机器学习导论-高斯混合模型

Introduction to Machine Learning-GMM

李文斌

https://cs.nju.edu.cn/liwenbin

liwenbin@nju.edu.cn

2024年04月23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

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

$\mathbf{A}$	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 **or** 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

The Matrix Cookbook. 2012

相关概念

❑ 符号和术语

$\det(\mathbf{A})$	Determinant of $\mathbf{A}$
$\mathrm{Tr}(\mathbf{A})$	Trace of the matrix $\mathbf{A}$
$\mathrm{diag}(\mathbf{A})$	Diagonal matrix of the matrix $\mathbf{A}$, i.e. $(\mathrm{diag}(\mathbf{A}))_{ij} = \delta_{ij}A_{ij}$
$\mathrm{eig}(\mathbf{A})$	Eigenvalues of the matrix $\mathbf{A}$
$\mathrm{vec}(\mathbf{A})$	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
$\mathbf{\Sigma}$	A positive definite matrix
$\mathbf{\Lambda}$	A diagonal matrix

The Matrix Cookbook. 2012

相关概念

❑ 一个带约束的数学优化问题

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

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

相关概念

❑ 一个**不带约束的**数学优化问题

❑ 最小二乘（least-norm-squared）问题

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

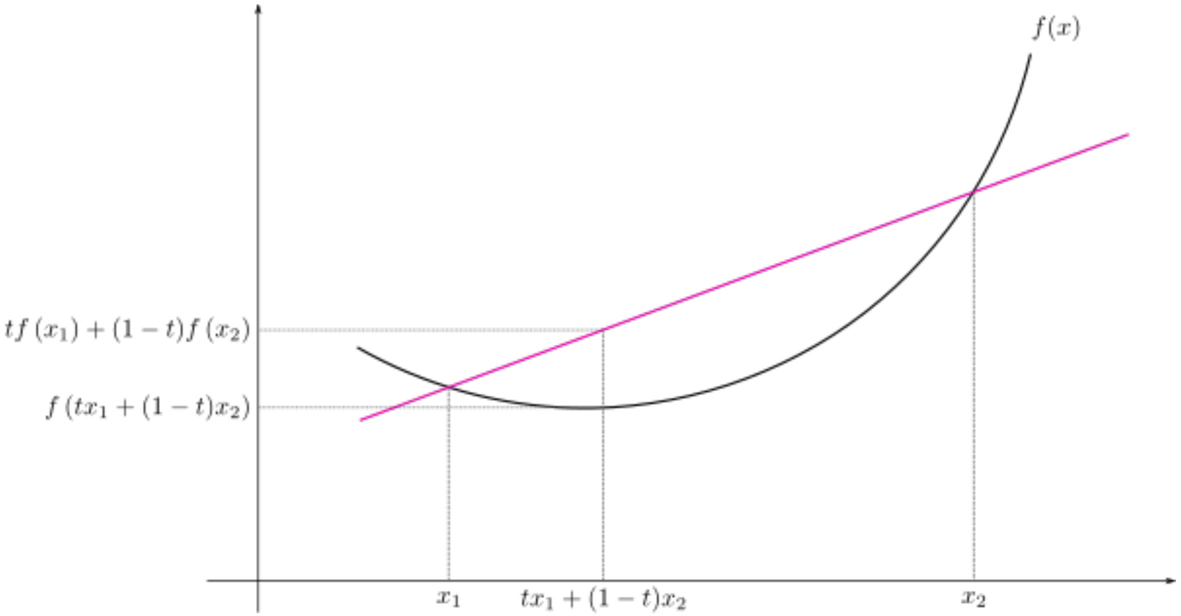
- 优化变量：$\boldsymbol{x} = (x_1, \dots, x_n)^T \in \mathbb{R}^n$
- 系数矩阵：$A \in \mathbb{R}^{k*n}$
- A 的第i行：$a_i^T \in \mathbb{R}^{1*n}$
- 最优解：$\boldsymbol{x}^* = (x_1, \dots, x_n)^T \in \mathbb{R}^n$

