boldsymbol{\mu} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i$$

多维高斯参数估计

- 在独立采样假设下进行，最大似然法进行估计

$$L(\boldsymbol{\mu},\boldsymbol{\Sigma})=\sum_{i=1}^{N}\left(-\frac{D\ln 2\pi+\ln|\boldsymbol{\Sigma}|}{2}-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})\right)$$

$$=\sum_{i=1}^{N}\left(-\frac{D\ln 2\pi+\ln\frac{1}{|\boldsymbol{\Sigma}^{-1}|}}{2}-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})\right)$$

$$=L(\boldsymbol{\mu},\boldsymbol{\Sigma}^{-1})$$

带参估计

多维高斯参数估计

- 在独立采样假设下进行，最大似然法进行估计

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}) = \sum_{i=1}^{N} \left(-\frac{D \ln 2\pi + \ln \frac{1}{|\boldsymbol{\Sigma}^{-1}|}}{2} - \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right)$$

$$\nabla_{\boldsymbol{\Sigma}^{-1}} L = \sum_{i=1}^{N} \left(\frac{1}{2} \boldsymbol{\Sigma} - \frac{1}{2} \nabla_{\boldsymbol{\Sigma}^{-1}} \langle (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T, \boldsymbol{\Sigma}^{-1} \rangle \right)$$

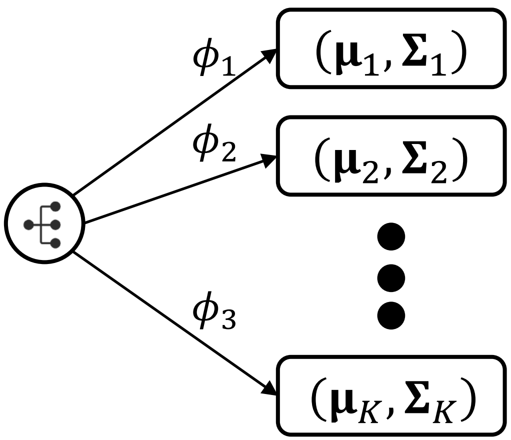
$$= \sum_{i=1}^{N} \left(\frac{1}{2} \boldsymbol{\Sigma} - \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T \right) = \mathbf{0}$$

$$\boldsymbol{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T \qquad \boldsymbol{\mu} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i$$

带参估计

混合高斯模型（GMM）

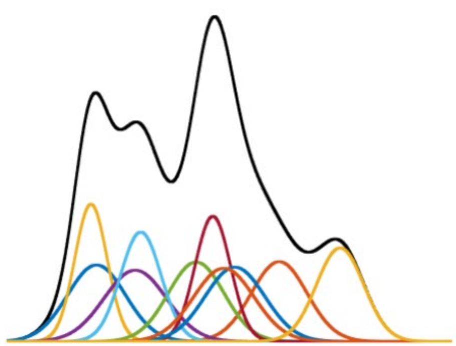
- 参数集合 $\Theta = \{(\phi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}_{k=1}^{K}$



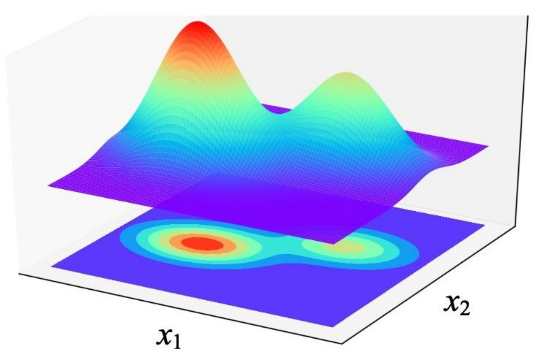
$$\mathcal{N}(X = \mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}\sqrt{|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

$$P(X = \mathbf{x}|\Theta) = \sum_{k=1}^{K} \phi_k \mathcal{N}(X = \mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

