

# 概率与学习

高 阳，李文斌

<http://cs.nju.edu.cn/rl>

2023年10月24日

# 大纲

## 相关概念

高斯混合模型

最大似然估计

期望最大化算法

# 相关概念

- 符号和术语（实数）

标量（scalar）：一个实数  $x \in \mathbb{R}$

向量（vector）：一个实数序列构成一个向量（粗斜体）  $\mathbf{x} \in \mathbb{R}^d$

$$\mathbf{x} = (x_1, x_2, \dots, x_d)^T$$

矩阵（matrix）：一个实数构成的矩形数组（大写粗体）  $\mathbf{X} \in \mathbb{R}^{m \times n}$

$$\mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix}$$

# 相关概念

- 符号和术语（实数）

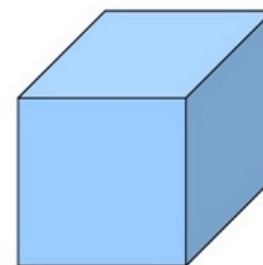
张量（tensor）：一个泛化的实数构成的 $n$ -维数组



1d-tensor



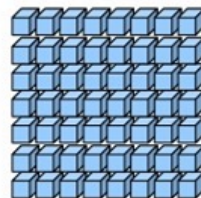
2d-tensor



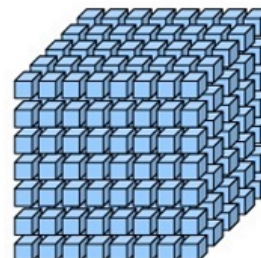
3d-tensor



4d-tensor



5d-tensor



6d-tensor

例子：  $\mathbf{X} \in \mathbb{R}^{H \times W \times D}$  是一个3阶张量

# 相关概念

- 符号和术语

$\mathbf{A}$	Matrix
$\mathbf{A}_{ij}$	Matrix indexed for some purpose
$\mathbf{A}_i$	Matrix indexed for some purpose
$\mathbf{A}^{ij}$	Matrix indexed for some purpose
$\mathbf{A}^n$	Matrix indexed for some purpose <b>or</b> The n.th power of a square matrix
$\mathbf{A}^{-1}$	The inverse matrix of the matrix $\mathbf{A}$
$\mathbf{A}^+$	The pseudo inverse matrix of the matrix $\mathbf{A}$ (see Sec. 3.6)
$\mathbf{A}^{1/2}$	The square root of a matrix (if unique), not elementwise
$(\mathbf{A})_{ij}$	The $(i, j)$ .th entry of the matrix $\mathbf{A}$
$A_{ij}$	The $(i, j)$ .th entry of the matrix $\mathbf{A}$
$[\mathbf{A}]_{ij}$	The $ij$ -submatrix, i.e. $\mathbf{A}$ with i.th row and j.th column deleted
$\mathbf{a}$	Vector (column-vector)
$\mathbf{a}_i$	Vector indexed for some purpose
$a_i$	The i.th element of the vector $\mathbf{a}$
$a$	Scalar

# 相关概念

## • 符号和术语

$\det(\mathbf{A})$	Determinant of $\mathbf{A}$
$\text{Tr}(\mathbf{A})$	Trace of the matrix $\mathbf{A}$
$\text{diag}(\mathbf{A})$	Diagonal matrix of the matrix $\mathbf{A}$ , i.e. $(\text{diag}(\mathbf{A}))_{ij} = \delta_{ij} A_{ij}$
$\text{eig}(\mathbf{A})$	Eigenvalues of the matrix $\mathbf{A}$
$\text{vec}(\mathbf{A})$	The vector-version of the matrix $\mathbf{A}$ (see Sec. 10.2.2)
$\sup$	Supremum of a set
$\ \mathbf{A}\ $	Matrix norm (subscript if any denotes what norm)
$\mathbf{A}^T$	Transposed matrix
$\mathbf{A}^{-T}$	The inverse of the transposed and vice versa, $\mathbf{A}^{-T} = (\mathbf{A}^{-1})^T = (\mathbf{A}^T)^{-1}$
$\mathbf{A}^*$	Complex conjugated matrix
$\mathbf{A}^H$	Transposed and complex conjugated matrix (Hermitian)
$\mathbf{A} \circ \mathbf{B}$	Hadamard (elementwise) product
$\mathbf{A} \otimes \mathbf{B}$	Kronecker product
$\mathbf{0}$	The null matrix. Zero in all entries.
$\mathbf{I}$	The identity matrix
$\mathbf{J}^{ij}$	The single-entry matrix, 1 at $(i, j)$ and zero elsewhere
$\Sigma$	A positive definite matrix
$\Lambda$	A diagonal matrix

# 相关概念

- 一个带约束的数学优化问题

$$\begin{array}{ll}\text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq b_i, \quad i = 1, \dots, m.\end{array}$$

- 优化变量:  $\mathbf{x} = (x_1, \dots, x_n)^T \in \mathbb{R}^n$
- 目标函数:  $f_0: \mathbb{R}^n \rightarrow \mathbb{R}$
- 约束函数:  $f_i: \mathbb{R}^n \rightarrow \mathbb{R}, i = 1, \dots, m$
- 最优解 :  $\mathbf{x}^* = (x_1, \dots, x_n)^T \in \mathbb{R}^n$

# 相关概念

- 一个不带约束的数学优化问题

## □ 最小二乘 (least-squares) 问题

$$\text{minimize } f_0(x) = \|Ax - b\|_2^2 = \sum_{i=1}^k (a_i^T x - b_i)^2$$

- 优化变量:  $x = (x_1, \dots, x_n)^T \in \mathbb{R}^n$
- 系数矩阵:  $A \in \mathbb{R}^{k \times n}$
- $A$  的第  $i$  行:  $a_i^T \in \mathbb{R}^{1 \times n}$
- 最优解 :  $x^* = (x_1, \dots, x_n)^T \in \mathbb{R}^n$

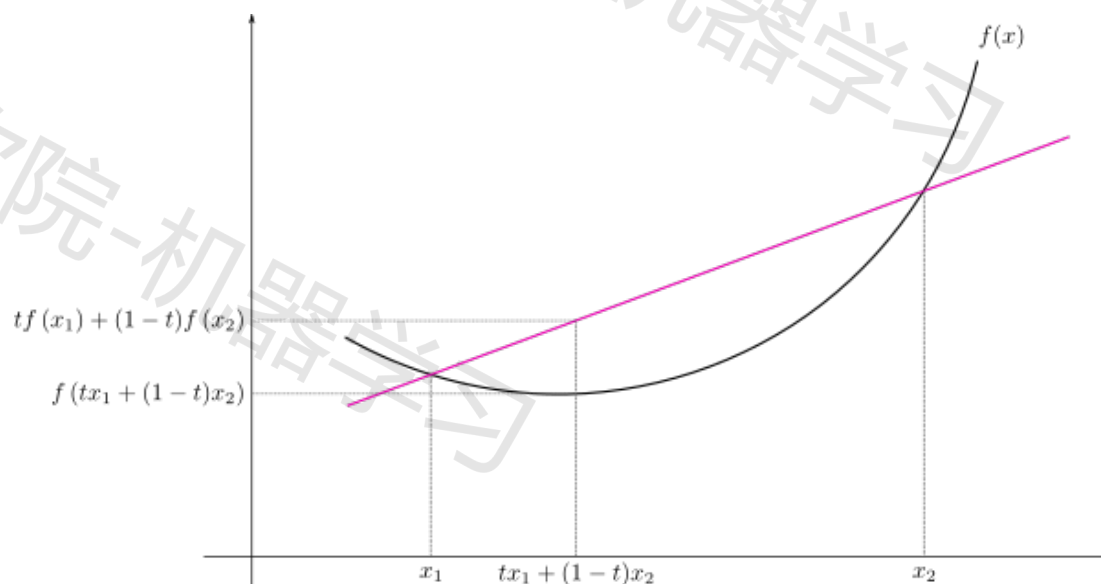


# 相关概念

- 凸函数

$X$  是一个凸集合,  $f: X \rightarrow \mathbb{R}$  表示定义在  $X$  上的一个函数

$$\forall x_1, x_2 \in X, \forall t \in [0, 1]: f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$