相关概念

❑ 凸函数

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

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

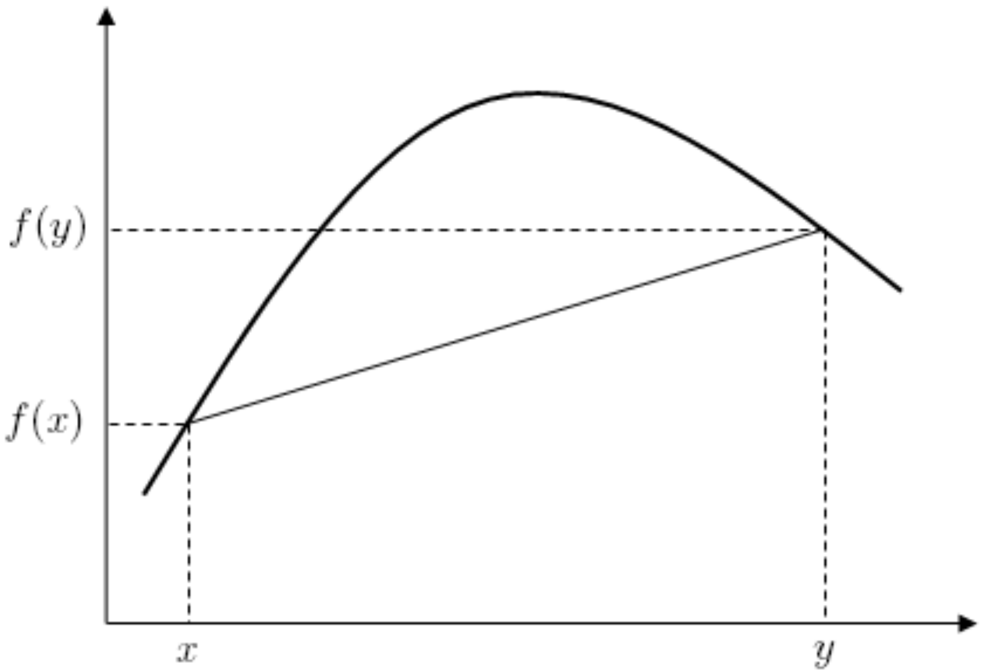


相关概念

❑ 凹函数

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

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



相关概念

❑ 判断函数的凹凸

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

如果二阶导数 $f''(x) \geq 0$ $f(x)$是凸函数

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

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

相关概念

❑ 判断函数的凹凸

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

如果二阶导数 $f''(x) \leq 0$ $f(x)$是凹函数

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

如果Hessian矩阵是非半正定的，即 $\nabla^2 f(\boldsymbol{x}) \preccurlyeq 0$ $f(x)$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.* $f(x) = \max\{x_1, \ldots, x_n\}$ is convex on $\mathbf{R}^n$.
- *Geometric mean.* The geometric mean $f(x) = \left(\prod_{i=1}^n x_i\right)^{1/n}$ is concave on $\mathbf{dom}\, f = \mathbf{R}^n_{++}$.
- *Log-determinant.* The function $f(X) = \log\det X$ is concave on $\mathbf{dom}\, f = \mathbf{S}^n_{++}$.

  - 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

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

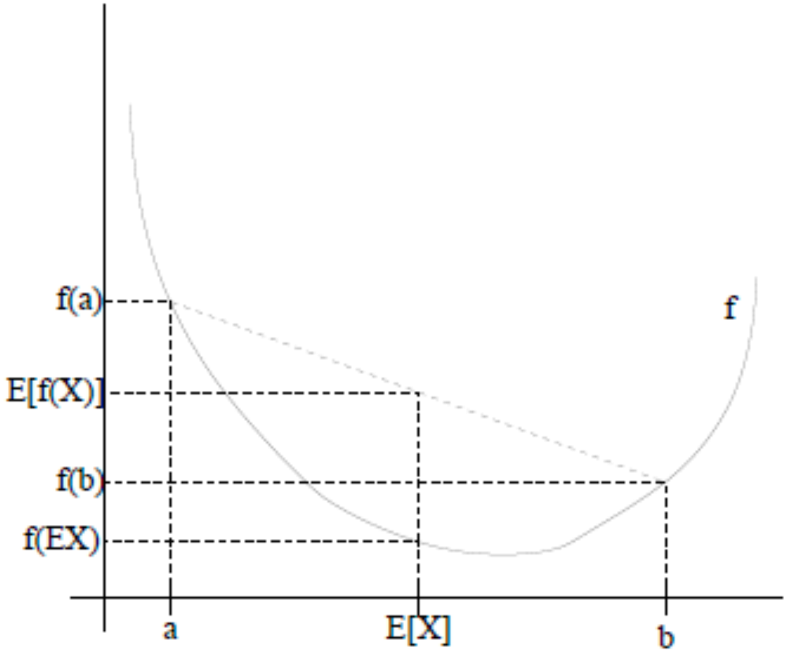
2. 假设X是离散的，pmf为$p_X(x)$。假设X的支撑用$\mathcal{S}_X$表示，如果$\Sigma_{x \in \mathcal{S}_X} |g(x)| p_X(x) < \infty$，那么$Y$的期望存在且为

$$E(Y) = \sum_{x \in \mathcal{S}_X} g(x) p_X(x)$$

- 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)

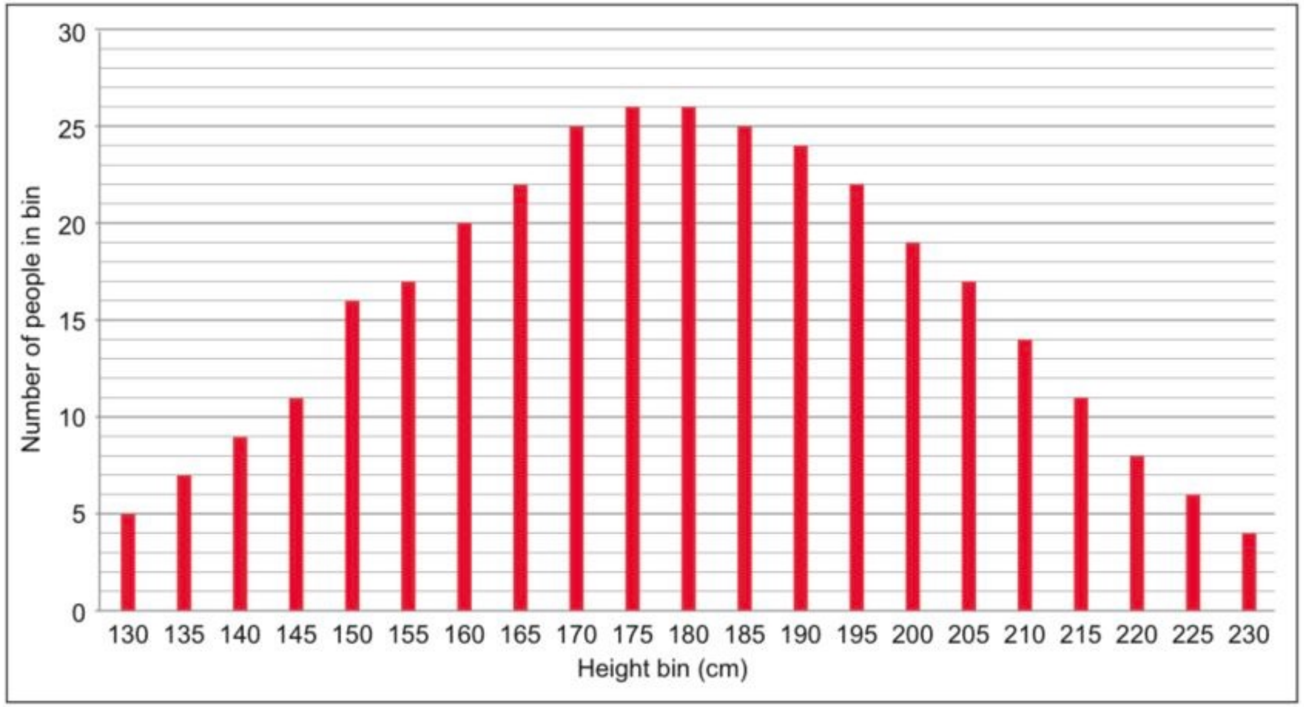


Figure 2.9 Histogram of a normal distribution, in this case the height of 334 fictitious people. The modal (most frequently occurring) bin is centered at 180 cm.