以 ϕ_k 为概率选择第 k 个高斯模型生成一个样本



一维混合高斯



二维混合高斯

带参估计

GMM参数估计

- 给定样本集$\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$
- 估计参数集合 $\Theta = \{(\phi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}_{k=1}^{K}$

$$L(\Theta) = \ln P(\mathcal{X}|\Theta) = \ln \prod_{i=1}^{N} \sum_{k=1}^{K} \phi_k \mathcal{N}(X = \mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$= \sum_{i=1}^{N} \ln \sum_{k=1}^{K} \phi_k \mathcal{N}(X = \mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$\max_{\Theta} \quad L(\Theta)$$

$$\text{s.t.} \quad \sum_{k=1}^{K} \phi_k = 1, \quad \phi_k > 0, \quad \boldsymbol{\Sigma}_k \succ \mathbf{0} \quad (k = 1, \dots, K)$$

带参估计

GMM参数估计

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

$$L(\Theta) = \sum_{i=1}^{N}\ln\sum_{k=1}^{K}\phi_k\mathcal{N}(X=\mathbf{x}_i|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$$

- $\boldsymbol{\mu}_k$无约束，直接求梯度?

$$\nabla_{\boldsymbol{\mu}_k}L = \sum_{i=1}^{N}\nabla_{\boldsymbol{\mu}_k}\ln\sum_{k=1}^{K}\phi_k\mathcal{N}(X=\mathbf{x}_i|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$$

$$= \sum_{i=1}^{N}\frac{\phi_k\nabla_{\boldsymbol{\mu}_k}\mathcal{N}(X=\mathbf{x}_i|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_{k=1}^{K}\phi_k\mathcal{N}(X=\mathbf{x}_i|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}$$

$$= \sum_{i=1}^{N}\frac{\frac{\phi_k}{(2\pi)^{D/2}\sqrt{|\boldsymbol{\Sigma}_k|}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T{\boldsymbol{\Sigma}_k}^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right\}\left(\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_k)\right)}{\sum_{k=1}^{K}\phi_k\mathcal{N}(X=\mathbf{x}_i|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)} = \mathbf{0}$$

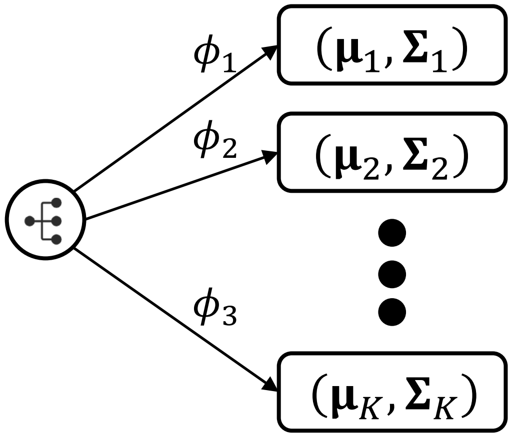
$$\boldsymbol{\mu}_k = \textbf{?}$$

带参估计

GMM参数估计

- 从生成的角度去分析

  以 ϕ_k 为概率选择第 k 个高斯模型生成一个样本



- 假设我们知道当时选择了哪个高斯模型

  记生成第 i 个样本 $\mathbf{x}_i$ 的高斯模型为 z_i

$$P(Z = z_i, X = \mathbf{x}_i | \Theta) = \phi_{z_i} \mathcal{N}\left(X = \mathbf{x}_i \middle| \boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}\right)$$

$$L(\Theta, \mathcal{Z}, \mathcal{X}) = \sum_{i=1}^{N} \left(\ln \phi_{z_i} + \ln \mathcal{N}\left(X = \mathbf{x}_i \middle| \boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}\right)\right)$$ 可解

- 然而 $\mathcal{Z} = \{\mathrm{z}_1, \mathrm{z}_2, \dots, \mathrm{z}_N\}$ 对我们来说并非已知