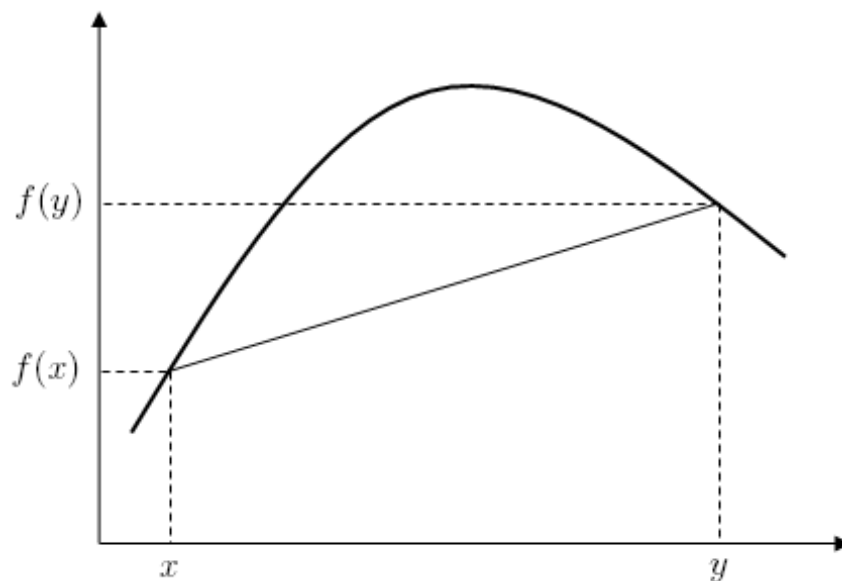


# 相关概念

- 凹函数

$X$ 是一个凸集合,  $f: X \rightarrow \mathbb{R}$ 表示定义在  $X$  上的一个函数

$$\forall x_1, x_2 \in X, \forall t \in [0, 1]: f(tx_1 + (1-t)x_2) \geq tf(x_1) + (1-t)f(x_2)$$



# 相关概念

- 判断函数的凹凸

$$x \in \text{dom } f, x \in \mathbb{R}$$

如果二阶导数  $f''(x) \geq 0$

$f(x)$ 是凸函数

$$\mathbf{x} \in \text{dom } f, \mathbf{x} \in \mathbb{R}^d$$

如果Hessian矩阵是半正定的，即  $\nabla^2 f(\mathbf{x}) \succcurlyeq 0$   $f(\mathbf{x})$ 是凸函数

# 相关概念

- 判断函数的凹凸

$$x \in \text{dom } f, x \in \mathbb{R}$$

如果二阶导数  $f''(x) \leq 0$

$f(x)$ 是凹函数

$$\mathbf{x} \in \text{dom } f, \mathbf{x} \in \mathbb{R}^d$$

如果Hessian矩阵是非半正定的, 即  $\nabla^2 f(\mathbf{x}) \not\leq 0$   $f(\mathbf{x})$ 是凹函数

# 相关概念

## • 常见例子

- *Exponential.*  $e^{ax}$  is convex on  $\mathbf{R}$ , for any  $a \in \mathbf{R}$ .
- *Powers.*  $x^a$  is convex on  $\mathbf{R}_{++}$  when  $a \geq 1$  or  $a \leq 0$ , and concave for  $0 \leq a \leq 1$ .
- *Powers of absolute value.*  $|x|^p$ , for  $p \geq 1$ , is convex on  $\mathbf{R}$ .
- *Logarithm.*  $\log x$  is concave on  $\mathbf{R}_{++}$ .
- *Norms.* Every norm on  $\mathbf{R}^n$  is convex.
- *Max function.*  $f(x) = \max\{x_1, \dots, x_n\}$  is convex on  $\mathbf{R}^n$ .
- *Geometric mean.* The geometric mean  $f(x) = (\prod_{i=1}^n x_i)^{1/n}$  is concave on  $\text{dom } f = \mathbf{R}_{++}^n$ .
- *Log-determinant.* The function  $f(X) = \log \det X$  is concave on  $\text{dom } f = \mathbf{S}_{++}^n$ .
  - $R$  实数;  $R_+$  非负实数;  $R_{++}$  正实数;  $R^n$  表示  $n$  维向量空间
  - $S_{++}^n$  是  $n \times n$  对称正定矩阵构成的空间

# 相关概念

- 推荐阅读

Boyd, Stephen, Stephen P. Boyd, and Lieven Vandenberghe.  
*Convex optimization*. Cambridge university press, 2004.

Kaare Brandt Petersen Michael Syskind Pedersen.  
*The Matrix Cookbook*. Technical University of Denmark. 2012

# 相关概念

## • 随机变量的期望

**定理1**：令 $X$ 表示一个随机变量，存在某个函数 $g$ 使得 $Y = g(X)$

1. 假设 $X$ 是连续的，pdf为 $f_X(x)$ 。如果 $\int_{-\infty}^{\infty} |g(x)|f_X(x)dx < \infty$ ，那么 $Y$ 的期望存在且为

$$E(Y) = \int_{-\infty}^{\infty} g(x)f_X(x)dx$$

2. 假设 $X$ 是离散的，pmf为 $p_X(x)$ 。假设 $X$ 的支撑用 $S_X$ 表示，如果 $\sum_{x \in S_X} |g(x)|p_X(x) < \infty$ ，那么 $Y$ 的期望存在且为

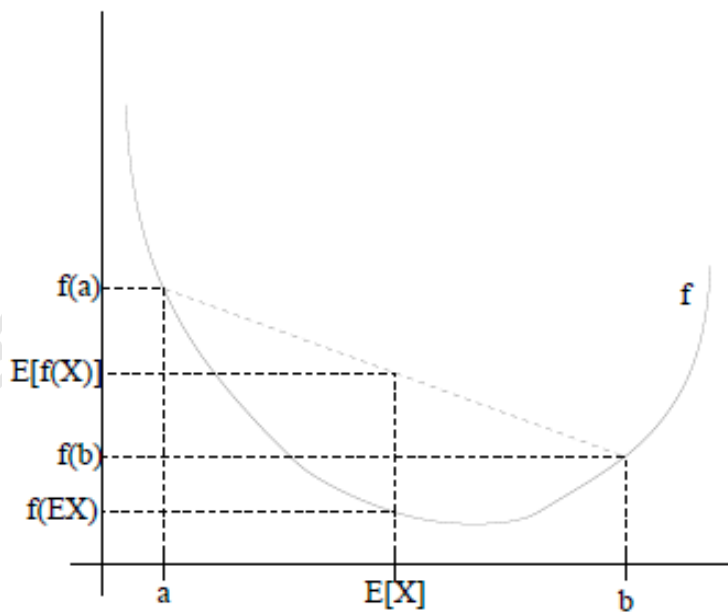
$$E(Y) = \sum_{x \in S_X} g(x)p_X(x)$$

- 概率密度函数：Probability density function (PDF)
- 概率质量函数：Probability mass function (PMF)

# 相关概念

- Jensen's inequality

- ✓ 如果 $X$ 是随机变量，并且 $f(X)$ 是凸函数，则 $E[f(X)] \geq f(E[X])$
- ✓ 如果 $X$ 是随机变量，并且 $f(X)$ 是凹函数，则 $E[f(X)] \leq f(E[X])$



$X$ 有0.5的概率是 $a$ ，有0.5的概率是 $b$ ，那么 $E[X] = \frac{a+b}{2}$



# 相关概念

- 高斯分布/正态分布  
(Gaussian distribution /Normal distribution)

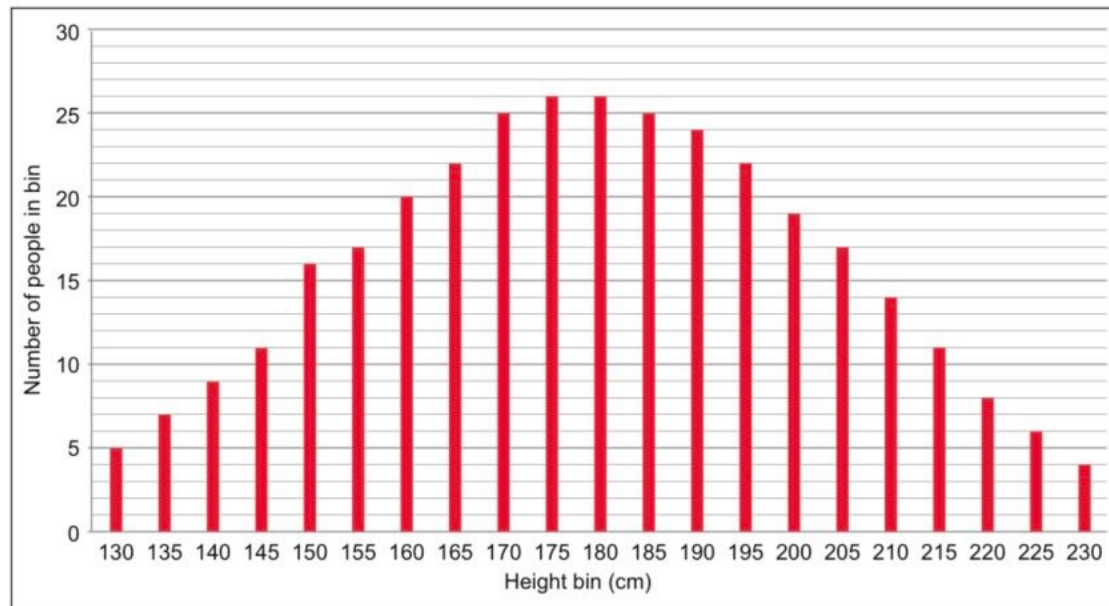


Figure 2.9 Histogram of a normal distribution, in this case the height of 334 fictitious people. The modal (most frequently occurring) bin is centered at 180 cm.