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)

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

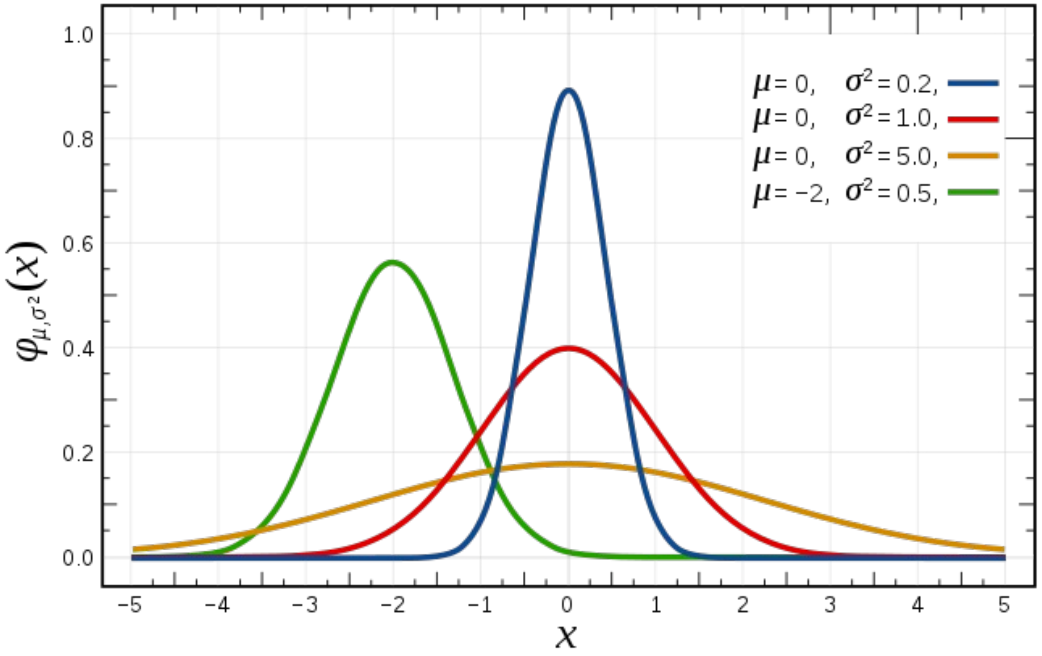
- μ 为随机变量X的均值，决定了分布的位置
- σ为随机变量X的标准差，决定了分布的幅度

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)

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

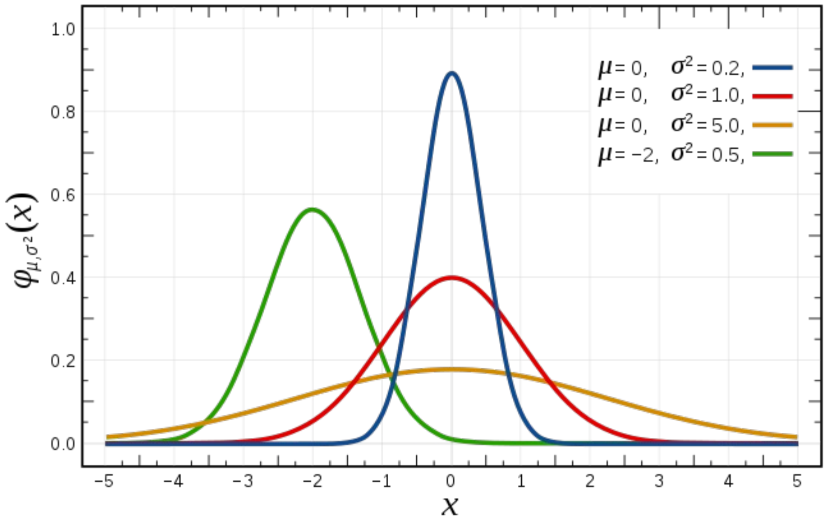


相关概念

❑ 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$\boldsymbol{X} = (X_1, \dots, X_d)^T$ 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)

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

  ○ $\boldsymbol{\mu} \in \mathbb{R}^d$ 为随机变量$\boldsymbol{X} \in \mathbb{R}^d$的均值向量

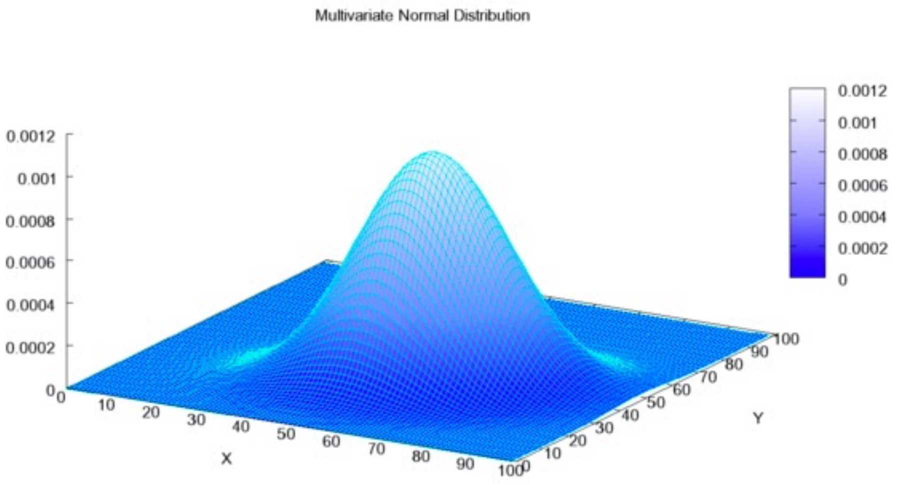
  ○ $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$为随机变量$\boldsymbol{X} \in \mathbb{R}^d$的协方差矩阵