带参估计

期望最大化算法（EM）

◆ 三个基本原理

- 分类原理 $P(X|\Theta) = \sum_{Z} P(Z, X|\Theta)$

例如

$$X = \begin{cases} 1, \text{一位同学成绩为优秀} \\ 0, \text{一位同学成绩非优秀} \end{cases} \qquad Z = \begin{cases} 1, \text{该同学为男生} \\ 0, \text{该同学为女生} \end{cases}$$

$$P(X = 1) = P(Z = 0, X = 1) + P(Z = 1, X = 1)$$

- 贝叶斯公式 $P(X|Z, \Theta) = \dfrac{P(Z, X|\Theta)}{P(Z|\Theta)}$

期望最大化算法（EM）

- 三个基本原理
  - Jessen’s 不等式

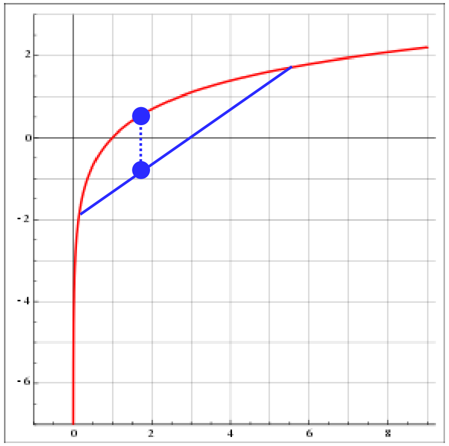
对于任意的凹函数 f 以及随机变量 Z，则有

$$\mathbb{E}_Z f(Z) = \sum_Z P(Z) f(Z) \leq f\left(\sum_Z P(Z) Z\right) = f(\mathbb{E}_Z Z)$$

当 f 为严格凹函数时，其中不等式中的等号成立当且仅当 Z 为常数

$\ln x$, $-e^x$ 都为常见的严格凹函数

丹麦业余数学家
Johan Jensen (1859-1925)



期望最大化算法（EM）

- EM 算法原理

$$P(X|\Theta) = \sum_Z P(Z,X|\Theta)$$

$$P(X|Z,\Theta) = \frac{P(Z,X|\Theta)}{P(Z|\Theta)}$$

$$\sum_Z P(Z)\ln(Z) \le \ln\left(\sum_Z P(Z)Z\right)$$

$$\ln P(X|\Theta) = \ln \sum_Z P(Z,X|\Theta)$$

$$= \ln \sum_Z \frac{P(Z,X|\Theta)}{P(Z|X,\Theta)} P(Z|X,\Theta)$$

$$\frac{P(Z,X|\Theta)}{P(Z|X,\Theta)} = P(X|\Theta)$$

$$= \sum_Z P(Z|X,\Theta) \ln \frac{P(Z,X|\Theta)}{P(Z|X,\Theta)}$$

带参估计

期望最大化算法（EM）

- EM 算法原理

$$P(X|\Theta) = \sum_Z P(Z, X|\Theta)$$

$$P(X|Z,\Theta) = \frac{P(Z,X|\Theta)}{P(Z|\Theta)}$$

$$\sum_Z P(Z)\ln(Z) \le \ln\left(\sum_Z P(Z)Z\right)$$

$$\ln P(X|\Theta) = \sum_Z P(Z|X,\Theta)\ln\frac{P(Z,X|\Theta)}{P(Z|X,\Theta)}$$

$$\ln P\left(X\middle|\widehat{\Theta}\right) = \sum_Z P\left(Z\middle|X,\widehat{\Theta}\right)\ln\frac{P\left(Z,X\middle|\widehat{\Theta}\right)}{P\left(Z\middle|X,\widehat{\Theta}\right)}$$

$$\boxed{\le \max_\Theta \sum_Z P\left(Z\middle|X,\widehat{\Theta}\right)\ln\frac{P(Z,X|\Theta)}{P\left(Z\middle|X,\widehat{\Theta}\right)}}$$

$$= \sum_Z P\left(Z\middle|X,\widehat{\Theta}\right)\ln\frac{P(Z,X|\Theta^\star)}{P\left(Z\middle|X,\widehat{\Theta}\right)}$$