由334个人的身高数据构成的正态分布直方图

# 相关概念

- 高斯分布/正态分布

(**Gaussian distribution /Normal distribution**)

- ✓ 正态分布是在统计以及许多统计测试中最广泛应用的一类分布
- ✓ 正态分布也是机器学习，统计模式识别和计算机视觉中使用最广泛的概率分布

# 相关概念

- 单变量高斯分布/正态分布

(Univariate Gaussian distribution / Normal distribution)

✓ 若一维随机变量 $X$ 服从高斯分布，则记为

$$X \sim N(\mu, \sigma^2)$$

标准正态分布  
 $\mu = 0, \sigma = 1$

✓ 概率密度函数(probability density function, PDF)

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

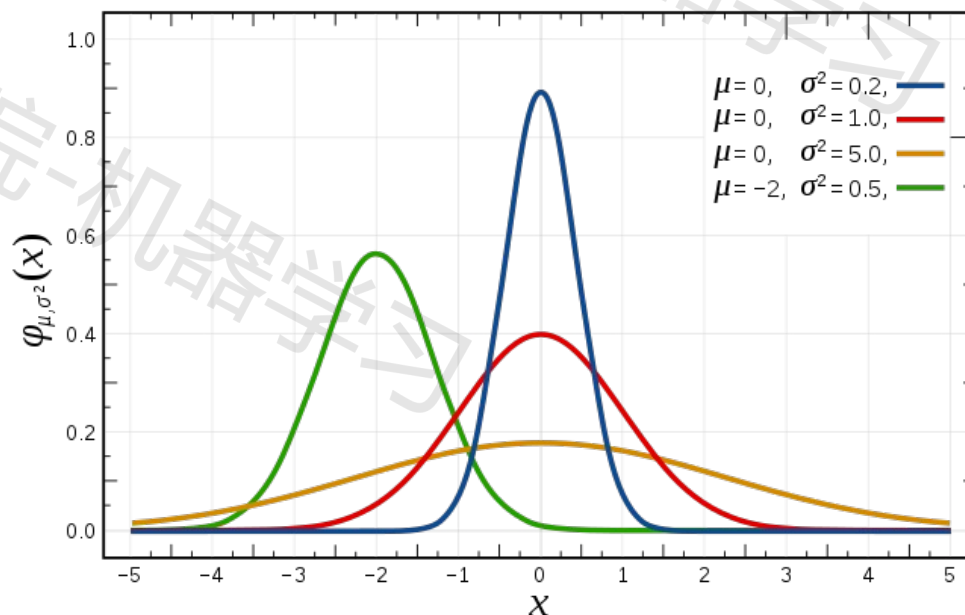
- $\mu$  为随机变量 $X$ 的均值，决定了分布的位置
- $\sigma$  为随机变量 $X$ 的标准差，决定了分布的幅度

# 相关概念

- 单变量高斯分布/正态分布

- ✓ 概率密度函数(probability density function, PDF)

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

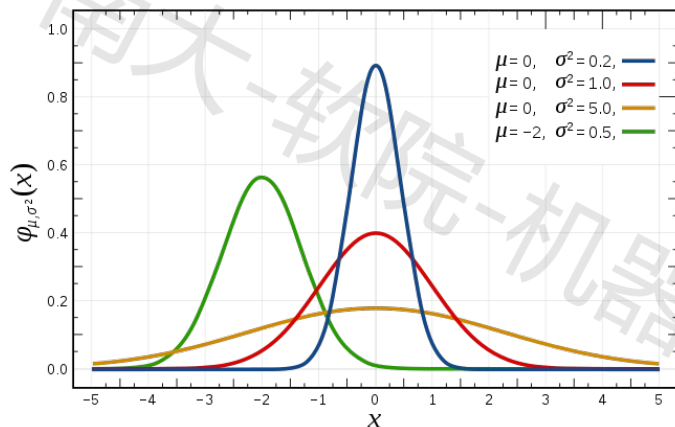


# 相关概念

- 单变量高斯分布/正态分布

- ✓ 概率密度函数(probability density function, PDF)

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



$$\forall -\infty < a < b < \infty,$$

$$\mathbb{P}[a < X \leq b] = F_X(b) - F_X(a) = \int_a^b p(x) dx$$

在 $(a, b]$ 范围概率

累积分布函数

# 相关概念

- 多变量高斯分布/正态分布

(Multivariate Gaussian distribution / Normal distribution)

✓ 若 $d$ 维随机变量 $\mathbf{X} = (X_1, \dots, X_d)^T$  服从高斯分布, 则

$$\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

✓ 概率密度函数(probability density function, PDF)

$$p_{\mathbf{X}}(x_1, \dots, x_k) = \frac{\exp(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}))}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}}$$

- $\boldsymbol{\mu} \in \mathbb{R}^d$  为随机变量 $\mathbf{X} \in \mathbb{R}^d$ 的均值向量
- $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ 为随机变量 $\mathbf{X} \in \mathbb{R}^d$ 的协方差矩阵

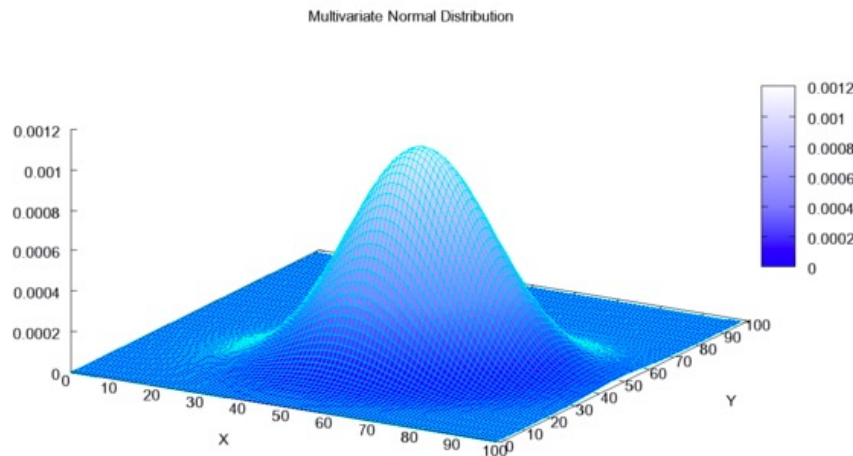
# 相关概念

- 多变量高斯分布/正态分布

(Multivariate Gaussian distribution / Normal distribution)

✓ 概率密度函数(probability density function, PDF)

$$p_X(x_1, \dots, x_k) = \frac{\exp(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}))}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}}$$



2变量正态分布

# 相关概念

- 多变量高斯分布/正态分布

(Multivariate Gaussian distribution / Normal distribution)

✓ 若 $d$ 维随机变量 $\mathbf{X} = (X_1, \dots, X_d)^T$  服从高斯分布, 则

$$\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

✓ 概率密度函数(probability density function, PDF)

$$p_{\mathbf{X}}(x_1, \dots, x_k) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}}$$

马氏距离的平方  
计算了 $\mathbf{x}$ 和 $\boldsymbol{\mu}$ 之间的距离

- $\boldsymbol{\mu} \in \mathbb{R}^d$  为随机变量 $\mathbf{X} \in \mathbb{R}^d$ 的均值向量
- $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ 为随机变量 $\mathbf{X} \in \mathbb{R}^d$ 的协方差矩阵