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)

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



2变量的正态分布

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$\boldsymbol{X} = (X_1, \dots, X_d)^T$ 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)

$$p_{\boldsymbol{X}}(x_1, \dots, x_k) = \frac{\exp\left(\boxed{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\text{) 马氏距离的平方}}\right.}{\sqrt{(2\pi)^d|\Sigma|}}$$

计算了x和μ之间的距离

○ $\boldsymbol{\mu} \in \mathbb{R}^d$ 为随机变量$\boldsymbol{X} \in \mathbb{R}^d$的均值向量

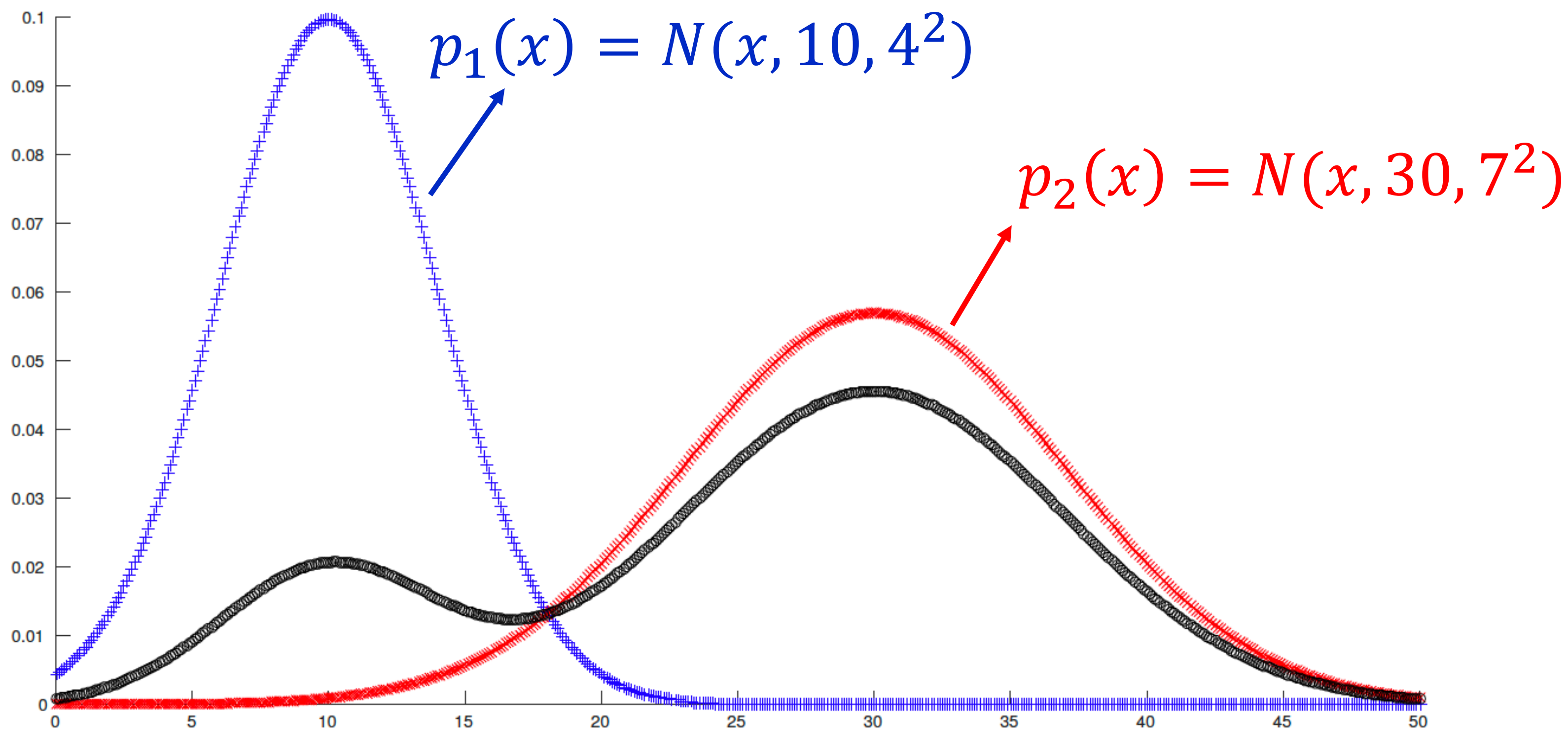
○ $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$为随机变量$\boldsymbol{X} \in \mathbb{R}^d$的协方差矩阵