$$\Theta^\star = \arg\max_\Theta \sum_Z P\left(Z\middle|X,\widehat{\Theta}\right)\ln\frac{P(Z,X|\Theta)}{P\left(Z\middle|X,\widehat{\Theta}\right)}$$

$$\le \ln\sum_Z P\left(Z\middle|X,\widehat{\Theta}\right)\frac{P(Z,X|\Theta^\star)}{P\left(Z\middle|X,\widehat{\Theta}\right)}$$

$$= \ln\sum_Z P(Z,X|\Theta^\star) = \ln P(X|\Theta^\star)$$

带参估计

期望最大化算法（EM）

- EM 算法原理

Θ*相同

$$\max_{\Theta}\sum_{Z}P\left(Z\middle|X,\widehat{\Theta}\right)\ln\frac{P(Z,X|\Theta)}{P\left(Z\middle|X,\widehat{\Theta}\right)} \Longleftrightarrow \max_{\Theta}\sum_{Z}P\left(Z\middle|X,\widehat{\Theta}\right)\ln P(Z,X|\Theta)$$

$$= \mathbb{E}_{P\left(Z\middle|X,\widehat{\Theta}\right)}\ln P(Z,X|\Theta)$$

- 算法流程
  - 给定样本集合$\mathcal{X}=\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N\}$
  - 初始化概率模型参数 $\widehat{\Theta}$
  - 基于 $\mathcal{X}$ 和 $\widehat{\Theta}$ 估计条件期望（E 步）

$$\mathbb{E}_{P\left(Z\middle|X,\widehat{\Theta}\right)}\ln P(Z,X|\Theta)$$

不收敛则循环

  - 求解条件期望最大化（M 步）

$$\widehat{\Theta}\longleftarrow\arg\max_{\Theta}\mathbb{E}_{P\left(Z\middle|X,\widehat{\Theta}\right)}\ln P(Z,X|\Theta)$$

带参估计

EM算法估计GMM参数

- 基于 $\mathcal{X}$ 和 $\widehat{\Theta}$ 估计条件期望（E 步）

$$\mathbb{E}_{P\left(Z\middle|X,\widehat{\Theta}\right)}\ln P(Z,X|\Theta)$$

$\Theta=\{(\phi_k,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\}_{k=1}^{K}$ $\quad$ Z 表示 X 来自的高斯模型编号

$$P\left(Z=k\middle|X=\mathbf{x},\widehat{\Theta}\right)=\frac{P\left(Z=k,X=\mathbf{x}\middle|\widehat{\Theta}\right)}{P\left(X=\mathbf{x}\middle|\widehat{\Theta}\right)}=\frac{\hat{\phi}_k\mathcal{N}\left(X=\mathbf{x}\middle|\widehat{\boldsymbol{\mu}}_k,\widehat{\boldsymbol{\Sigma}}_k\right)}{\sum_{k=1}^{K}\hat{\phi}_k\mathcal{N}\left(X=\mathbf{x}\middle|\widehat{\boldsymbol{\mu}}_k,\widehat{\boldsymbol{\Sigma}}_k\right)}$$

$$P\left(Z=k\middle|X=\mathbf{x}_i,\widehat{\Theta}\right)=\gamma_{ik}$$

$$\begin{aligned}\mathbb{E}_{P\left(Z\middle|X,\widehat{\Theta}\right)}\ln P(Z,X|\Theta)&=\mathbb{E}_{P\left(Z\middle|X,\widehat{\Theta}\right)}\sum_{i=1}^{N}\ln P(Z,X=\mathbf{x}_i|\Theta)\\&=\sum_{k=1}^{K}\sum_{i=1}^{N}P\left(Z=k\middle|X=\mathbf{x}_i,\widehat{\Theta}\right)\ln P(Z=k,X=\mathbf{x}_i|\Theta)\\&=\sum_{k=1}^{K}\sum_{i=1}^{N}\gamma_{ik}\ln P(Z=k,X=\mathbf{x}_i|\Theta)\end{aligned}$$