# 大纲

相关概念

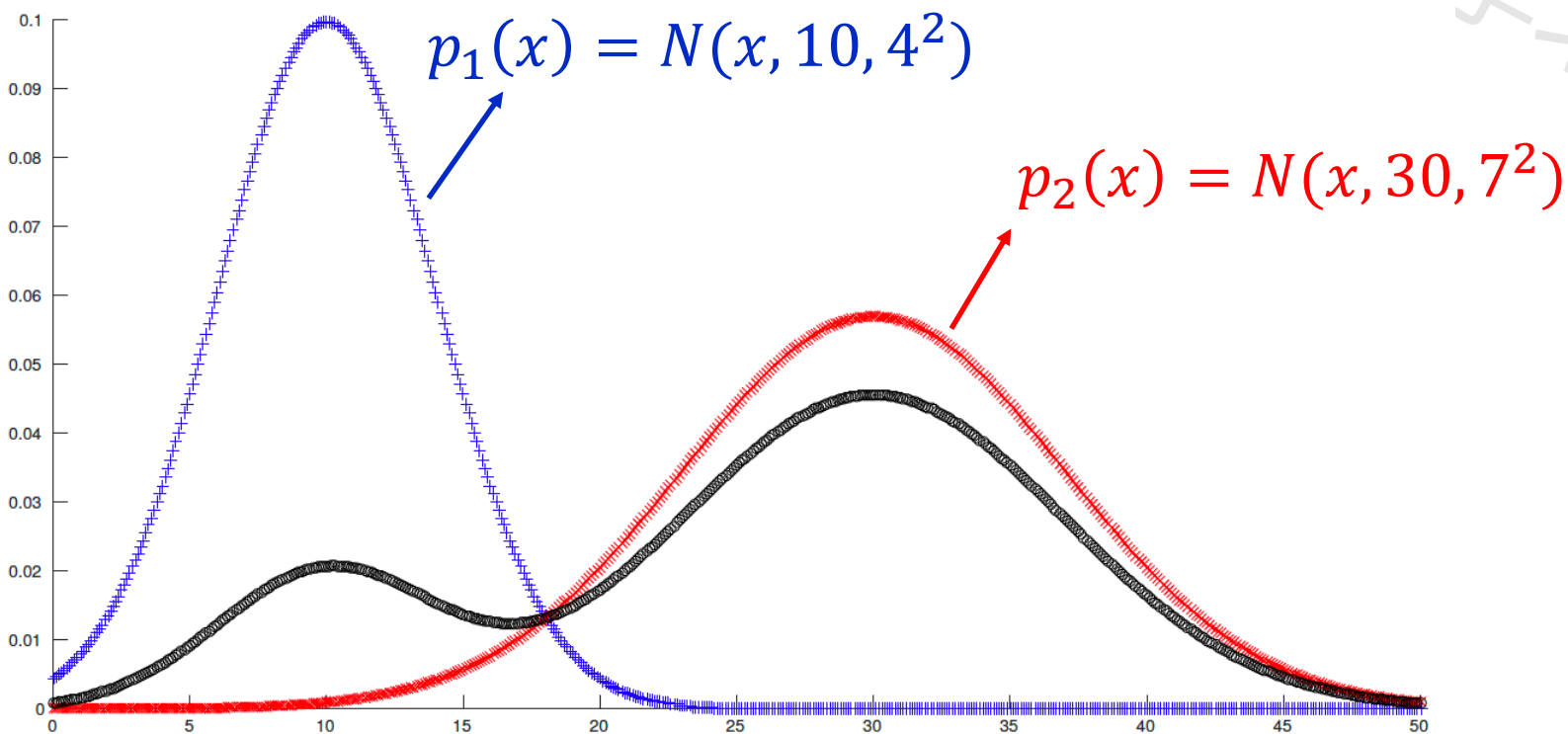
高斯混合模型

最大似然估计

期望最大化算法

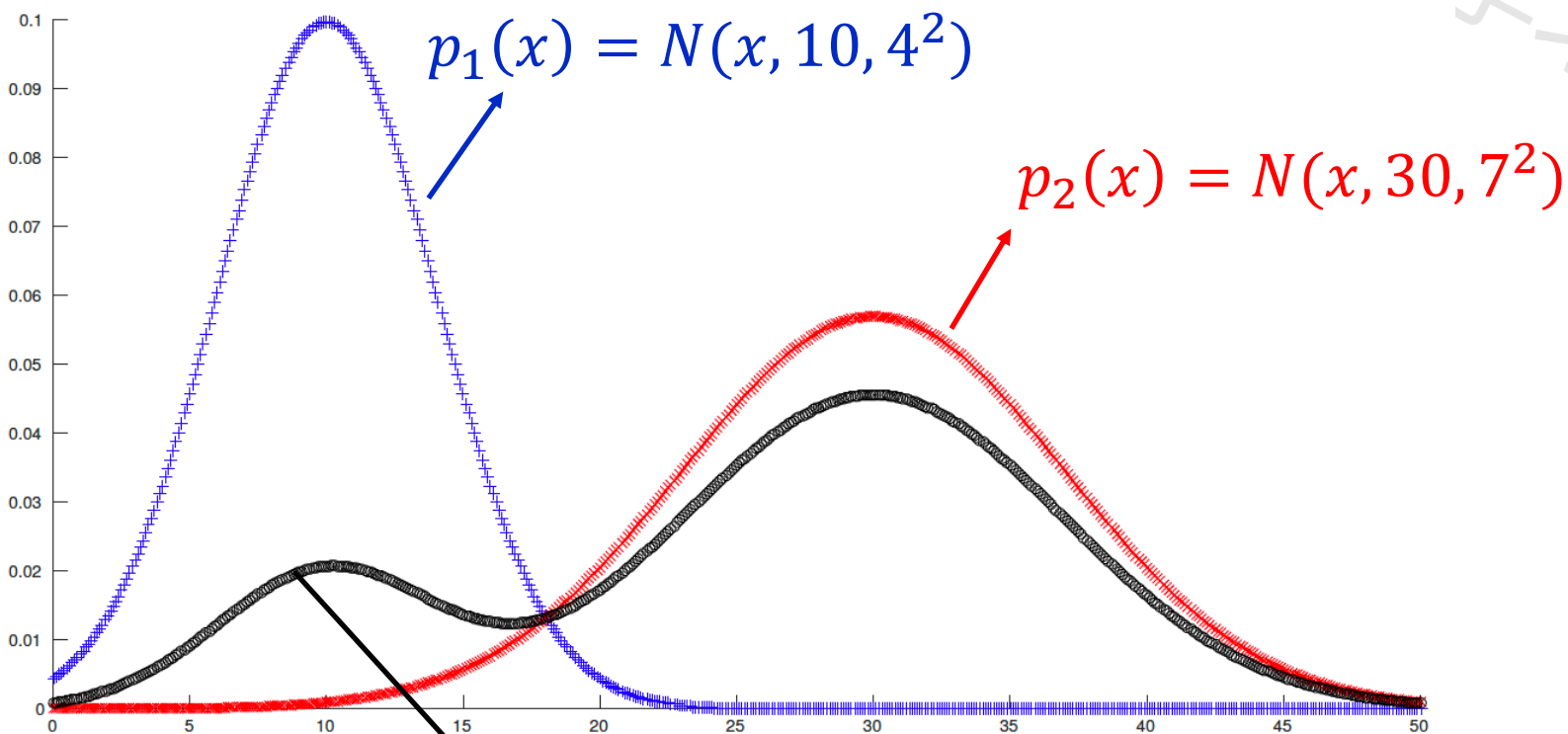
# 高斯混合模型

- 高斯混合模型 (Gaussian Mixture Model, GMM)