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

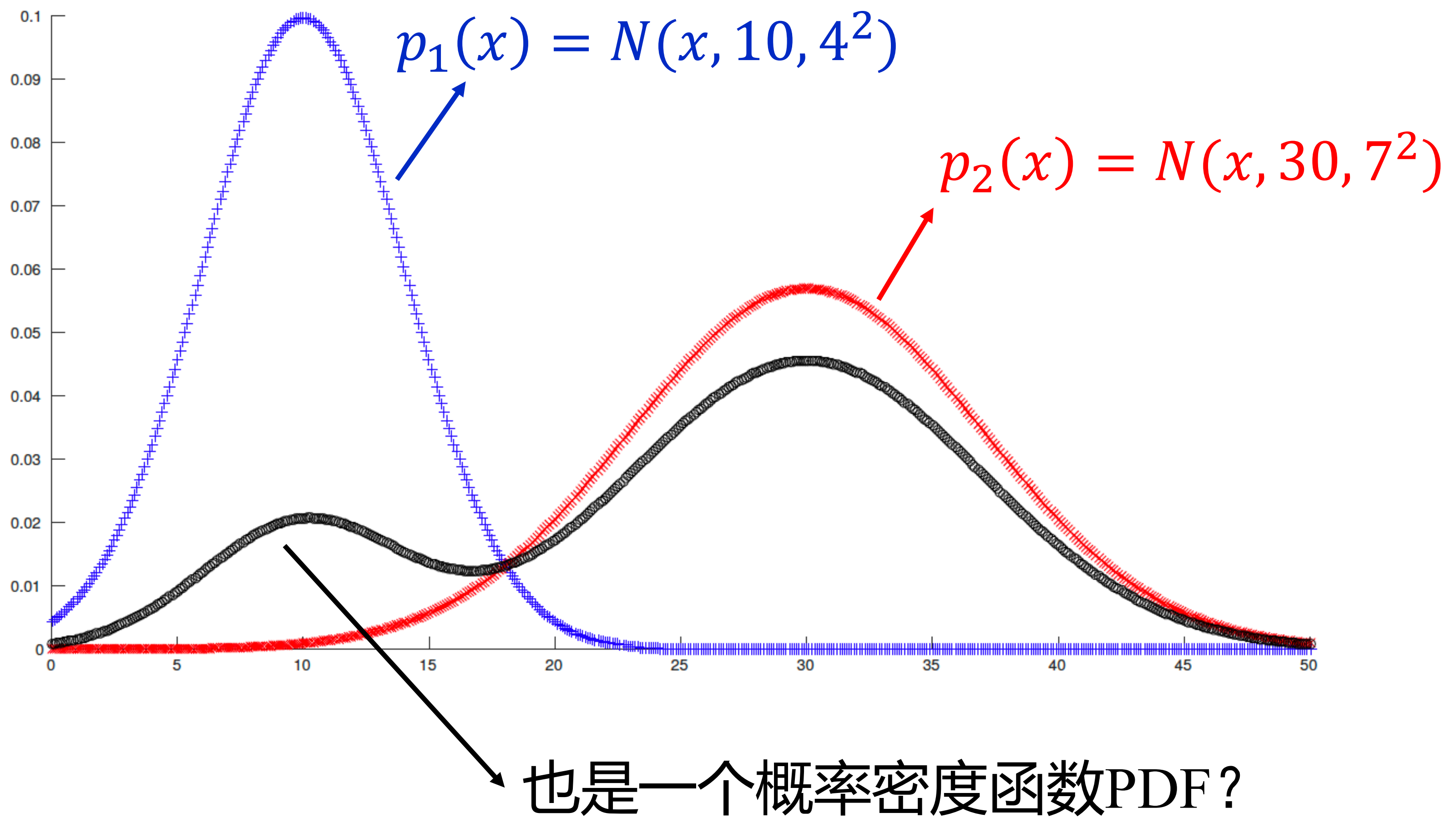
高斯混合模型

- 高斯混合模型（Gaussian Mixture Model, GMM）



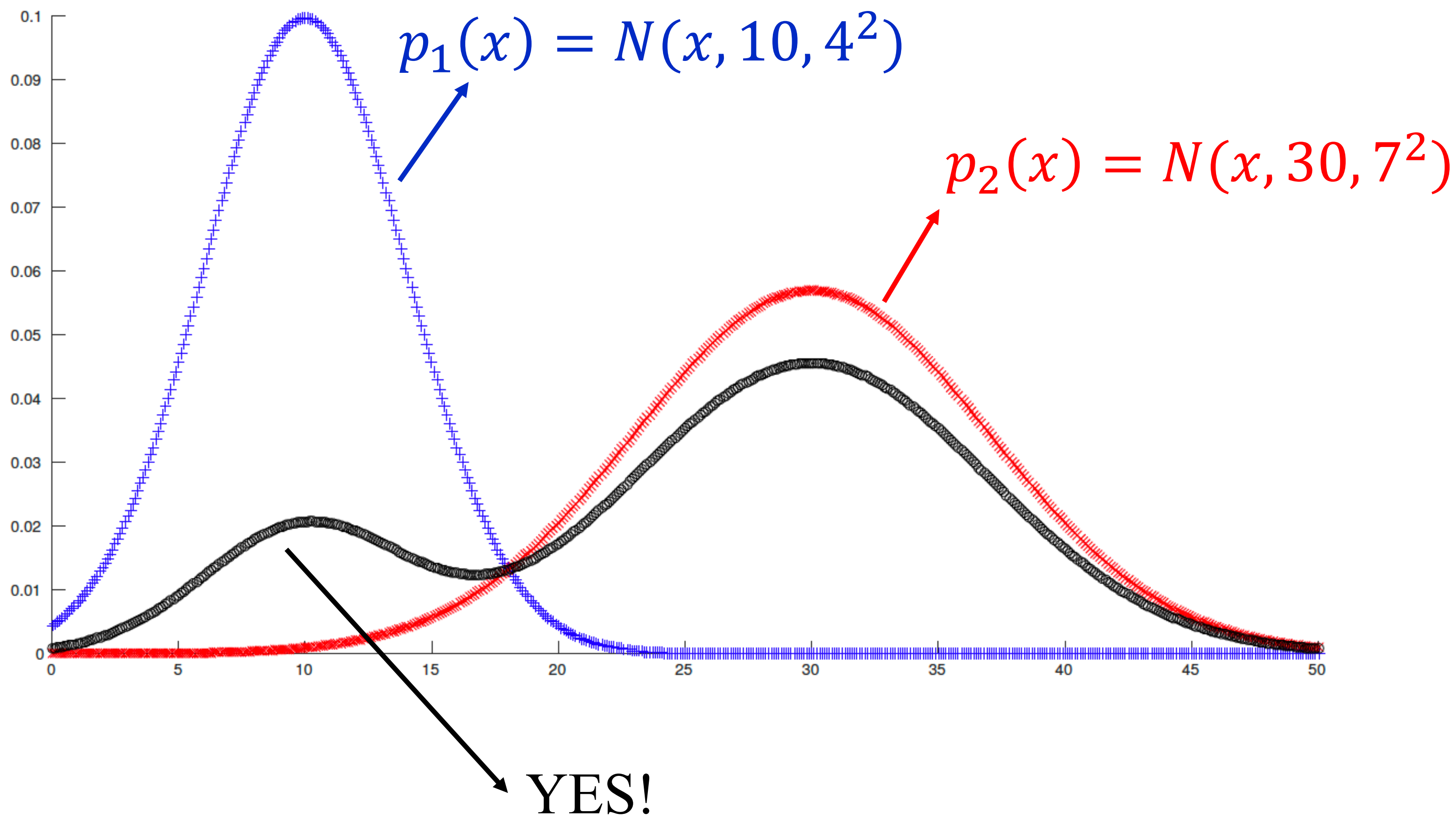
高斯混合模型

- 高斯混合模型（Gaussian Mixture Model, GMM）



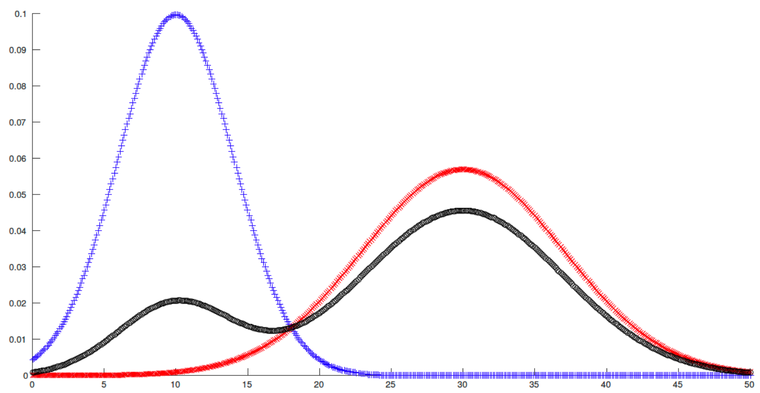
高斯混合模型

- 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)

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

- 随机变量$\boldsymbol{x} \in \mathbb{R}^d$
- 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）

$$p(x) = \sum_{i=1}^{N} \alpha_i N(\boldsymbol{x}; \boldsymbol{\mu}_i, \Sigma_i)$$

$$= \sum_{i=1}^{N} \frac{\alpha_i}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T \Sigma_i^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_i)\right)$$

$$\downarrow$$

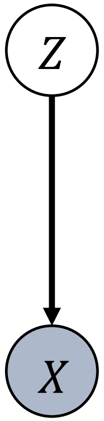
参数： $\boldsymbol{\theta} = \{\alpha_i, \boldsymbol{\mu}_i, \Sigma_i\}_{i=1}^{N}$

- 随机变量$\boldsymbol{x} \in \mathbb{R}^d$
- 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$\mathcal{Z} \in \{1,2,\dots,N\}$符合多项式离散分布
- $\mathcal{Z}$取值为i的概率为：

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

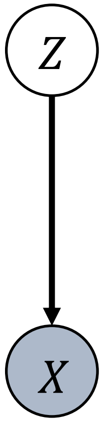
✓ Two-step sampling，从GMM里采样一个样本x

- 从$\mathcal{Z}$中采样，得到一个值i，其中$(1 \le i \le N)$
- 从第i个高斯分布$N(\boldsymbol{\mu}_i, \Sigma_i)$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



Hidden variable ?

Observable variable

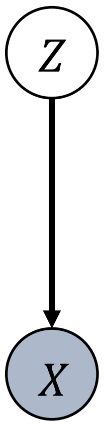
✓ Two-step sampling，从GMM里采样一个样本x

- 从z中采样，得到一个值i，其中$(1 \le i \le N)$
- 从第i个高斯分布$N(\boldsymbol{\mu}_i, \Sigma_i)$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



Hidden variable → $\mathcal{Z} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_M\}$