带参估计

EM算法估计GMM参数

- 求解条件期望最大化（M 步）

$$\max_{\Theta} \mathbb{E}_{P\left(Z \middle| X, \widehat{\Theta}\right)} \ln P(Z, X|\Theta)$$

$$\mathbb{E}_{P\left(Z \middle| X, \widehat{\Theta}\right)} \ln P(Z, X|\Theta) = \sum_{k=1}^{K} \sum_{i=1}^{N} \gamma_{ik} \ln P(Z = k, X = \mathbf{x}_i|\Theta)$$

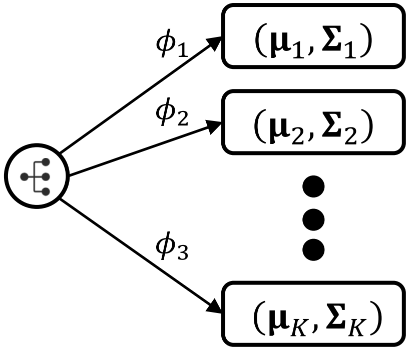
$$= \sum_{k=1}^{K} \sum_{i=1}^{N} \gamma_{ik} (\ln \phi_k + \ln \mathcal{N}(X = \mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))$$

- 参考多维高斯计算 $\boldsymbol{\mu}$ 和 $\boldsymbol{\Sigma}$ 的过程可解出

$$\boldsymbol{\mu}_k{}^{\star} = \frac{\sum_{i=1}^{N} \gamma_{ik} \mathbf{x}_i}{\sum_{i=1}^{N} \gamma_{ik}} \qquad \boldsymbol{\Sigma}_k{}^{\star} = \frac{\sum_{i=1}^{N} \gamma_{ik} (\mathbf{x}_i - \boldsymbol{\mu}_k{}^{\star})(\mathbf{x}_i - \boldsymbol{\mu}_k{}^{\star})^T}{\sum_{i=1}^{N} \gamma_{ik}}$$

带参估计

EM算法估计GMM参数



- 求解条件期望最大化（M 步）
  - 求解 ϕ_k

$$\mathbb{E}_{P(Z|X,\widehat{\Theta})}\ln P(Z,X|\Theta)=\sum_{k=1}^{K}\sum_{i=1}^{N}\gamma_{ik}(\ln\phi_k+\ln\mathcal{N}(X=\mathbf{x}_i|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k))=f(\phi_1,\dots,\phi_K)$$

$$\boxed{\max_{\phi_1,\dots,\phi_K} f(\phi_1,\dots,\phi_K) \quad \text{s.t.} \ \sum_{k=1}^{K}\phi_k=1,\phi_k\geq 0 \ \ (k=1,\dots,K)}$$

构造拉格朗日函数： $L(\phi_1,\dots,\phi_K,\lambda)=f(\phi_1,\dots,\phi_K)+\lambda\left(\sum_{k=1}^{K}\phi_k-1\right)$

最优性条件：

$$\begin{cases}\dfrac{\partial L}{\partial\phi_k}=\dfrac{\sum_{i=1}^{N}\gamma_{ik}}{\phi_k}+\lambda=0 \quad (k=1,\dots,K)\\ \sum_{k=1}^{K}\phi_k=1\end{cases} \Rightarrow \quad {\phi_k}^{\star}=\frac{\sum_{i=1}^{N}\gamma_{ik}}{\sum_{i=1}^{N}\sum_{k=1}^{K}\gamma_{ik}} \quad (k=1,\dots,K)$$

带参估计

EM算法估计GMM参数

- 算法流程
  - 给定样本集合$\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$
  - 初始化概率模型参数 $\widehat{\Theta} = \left\{\left(\hat{\phi}_k, \widehat{\boldsymbol{\mu}}_k, \widehat{\boldsymbol{\Sigma}}_k\right)\right\}_{k=1}^{K}$
  - 基于 $\mathcal{X}$ 和 $\widehat{\Theta}$ 估计条件期望（E 步）

$$\gamma_{ik} \leftarrow \frac{\hat{\phi}_k \mathcal{N}\left(X = \mathbf{x}_i \middle| \widehat{\boldsymbol{\mu}}_k, \widehat{\boldsymbol{\Sigma}}_k\right)}{\sum_{k=1}^{K} \hat{\phi}_k \mathcal{N}\left(X = \mathbf{x}_i \middle| \widehat{\boldsymbol{\mu}}_k, \widehat{\boldsymbol{\Sigma}}_k\right)}$$