# 高斯混合模型

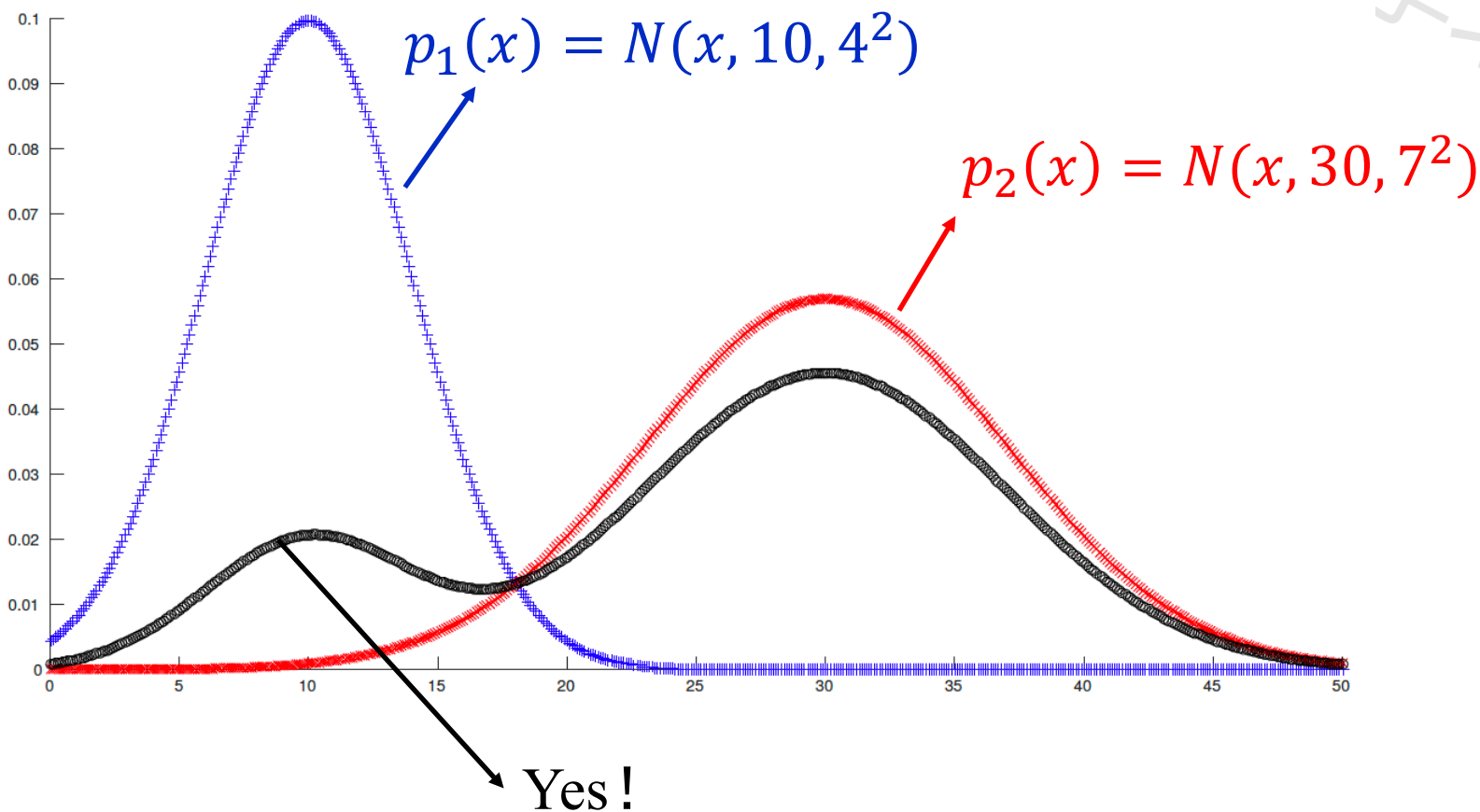
- 高斯混合模型 (Gaussian Mixture Model, GMM)



也是一个概率密度函数PDF?

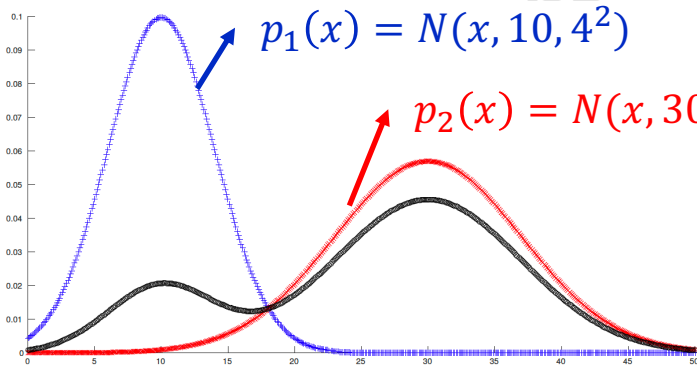
# 高斯混合模型

- 高斯混合模型 (Gaussian Mixture Model, GMM)



# 高斯混合模型

- 高斯混合模型 (Gaussian Mixture Model, GMM)



事实上,  $p_3(x)$  是这两个高斯分布的一个加权:

$$p_3(x) = 0.2p_1(x) + 0.8p_2(x)$$

高斯混合模型

# 高斯混合模型

- 高斯混合模型 (**Gaussian Mixture Model, GMM**)

- ✓ 概率密度函数(probability density function, PDF)

$$\begin{aligned} p(\mathbf{x}) &= \sum_{i=1}^N \alpha_i N(\mathbf{x}; \boldsymbol{\mu}_i, \Sigma_i) \\ &= \sum_{i=1}^N \frac{\alpha_i}{(2\pi)^{d/2} |\Sigma_i|^{1/2}} \exp \left( -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_i)^T \Sigma_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) \right) \end{aligned}$$

- 随机变量  $\mathbf{x} \in \mathbb{R}^d$
- $N$  表示有  $N$  个高斯分布组成成分
- $\forall i, \alpha_i \geq 0, \sum_{i=1}^N \alpha_i = 1$

# 高斯混合模型

- 高斯混合模型 (**Gaussian Mixture Model, GMM**)

$$\begin{aligned} p(\mathbf{x}) &= \sum_{i=1}^N \alpha_i N(\mathbf{x}; \boldsymbol{\mu}_i, \Sigma_i) \\ &= \sum_{i=1}^N \frac{\alpha_i}{(2\pi)^{d/2} |\Sigma_i|^{1/2}} \exp \left( -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_i)^T \Sigma_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) \right) \end{aligned}$$



参数:  $\theta = \{\alpha_i, \boldsymbol{\mu}_i, \Sigma_i\}_{i=1}^N$

- 随机变量  $\mathbf{x} \in \mathbb{R}^d$
- $N$  表示有  $N$  个高斯分布组成成分
- $\forall i, \alpha_i \geq 0, \sum_{i=1}^N \alpha_i = 1$

# 高斯混合模型

- 高斯混合模型（**Gaussian Mixture Model, GMM**）

- ✓ 将GMM看成一个图模型



- 假设随机变量  $Z \in \{1, 2, \dots, N\}$  符合多项式离散分布
- $Z$  取值为  $i$  的概率为：

$$Pr(Z = i) = \alpha_i$$

- ✓ Two-step sampling, 从GMM里采样一个样本  $x$

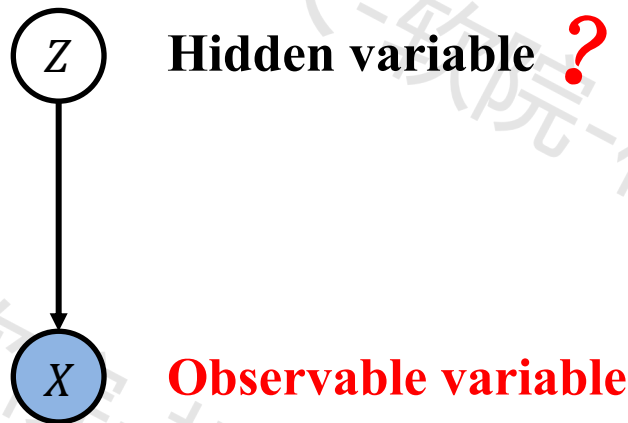
- 从  $Z$  中采样，得到一个值  $i$ ，其中  $(1 \leq i \leq N)$
- 从第  $i$  个高斯分布  $N(\mu_i, \Sigma_i)$  里采样  $x$



# 高斯混合模型

- 高斯混合模型 (Gaussian Mixture Model, GMM)