One-hot vector: $z_i = [0,1,\ldots,0] \in \mathbb{R}^N$

Observable variable → $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M\}$

估计

- $\boldsymbol{z}_i$和 $\boldsymbol{x}_i$一一对应
- N表示GMM中高斯成分的个数

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

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**

  $\boldsymbol{z}_i$和 $\boldsymbol{x}_i$一一对应，如何估计参数θ？

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

$\boldsymbol{z}_i$和 $\boldsymbol{x}_i$一一对应，如何估计参数θ？

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

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

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**

$\boldsymbol{z}_i$和 $\boldsymbol{x}_i$一一对应，如何估计参数θ？

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

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

- 第二步：统计和计算每个高斯分量的参数

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

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

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

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**）：

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

- θ固定（数据分布假设固定），$p(X|\theta)$看作是X的函数，即为概率函数
- X固定（观测数据固定），$\mathcal{L}(\theta|X)$看作是θ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： $\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, \mathcal{L}(\theta|X)$

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（单变量）**：

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

$$\Downarrow$$

最大对数似然估计MLE：

$$\hat{\theta} = \underset{\theta\in\Theta}{\operatorname{argmax}}\, \ln \mathcal{L}(\theta|X) = \underset{\theta\in\Theta}{\operatorname{argmax}} \sum_{j=1}^{M} \ln p(x_j|\theta)$$

