聚类习题

浙江大学

赵洲

聚类任务题目

- 聚类任务中，样本的簇标记是事先已知的，聚类过程只是将样本分配到相应的簇中。

聚类任务题目

- **答案：错误。**
- ● 分析：聚类任务的核心特点是无监督学习。样本的簇标记在聚类开始时是未知的，聚类算法通过样本间的相似性或距离，自动将样本划分到不同的簇中，而不是依据事先定义的簇标记进行分配。与此相反，监督学习则需要事先知道样本的类别标签。因此，题干表述与聚类任务的定义相悖。

性能度量题目

- Jaccard 系数（JC）是聚类性能度量的内部指标，其值越大表示聚类结果越好。

性能度量题目

- **答案：错误。**
- **• 分析：Jaccard 系数是一种外部指标**，用于评估聚类结果与真实分簇（或人工标注）的相似性。外部指标需要先验的真实标签作为对比标准。内部指标则不需要真实标签，而是基于簇内和簇间的样本关系度量聚类质量，例如轮廓系数（Silhouette Coefficient）。因此，题干中将 Jaccard 系数归为内部指标是错误的

$$\mathrm{JC} = \frac{a}{a+b+c}$$

原型聚类题目

- k 均值算法中，初始均值向量的选择对最终聚类结果没有影响。

原型聚类题目

- **答案：错误。**

- **分析**：k 均值算法的核心步骤包括初始中心点的选择和样本到中心点的分配。初始均值向量不同可能导致算法陷入局部最优，从而产生不同的聚类结果。常见的方法包括多次运行 k 均值并选择最优结果，或使用改进算法（如 k-means++）优化初始点选择。

距离计算题目

- 闵可夫斯基距离在任何情况下都满足距离度量的直递性。

距离计算题目

- **答案：错误。**

- • 分析闵可夫斯基距离是一种广义的距离度量，其中参数 p 控制距离的类型。当 p >= 1 时，闵可夫斯基距离满足三角不等式，即具有距离的直递性；但当 0 <= p < 1 时，三角不等式不成立，因此此时不满足距离的直递性。例如，当 p=0.5 时，计算结果可能违背直递性。

聚类任务题目

- - 以下关于聚类任务的描述，正确的是（ ）
- A. 聚类只能作为单独过程，不能用于其他学习任务的前驱过程
- B. 聚类试图将数据集划分为相交的子集，每个子集称为一个簇
- C. 聚类过程能自动形成簇结构，簇所对应的概念语义需使用者把握和命名
- D. 聚类任务中训练样本必须有标记信息

聚类任务题目

- - 以下关于聚类任务的描述，正确的是（C）
- A. 聚类只能作为单独过程，不能用于其他学习任务的前驱过程
- B. 聚类试图将数据集划分为相交的子集，每个子集称为一个簇
- C. 聚类过程能自动形成簇结构，簇所对应的概念语义需使用者把握和命名
- D. 聚类任务中训练样本必须有标记信息

聚类任务题目

- A.聚类只能作为单独过程，不能用于其他学习任务的前驱过程

- 错误。聚类不仅可以作为独立的数据分析方法，用于探索数据的内部结构，还可以作为监督学习的前驱过程，例如在半监督学习中使用聚类结果生成伪标签。

聚类任务题目

- B. 聚类试图将数据集划分为相交的子集，每个子集称为一个簇

- **错误。聚类试图将数据划分为不相交**的子集（硬聚类）或部分相交的子集（软聚类），题干中的“相交”表述不准确。

聚类任务题目

- C. 聚类过程能自动形成簇结构，簇所对应的概念语义需使用者把握和命名

- 正确。聚类的目标是根据数据间的相似性自动形成簇结构，但簇的语义信息通常需要使用者根据领域知识理解和命名。

聚类任务题目

- D. 聚类任务中训练样本必须有标记信息

- 错误。聚类是一种无监督学习方法，训练样本无需标记信息。
- 聚类的定义：聚类是一种无监督学习任务，其目标是根据数据内部的特征和相似性将样本自动划分为不同的簇，而不需要事先知道样本的标签

性能度量题目

- - 以下哪个性能度量指标的值越小表示聚类结果越好?（ ）
- A. Rand指数（RI）
- B. FM指数（FMI）
- C. Dunn指数（DI）
- D. DB指数（DBI）

性能度量题目

- - 以下哪个性能度量指标的值越小表示聚类结果越好?（ D）
- A. Rand指数（RI）
- B. FM指数（FMI）
- C. Dunn指数（DI）
- D. DB指数（DBI）

性能度量题目

- A. Rand指数（RI）

- 错误。Rand 指数（RI）的值越大表示聚类结果越好，因为它衡量聚类结果和真实分簇间的一致性。

$$\mathrm{RI} = \frac{2(a+d)}{m(m-1)}$$

性能度量题目

- B. FM指数（FMI）

- 错误。FM 指数（FMI）越大越好，同样是衡量聚类结果与真实分簇间一致性的指标

$$\mathrm{FMI} = \sqrt{\frac{a}{a+b} \cdot \frac{a}{a+c}}$$

性能度量题目

- C. Dunn指数（DI）

- 错误。Dunn 指数（DI）越大越好，它评估的是簇的紧密性和分离性，值越大说明聚类效果越好。

$$\mathrm{DI} = \min_{1 \leqslant i \leqslant k} \left\{ \min_{j \neq i} \left(\frac{d_{\min}(C_i, C_j)}{\max_{1 \leqslant l \leqslant k} \operatorname{diam}(C_l)} \right) \right\}$$