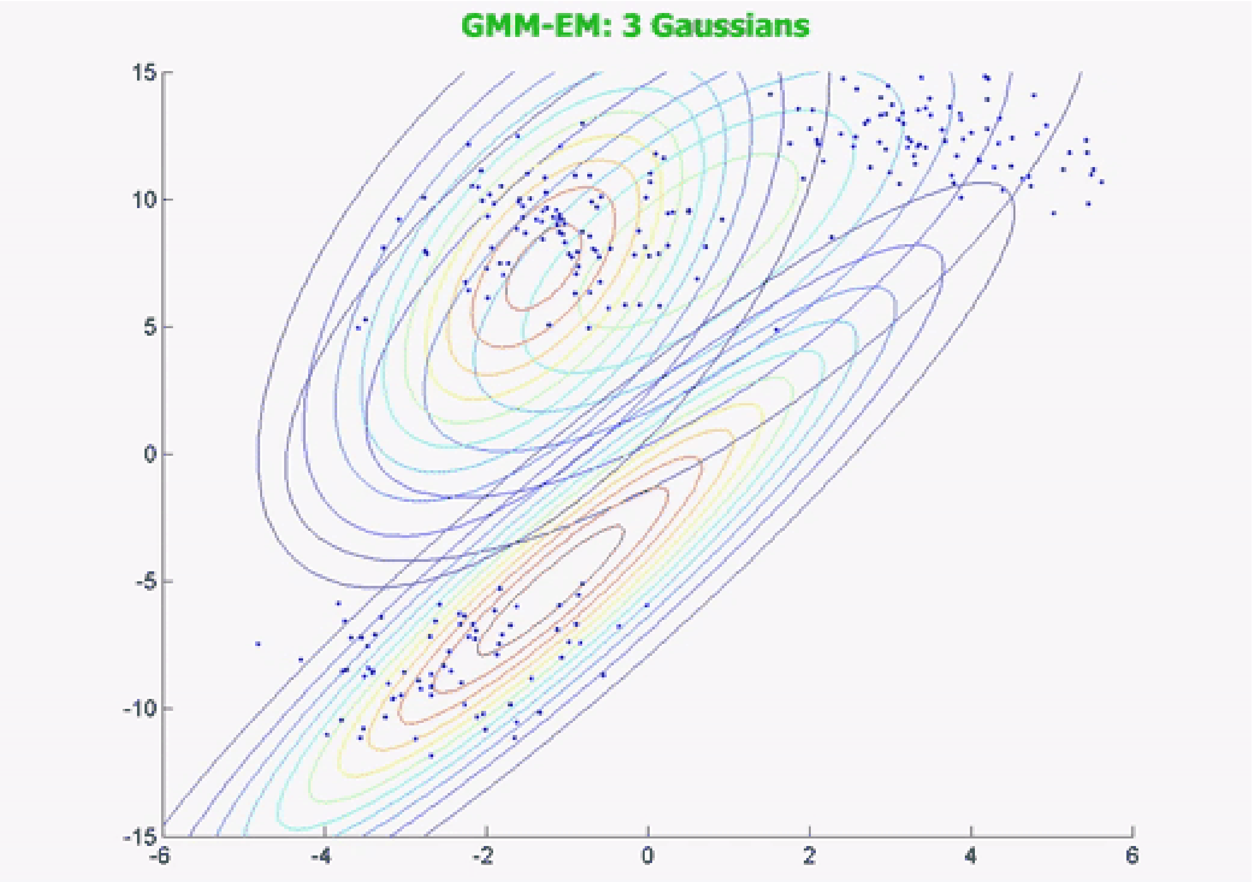
  - 求解条件期望最大化（M 步）

$$\widehat{\boldsymbol{\mu}}_k \leftarrow \frac{\sum_{i=1}^{N} \gamma_{ik}\, \mathbf{x}_i}{\sum_{i=1}^{N} \gamma_{ik}} \qquad \widehat{\boldsymbol{\Sigma}}_k \leftarrow \frac{\sum_{i=1}^{N} \gamma_{ik}\, (\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k)^T}{\sum_{i=1}^{N} \gamma_{ik}} \qquad \hat{\phi}_k \leftarrow \frac{\sum_{i=1}^{N} \gamma_{ik}}{\sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_{ik}}$$