求解：

- 求导，令导数为0
- 求解方程，得到最优θ^*

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

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

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

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

$$= \sum_{X} \ln \sum_{\mathcal{Z}} p(\mathcal{Z}|\alpha)\, p(X|\mathcal{Z}; \boldsymbol{\mu}, \Sigma)$$

$$= \sum_{X} \ln \sum_{\mathcal{Z}} p(X, \mathcal{Z}|\theta)$$

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

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

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）

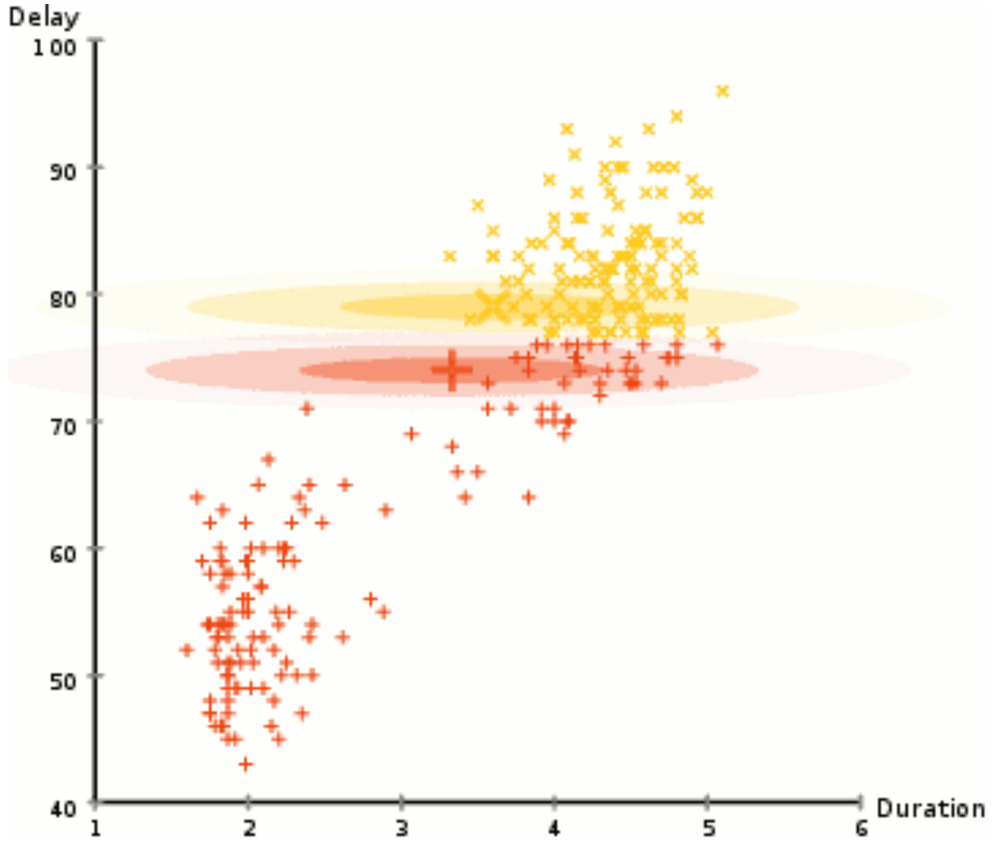
✓ **核心思想**:

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）

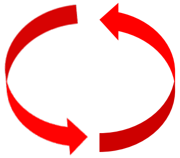


Wiki: EM clustering of Old Faithful eruption data

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$\boldsymbol{\theta}^{(t)}$，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$\boldsymbol{\theta}^{(t+1)}$

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$\mathcal{Z}$的分布$Q(\mathcal{Z}|\theta)$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：**

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

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

$$= \sum_{X} \ln E_{Q}[\frac{p(X, \mathcal{Z}|\theta)}{Q(\mathcal{Z}|\theta)}]$$

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$\Sigma_{x \in S_X} |g(x)| p_X(x) < \infty$，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:

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

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

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

期望最大化(EM)算法

❑ EM优化分析

✓ **高斯混合模型：**

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

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

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

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

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

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

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

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

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

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

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

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

固定θ，即为Z的后验概率

期望最大化(EM)算法

❑ EM优化分析

✓ **高斯混合模型：**

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

下界

固定θ，计算$Q(\mathcal{Z}|\theta) = p(\mathcal{Z}|X,\theta)$，就可以得到$\ell(\theta)$的下界 → E-Step

✓ 然后继续优化这个下界

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

$$= \underset{\theta\in\Theta}{\operatorname{argmax}} \sum_{X}\sum_{\mathcal{Z}} p(\mathcal{Z}|X,\theta)\ln\frac{p(X,\mathcal{Z}|\theta)}{p(\mathcal{Z}|X,\theta)}$$ → M-Step

期望最大化(EM)算法

❑ 算法流程

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

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

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

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

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

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x_j来自第i个高斯分布的期望：**

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

其中， $1 \le i \le N, 1 \le j \le M$

z_{ij}取值为0或者1； $z_{ij} = 1$，当且仅当$\boldsymbol{x}_j$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$\theta^{(t+1)}$：**