- ✓ 将GMM看成一个图模型



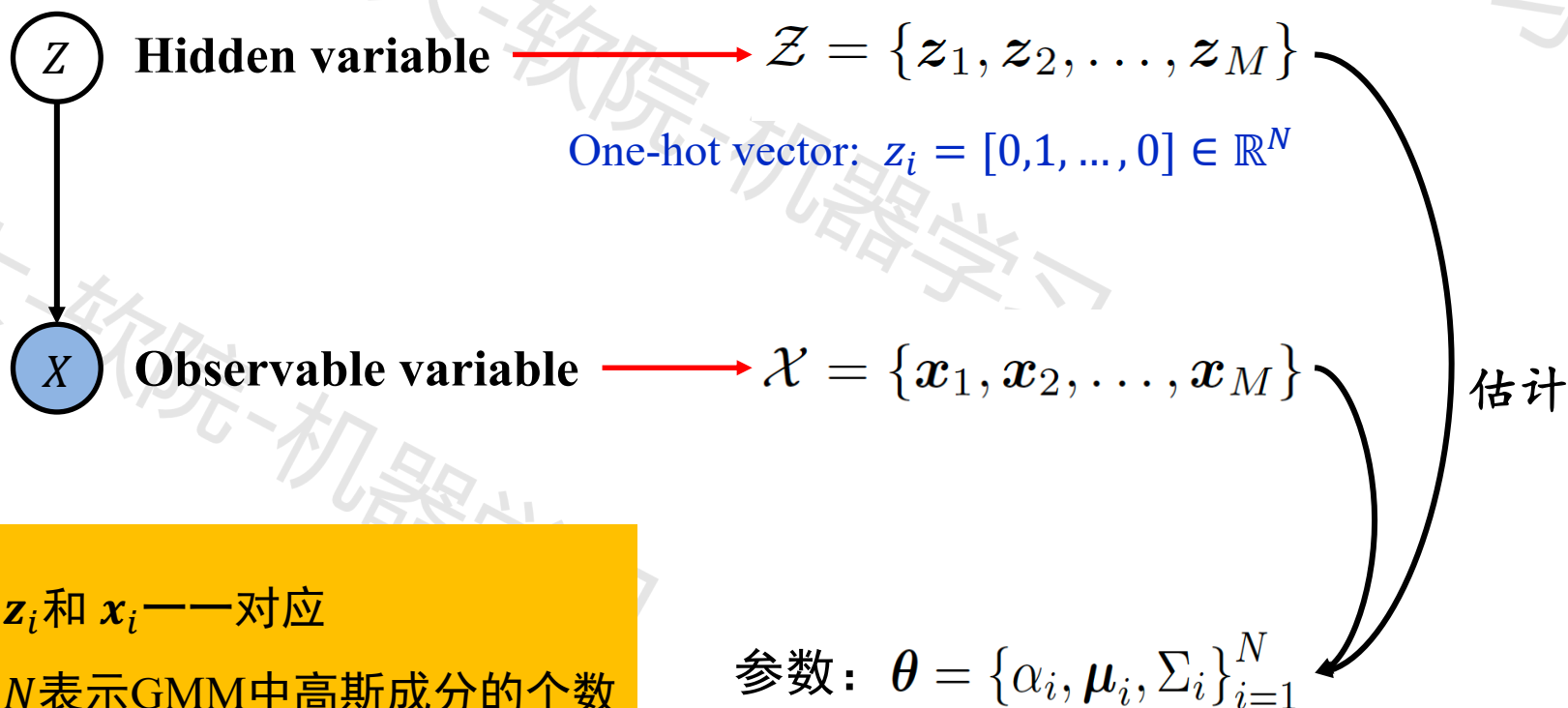
- ✓ Two-step sampling, 从GMM里采样一个样本 $x$

- 从 $Z$ 中采样, 得到一个值 $i$ , 其中  $(1 \leq i \leq N)$
- 从第 $i$ 个高斯分布 $N(\mu_i, \Sigma_i)$ 里采样 $x$

# 高斯混合模型

- 高斯混合模型（Gaussian Mixture Model, GMM）

✓ 将GMM看成一个图模型



- $z_i$  和  $x_i$  一一对应
- $N$  表示GMM中高斯成分的个数

# 高斯混合模型

- 高斯混合模型 (**Gaussian Mixture Model, GMM**)

✓ 假设一个特殊情形： $z$ 已知

$z_i$ 和  $x_i$ 一一对应，如何估计参数 $\theta$ ？

# 高斯混合模型

- 高斯混合模型（**Gaussian Mixture Model, GMM**）

✓ 假设一个特殊情形： $z$ 已知

$z_i$ 和  $x_i$ 一一对应，如何估计参数 $\theta$ ？

- 第一步：找到所有从第 $i$ 个高斯分量得到的采样，构成子集 $\mathcal{X}_i$

$$\mathcal{X}_i = \{\mathbf{x}_j | z_j = i, 1 \leq j \leq M\}$$

# 高斯混合模型

- 高斯混合模型 (Gaussian Mixture Model, GMM)

✓ 假设一个特殊情形： $z$ 已知

$z_i$ 和  $x_i$ 一一对应，如何估计参数 $\theta$ ？

- 第一步：找到所有从第 $i$ 个高斯分量得到的采样，构成子集 $\mathcal{X}_i$

$$\mathcal{X}_i = \{\mathbf{x}_j | z_j = i, 1 \leq j \leq M\}$$

- 第二步：统计和计算每个高斯分量的参数

$$\hat{\alpha}_i = \frac{m_i}{\sum_{j=1}^N m_j} = \frac{|\mathcal{X}_i|}{M}$$

$$\hat{\boldsymbol{\mu}}_i = \frac{1}{m_i} \sum_{\mathbf{x} \in \mathcal{X}_i} \mathbf{x}$$

$$\hat{\boldsymbol{\Sigma}}_i = \frac{1}{m_i} \sum_{\mathbf{x} \in \mathcal{X}_i} (\mathbf{x} - \hat{\boldsymbol{\mu}}_i)(\mathbf{x} - \hat{\boldsymbol{\mu}}_i)^T$$

# 大纲

相关概念

高斯混合模型

最大似然估计

期望最大化算法

# 最大似然估计

- 最大似然估计 (Maximum likelihood estimation, MLE)

- ✓ **定义**: MLE是通过最大化一个似然函数来估计一个概率分布的参数, 使得在假设的统计模型下, 观测数据最有可能出现。

- ✓ **单高斯模型为例 (单变量)** :

似然函数 (Likelihood function) :

$$\mathcal{L}(\theta|X) = p(X|\theta)$$

- $\theta$ 固定 (数据分布假设固定),  $p(X|\theta)$ 看作是 $X$ 的函数, 即为概率密度函数PDF
    - $X$ 固定 (观测数据固定),  $\mathcal{L}(\theta|X)$ 看作是 $\theta$ 的函数, 即为似然函数

# 最大似然估计

- 最大似然估计 (Maximum likelihood estimation, MLE)

✓ 单高斯模型为例（单变量）：

似然函数 (Likelihood function) :  $\mathcal{L}(\theta|X) = p(X|\theta)$

最大似然估计MLE:

$$\hat{\theta} = \operatorname{argmax}_{\theta \in \Theta} \mathcal{L}(\theta|X)$$

假设数据点i.i.d

$$\mathcal{L}(\theta|X) = \prod_{j=1}^M p(x_j|\theta)$$

乘积很小



$$\ln \mathcal{L}(\theta|X) = \sum_{j=1}^M \ln p(x_j|\theta)$$

对数似然函数



# 最大似然估计

- 最大似然估计 (Maximum likelihood estimation, MLE)

✓ 单高斯模型为例:

最大似然估计MLE:  $\hat{\theta} = \operatorname{argmax}_{\theta \in \Theta} \mathcal{L}(\theta|X)$



最大对数似然估计MLE:  $\hat{\theta} = \operatorname{argmax}_{\theta \in \Theta} \ln \mathcal{L}(\theta|X)$   
 $= \operatorname{argmax}_{\theta \in \Theta} \sum_{j=1}^M \ln p(x_j|\theta)$

求解:

- 求导, 令导数为0
- 求解方程, 得到最优 $\theta^*$

$$\theta = (\mu, \sigma^2)$$

$$\theta^* = (\mu^*, \sigma^{*2})$$

# 最大似然估计

- 最大似然估计 (Maximum likelihood estimation, MLE)

- ✓ 高斯混合模型:

最大对数似然估计MLE:  $\hat{\theta} = \operatorname{argmax}_{\theta \in \Theta} \ln \mathcal{L}(\theta|X)$

$$\ln \mathcal{L}(\theta|X) = \sum_{j=1}^M \ln p(\mathbf{x}_j|\theta) = \sum_{j=1}^M \ln \sum_{i=1}^N \alpha_i N(\mathbf{x}_j; \boldsymbol{\mu}_i, \Sigma_i)$$

$$= \sum_{\mathbf{x}} \ln \sum_{\mathbf{z}} p(\mathbf{z}|\alpha) p(\mathbf{x}|\mathbf{z}; \boldsymbol{\mu}, \Sigma)$$

$$= \sum_{\mathbf{x}} \ln \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}|\theta)$$

$$Pr(\mathbf{Z} = i) = \alpha_i$$

$$\theta = \{\alpha, \boldsymbol{\mu}, \Sigma\}$$

# 大纲

相关概念

高斯混合模型

最大似然估计

期望最大化算法

# 期望最大化(EM)算法

- EM算法 (Expectation-Maximization algorithm)