  - 重复E步与M步直到收敛

带参估计

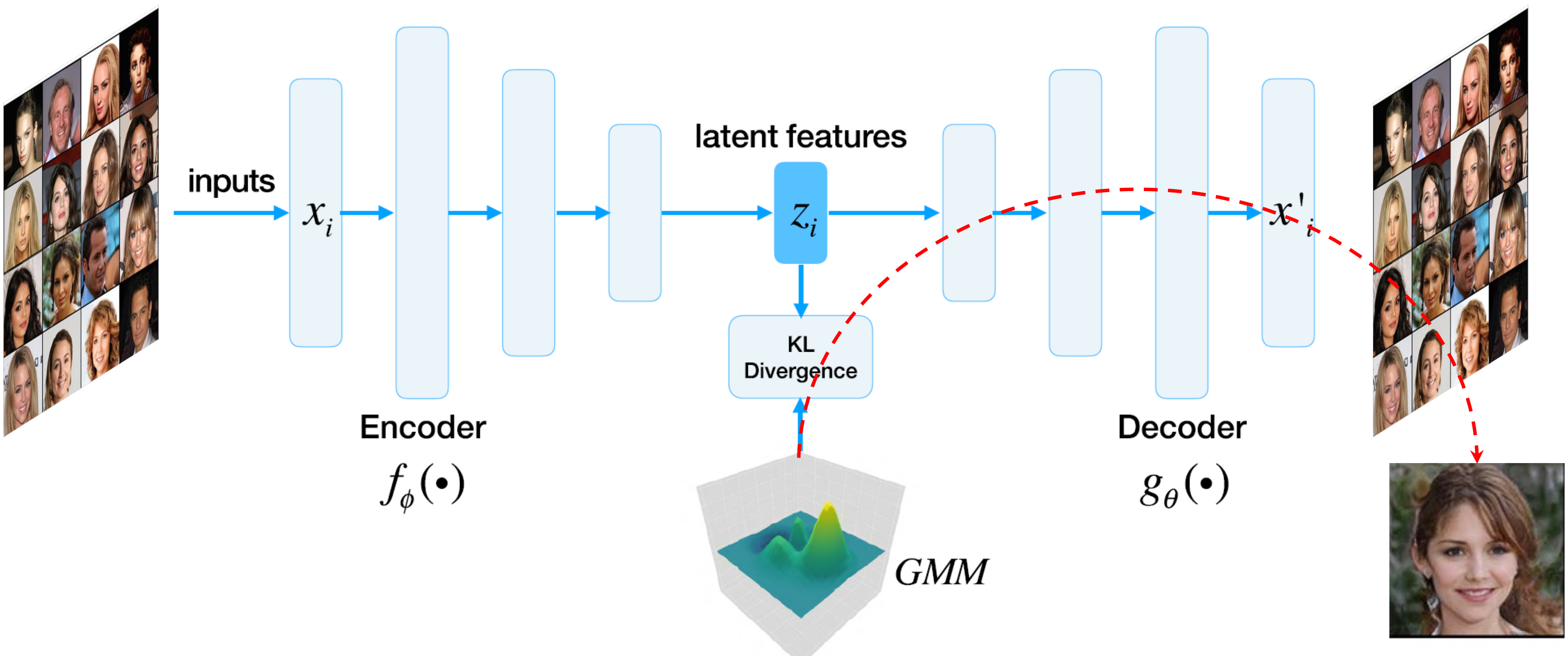
EM算法估计GMM参数



带参估计

高斯混合模型应用

- 图像生成（VAE+GMM）



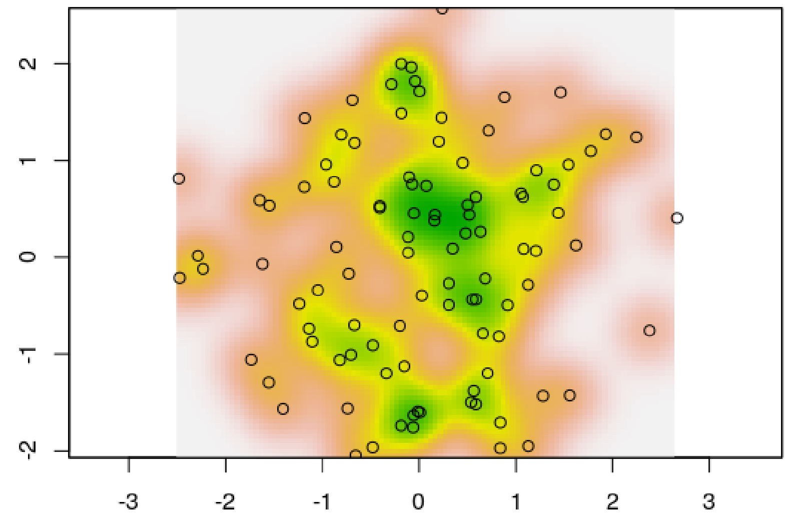
内容概要

- 内容回顾
- 带参分布估计
- **非参分布估计**
- 总结

非参估计

特点

- 分布 $P(X)$ 的形式未知
- 利用 $\mathcal{X}$ 本身进行分布表示
- 需要保留完整数据集 $\mathcal{X}$



基于核密度函数的非参估计

- 给定样本集$\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$

$$f(\mathbf{x}) = \frac{1}{Nh}\sum_{i=1}^{N} Ker\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)$$

- 其中 $h > 0$ 称为带宽， Ker称为**核密度函数**应满足

$$\int Ker(\mathbf{x})\ \mathrm{d}\mathbf{x} = 1$$

注意这里的核密度函数与之前讲到的核函数不同

非参估计

基于核密度函数的非参估计

- 核密度的归一性

$$\int f(\mathbf{x})\ \mathrm{d}\mathbf{x} = \int \frac{1}{Nh}\sum_{i=1}^{N} Ker\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)\ \mathrm{d}\mathbf{x}$$