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

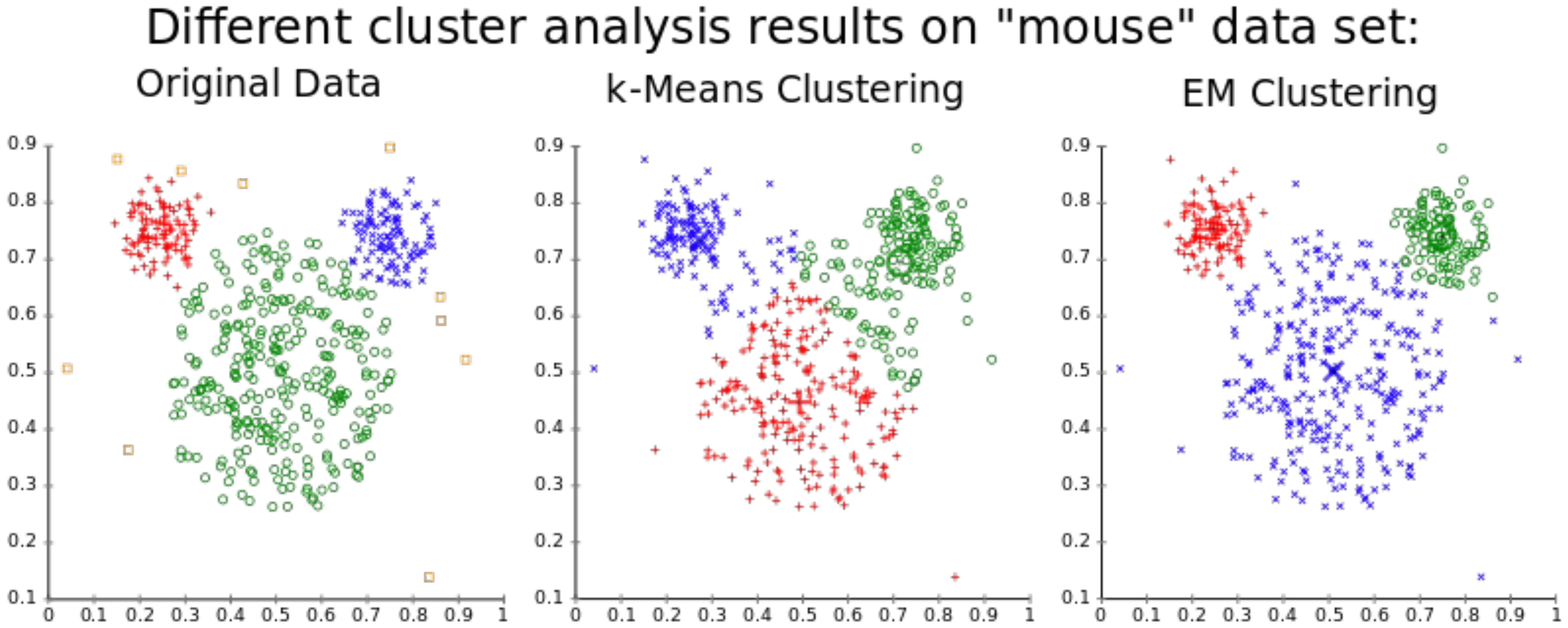
$$\boldsymbol{\mu}_i^{(t+1)} = \frac{\sum_{j=1}^{M} \gamma_{ij} \boldsymbol{x}_j}{m_i} \, ,$$

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

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

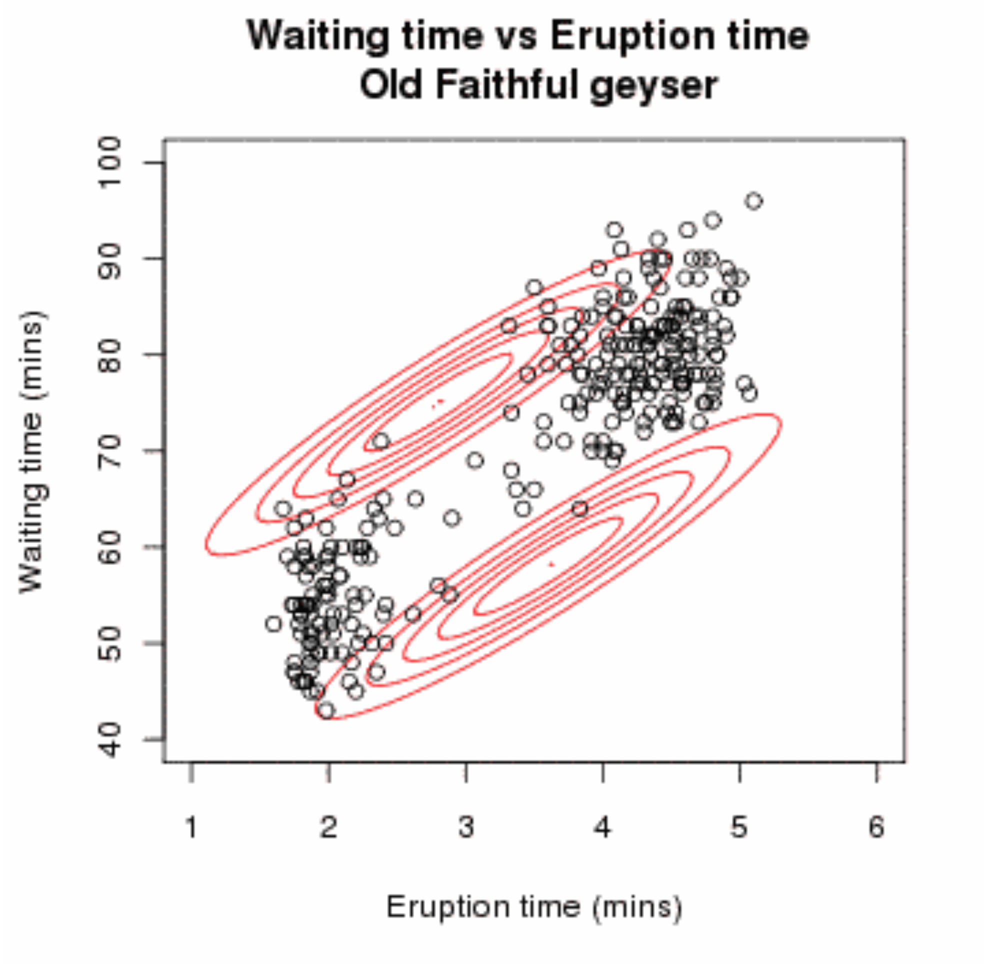
期望最大化(EM)算法

- 应用



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：

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

- 样本$\boldsymbol{x}_i$是可观测变量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θ
- 聚类准则函数看作θ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$\mu_1,\ldots\ \mu_k$
  - 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

谢 谢！

联系方式：liwenbin@nju.edu.cn

更多信息：https://cs.nju.edu.cn/liwenbin