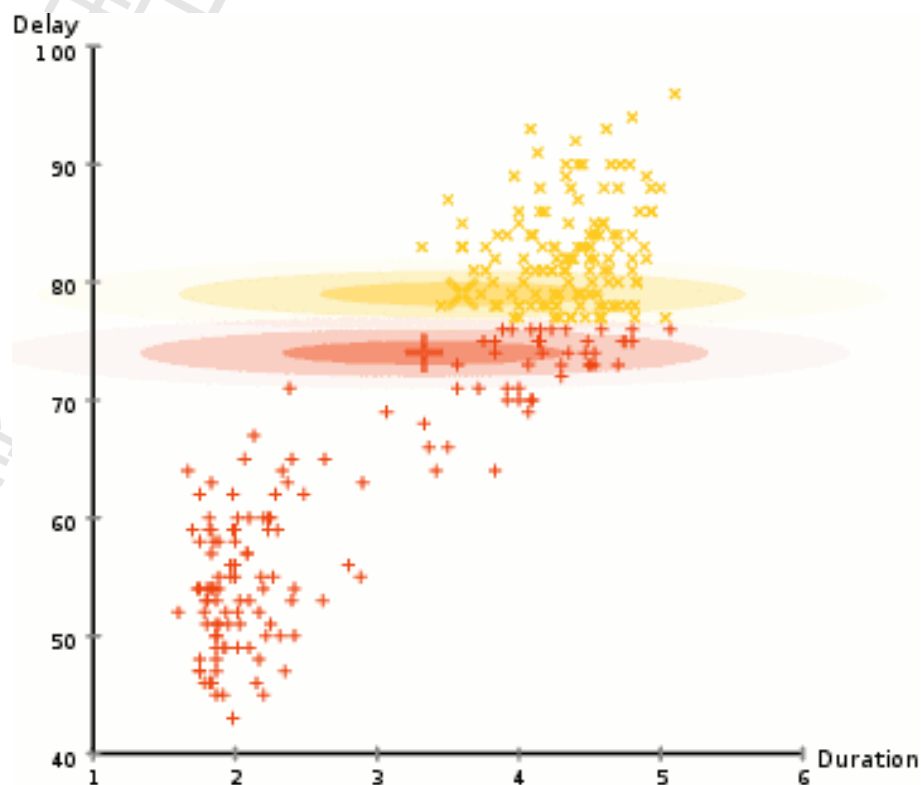
- ✓ 核心思想:

EM算法是一个迭代的方法，采用最大似然估计MLE对统计模型中的参数进行估计，特别是针对包含无法观测隐变量的模型。

通常引入隐含变量后会有两个参数，EM算法首先会固定其中的第一个参数，然后使用MLE计算第二个变量值；接着通过固定第二个变量，再使用MLE估测第一个变量值，依次迭代，直至收敛到局部最优解。

# 期望最大化 (EM) 算法

- EM算法 (Expectation-Maximization algorithm)



Wiki: EM clustering of Old Faithful eruption data

# 期望最大化 (EM) 算法

- EM算法 (Expectation-Maximization algorithm)

**E-Step:** 利用可观测数据 $\mathcal{X}$ 和当前估计的参数为 $\theta^{(t)}$ , 估计更好的隐藏变量 $z$



**M-Step:** 利用可观测数据 $\mathcal{X}$ 和当前估计的隐藏变量 $z$ , 估计更好的参数 $\theta^{(t+1)}$

**Repeat:** 重复上述两个步骤, 直至收敛

# 期望最大化 (EM) 算法

- EM优化分析

- ✓ 假设隐变量 $z$ 的分布 $Q(z|\theta)$ 是一个任意的离散分布

满足: 
$$\sum_z Q(z|\theta) = 1, Q(z|\theta) \geq 0$$

- ✓ 高斯混合模型:

$$\begin{aligned}\ell(\theta) &= \ln \mathcal{L}(\theta|X) = \sum_X \ln \sum_z p(X, z|\theta) \\ &= \sum_X \ln \sum_z Q(z|\theta) \frac{p(X, z|\theta)}{Q(z|\theta)} \\ &= \sum_X \ln E_Q \left[ \frac{p(X, z|\theta)}{Q(z|\theta)} \right]\end{aligned}$$

**定理1**：令 $X$ 表示一个随机变量，存在某个函数 $g$ 使得 $Y = g(X)$

1. 假设 $X$ 是连续的，pdf为 $f_X(x)$ 。如果 $\int_{-\infty}^{\infty} |g(x)|f_X(x)dx < \infty$ ，那么 $Y$ 的期望存在且为

$$E(Y) = \int_{-\infty}^{\infty} g(x)f_X(x)dx$$

2. 假设 $X$ 是离散的，pmf为 $p_X(x)$ 。假设 $X$ 的支撑用 $S_X$ 表示，如果 $\sum_{x \in S_X} |g(x)|p_X(x) < \infty$ ，那么 $Y$ 的期望存在且为

$$E(Y) = \sum_{x \in S_X} g(x)p_X(x)$$

✓ 高斯混合模型：

$$\begin{aligned}\ell(\theta) &= \ln \mathcal{L}(\theta|X) = \sum_X \ln \sum_Z p(X, Z|\theta) \\ &= \sum_X \ln \sum_Z \boxed{Q(Z|\theta)} \boxed{\frac{p(X, Z|\theta)}{Q(Z|\theta)}} \\ &= \sum_X \ln E_Q \left[ \frac{p(X, Z|\theta)}{Q(Z|\theta)} \right]\end{aligned}$$



# 期望最大化 (EM) 算法

- EM优化分析

- ✓ 高斯混合模型:

$$\ell(\theta) = \ln \mathcal{L}(\theta|X) = \sum_X \ln \sum_Z p(X, Z|\theta)$$

$$= \sum_X \ln \sum_Z Q(Z|\theta) \frac{p(X, Z|\theta)}{Q(Z|\theta)}$$

$$= \sum_X \ln E_Q \left[ \frac{p(X, Z|\theta)}{Q(Z|\theta)} \right]$$

$$\geq \sum_X E_Q \left[ \ln \frac{p(X, Z|\theta)}{Q(Z|\theta)} \right]$$

利用Jensen不等式, 因为  $\ln(\cdot)$  函数是凹函数, 所以  $\ln(E[X]) \geq E[\ln(X)]$

$$= \sum_X \sum_Z Q(Z|\theta) \ln \frac{p(X, Z|\theta)}{Q(Z|\theta)}$$

$$E[\ln g(Z)] = \sum_Z p(Z) \ln g(Z)$$

# 期望最大化 (EM) 算法

- EM优化分析

- ✓ 高斯混合模型:

$$\ell(\theta) \geq \sum_X \sum_Z Q(Z|\theta) \ln \frac{p(X, Z|\theta)}{Q(Z|\theta)}$$

下界?

什么时候上述不等式可以取等号?

$$X = E[X] \quad \checkmark$$

也就是说 $X$ 为常数时, 即:

$$\frac{p(X, Z|\theta)}{Q(Z|\theta)} = c$$

# 期望最大化 (EM) 算法

- EM优化分析

- ✓ 高斯混合模型:

$$\ell(\theta) \geq \sum_X \sum_Z Q(Z|\theta) \ln \frac{p(X, Z|\theta)}{Q(Z|\theta)}$$

下界?

上式取等号, 即

$$\frac{p(X, Z|\theta)}{Q(Z|\theta)} = c$$

$$Q(Z|\theta) = \frac{p(X, Z|\theta)}{c} = \frac{p(X, Z|\theta)}{c \cdot \sum_Z Q(Z|\theta)} \longrightarrow \sum_Z Q(Z|\theta) = 1$$

$$= \frac{p(X, Z|\theta)}{\sum_Z c \cdot Q(Z|\theta)} = \frac{p(X, Z|\theta)}{\sum_Z p(X, Z|\theta)} \longrightarrow \frac{p(X, Z|\theta)}{Q(Z|\theta)} = c$$

$$= \frac{p(X, Z|\theta)}{p(X|\theta)} = p(Z|X, \theta) \longrightarrow \text{固定 } \theta, \text{ 即为 } Z \text{ 的后验概率}$$

# 期望最大化 (EM) 算法

- EM优化分析

- ✓ 高斯混合模型:

$$\ell(\theta) = \sum_X \sum_Z p(Z|X, \theta) \ln \frac{p(X, Z|\theta)}{p(Z|X, \theta)}$$

下界

固定 $\theta$ , 计算 $Q(Z|\theta) = p(Z|X, \theta)$ ,  
就可以得到 $\ell(\theta)$ 的下界

→ E-Step

- ✓ 然后继续优化这个下界

$$\theta^* = \operatorname{argmax}_{\theta \in \Theta} \ell(\theta)$$

$$= \operatorname{argmax}_{\theta \in \Theta} \sum_X \sum_Z p(Z|X, \theta) \ln \frac{p(X, Z|\theta)}{p(Z|X, \theta)}$$