$$= \frac{1}{Nh}\sum_{i=1}^{N} \int Ker\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)\ \mathrm{d}\mathbf{x}$$

$$= \frac{1}{Nh}\sum_{i=1}^{N} h = 1$$

$$\int Ker(\mathbf{x})\ \mathrm{d}\mathbf{x} = 1$$

- 高斯核密度

$$Ker(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}} \exp\left\{-\frac{1}{2}\mathbf{x}^T\mathbf{x}\right\}$$

$$f(\mathbf{x}) = \frac{1}{Nh}\sum_{i=1}^{N} Ker\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{(2\pi)^{D/2}h} \exp\left\{-\frac{1}{2}\frac{(\mathbf{x}-\mathbf{x}_i)^T(\mathbf{x}-\mathbf{x}_i)}{h^2}\right\}$$

以每个样本点 $\mathbf{x}_i$ 为质心，以 h^2 为方差的 N 高斯混合模型

内容概要

- 内容回顾
- 带参分布估计
- 非参分布估计
- **总结**

总结

- 带参估计
  - 多维高斯形式与参数估计
  - 混合高斯模型
  - EM 算法原理
  - 基于 EM 算法的混合高斯模型
- 非参估计
  - 核密度估计法
  - 高斯核密度与混合高斯模型之间的关系


谢 谢!
问 题?