→ M-Step

# 期望最大化 (EM) 算法

- 算法流程

**Initialization:**  $t \leftarrow 0$ ;  $\theta^{(0)}$

**E-Step:** 根据观测数据  $X$  和上一次迭代的参数  $\theta^{(t)}$ , 计算隐藏变量  $Z$  的后验概率, 或者称为隐变量的期望值;

$$Q^t = p(Z|X, \theta^t) = \frac{p(X, Z|\theta^t)}{\sum_Z p(X, Z|\theta^t)}$$

**M-Step:** 在上述  $Z$  的后验概率的基础上, 进行最大化似然估计, 估计新的参数  $\theta^{(t+1)}$

$$\theta^{(t+1)} = \operatorname{argmax}_{\theta \in \Theta} \sum_X \sum_Z Q^t \ln \frac{p(X, Z|\theta)}{Q^t}$$

**Repeat:** 重复上述两个步骤, 直至收敛

# 期望最大化 (EM) 算法

- E-Step

✓ 计算每个样本  $\mathbf{x}_j$  来自第  $i$  个高斯分布的期望：

$$\gamma_{ij} = \mathbb{E} \left[ z_{ij} | \mathbf{x}_j, \boldsymbol{\theta}^{(t)} \right] = \frac{\alpha_i^{(t)} N(\mathbf{x}_j; \boldsymbol{\mu}_i^{(t)}, \boldsymbol{\Sigma}_i^{(t)})}{\sum_{k=1}^N \alpha_k^{(t)} N(\mathbf{x}_j; \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)})}$$

其中,  $1 \leq i \leq N, 1 \leq j \leq M$

$z_{ij}$  取值为0或者1;  $z_{ij} = 1$ , 当且仅当  $\mathbf{x}_j$  由第  $i$  个高斯分布产生

# 期望最大化 (EM) 算法

- M-Step

✓ 计算新一轮迭代的模型参数 $\theta^{(t+1)}$ ：

$$m_i = \sum_{j=1}^M \gamma_{ij},$$

$$\mu_i^{(t+1)} = \frac{\sum_{j=1}^M \gamma_{ij} \mathbf{x}_j}{m_i},$$

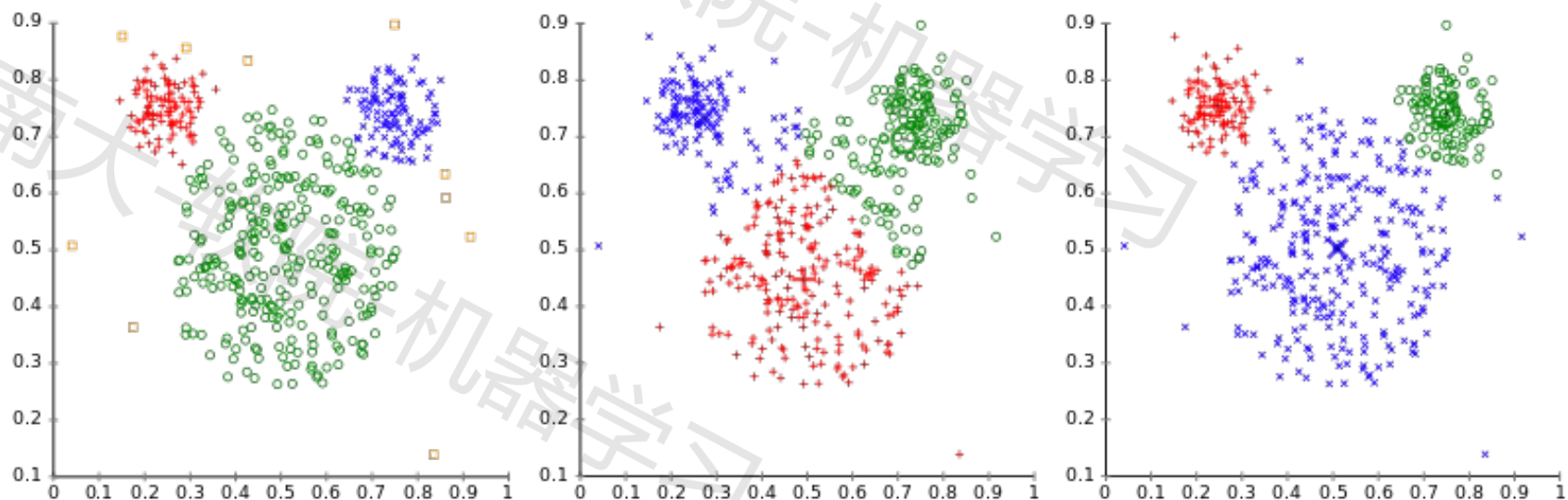
$$\Sigma_i^{(t+1)} = \frac{\sum_{j=1}^M \gamma_{ij} \left( \mathbf{x}_j - \mu_i^{(t+1)} \right) \left( \mathbf{x}_j - \mu_i^{(t+1)} \right)^T}{m_i}$$

$$\alpha_i = \frac{\sum_{j=1}^M \gamma_{ij}}{M} = \frac{m_i}{M}$$

# 期望最大化 (EM) 算法

- 应用

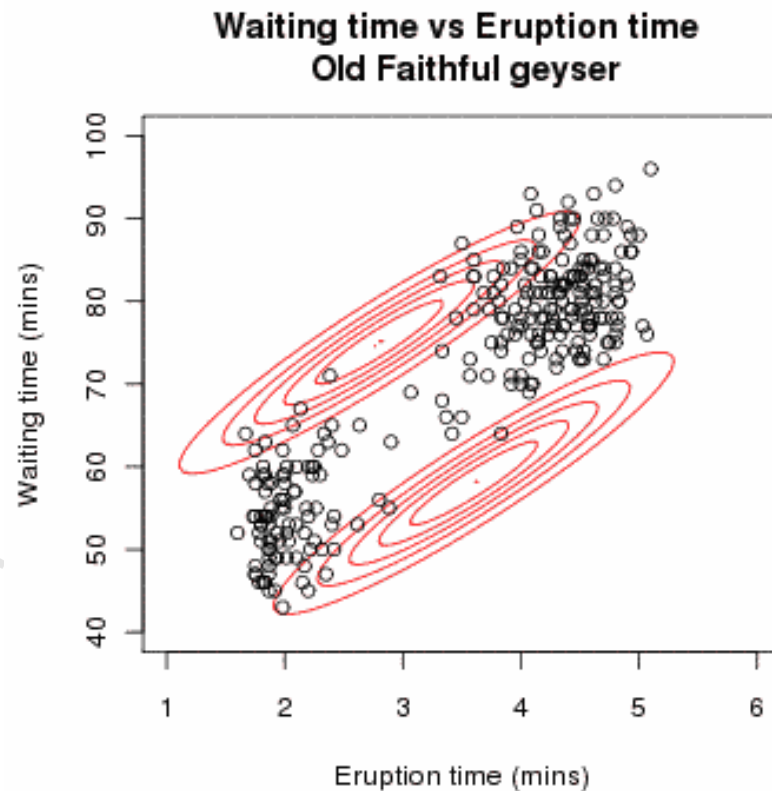
Different cluster analysis results on "mouse" data set:  
Original Data      k-Means Clustering      EM Clustering





# 期望最大化 (EM) 算法

- 应用



使用EM算法来拟合包含两个  
高斯分布的高斯混合模型

# 期望最大化 (EM) 算法

- 思考一下

K-means算法与EM算法的关系？

# 期望最大化 (EM) 算法

- K-means算法背后的EM思想

聚类准则函数：

$$J = \sum_{j=1}^c \sum_{x \in S_j} \|x - m_j\|^2$$

- 样本 $x_i$ 是可观测量 $X$ ；
- 类别标签（簇） $S_j$ 看作是隐藏变量 $Z$
- 簇中心/聚类均值 $m_j$ 看作参数 $\theta$
- 聚类准则函数看作 $\theta$ 的似然函数

# 期望最大化 (EM) 算法

- K-means算法背后的EM思想

- Step1: 选择一个聚类数量 $k$
- Step2: 初始化聚类中心 $\mu_1, \dots, \mu_k$ 
  - 随机选择 $k$ 个样本点, 设置这些样本点为中心
- Step3: 对每个样本点, 计算样本点到 $k$ 个聚类中心的距离 (使用某种距离度量方法), 将样本点分给距离它最近的聚类中心所属的聚类
- Step4: 重新计算聚类中心, 聚类中心为属于这一个聚类的所有样本的均值
- Step5: 如果没有发生样本所属的聚类改变的情况, 则退出, 否则, 返回Step3继续。

初始化

E-Step

M-Step

南大-软院-机器学习

南大-软院-机器学习

南大-软院-机器学习

谢谢！