性能度量题目

- D. DB 指数（DBI）

- 正确。DB 指数（DBI）是内部指标，衡量的是簇的分离度和紧密度的比值，值越小表示簇内更紧密且簇间更分离，因此 DBI 越小越好。

$$\mathrm{DBI} = \frac{1}{k}\sum_{i=1}^{k}\max_{j\neq i}\left(\frac{\mathrm{avg}(C_i)+\mathrm{avg}(C_j)}{d_{\mathrm{cen}}(\boldsymbol{\mu}_i,\boldsymbol{\mu}_j)}\right)$$

距离计算题目

- 对于定义域为{1，2，3}的离散属性，以下说法正确的是（ ）
- A. 它是无序属性，不能直接在属性值上计算距离
- B. 它是有序属性，能直接在属性值上计算距离，且“1”与“2”距离和“2”与“3”距离相等
- C. 它是有序属性，能直接在属性值上计算距离，且“1”与“2”距离小于“1”与“3”距离
- D. 它与连续属性性质完全不同，不能用任何距离计算方法处理

距离计算题目

- 对于定义域为{1，2，3}的离散属性，以下说法正确的是（C）
- A. 它是无序属性，不能直接在属性值上计算距离
- B. 它是有序属性，能直接在属性值上计算距离，且“1”与“2”距离和“2”与“3”距离相等
- C. 它是有序属性，能直接在属性值上计算距离，且“1”与“2”距离小于“1”与“3”距离
- D. 它与连续属性性质完全不同，不能用任何距离计算方法处理

距离计算题目

- A. 它是无序属性，不能直接在属性值上计算距离

- A. 错误。定义域为 {1, 2, 3} 的离散属性是有序的（数字具有自然顺序），因此可以直接计算距离。

距离计算题目

- B. 它是有序属性，能直接在属性值上计算距离，且“1”与“2”距离和“2”与“3”距离相等

- B. 错误。虽然它是有序属性，但距离的计算应该反映其有序性，例如“1”与“3”之间的距离应大于“1”与“2”的距离。

距离计算题目

- C. 它是有序属性，能直接在属性值上计算距离，且“1”与“2”距离小于“1”与“3”距离

- C. 正确。题干描述准确，有序属性允许直接在属性值上计算距离，且“1”与“2”的距离小于“1”与“3”的距离。

距离计算题目

- D. 它与连续属性性质完全不同，不能用任何距离计算方法处理

- D. 错误。尽管离散属性与连续属性有所区别，但有序离散属性仍然可以通过数值直接计算距离。

原型聚类题目

- 以下哪种原型聚类算法假设数据样本带有类别标记？（ ）
- A. k均值算法
- B. 学习向量量化（LVQ）
- C. 高斯混合聚类
- D. 以上都不是

原型聚类题目

- 以下哪种原型聚类算法假设数据样本带有类别标记?（B）
- A. k均值算法
- B. 学习向量量化（LVQ）
- C. 高斯混合聚类
- D. 以上都不是

原型聚类题目

- A. k均值算法

- A. **错误。**k-均值是一种无监督算法，不假设数据有类别标记。
- K-means 是一种无监督学习算法，用于将数据集划分为个簇，其目标是通过迭代优化，将每个样本分配到与其最近的簇，使得簇内样本之间的差异最小化，同时簇间差异最大化；具体过程包括随机初始化 个簇的质心，然后重复样本分配和质心更新的步骤，直到质心稳定或达到预设条件，其最终结果是 个由质心及其对应样本组成的簇。

原型聚类题目

- B. 学习向量量化（LVQ）

- B. 正确。学习向量量化（Learning Vector Quantization, LVQ）是一种基于原型的**监督学习算法**，用于分类任务，其核心思想是通过定义每个类别的原型向量来表示类别特征，并利用已标注的训练样本迭代优化这些原型向量的位置，使得它们更接近属于该类别的样本，同时远离其他类别的样本，从而实现分类；LVQ 的训练过程包括初始化原型向量、根据样本的类别调整最接近的原型向量的位置，直到收敛或达到预设条件，其结果是一组经过优化的原型向量，用于对新样本进行分类。

原型聚类题目

- C. 高斯混合聚类

- C. 错误。高斯混合聚类是一种无监督算法，不需要类别标记。
- 高斯混合聚类（Gaussian Mixture Clustering, GMM）是一种基于概率模型的无监督学习算法，其假设数据由多个高斯分布组成，每个分布代表一个簇，通过估计这些高斯分布的参数（包括均值、协方差矩阵和混合系数）来描述数据分布；具体方法通常采用期望最大化（EM）算法，迭代优化数据属于每个高斯分布的概率分布和模型参数，最终实现数据的软聚类，每个样本根据概率被分配到一个或多个簇中，适用于处理复杂形状的簇和重叠分布的数据集。

密度聚类题目

- DBSCAN 算法中，所有样本都一定会被划分到某个聚类簇中。

密度聚类题目

- **答案：错误。**

- • 分析：DBSCAN 根据样本密度分配簇结构，但对于密度过低（例如未达到 MinPts）的样本，会将其标记为噪声。噪声样本不属于任何簇，因此题干表述错误。
- DBSCAN（Density-Based Spatial Clustering of Applications with Noise）是一种基于密度的聚类算法，其核心思想是通过区域内样本点的密度分布识别聚类簇，适合处理任意形状的簇和含有噪声的数据集。DBSCAN 定义了两个关键参数： （邻域半径）和 （最小邻域点数）。算法通过判断某点在其 邻域内是否至少包含 个点来确定其为核心点、边界点或噪声点；核心点可以扩展形成簇，而边界点是邻近核心点的点，无法直接扩展簇；噪声点则不属于任何簇。DBSCAN 不需要预先指定簇的数量，能有效处理密度不均的数据，但对参数选择较为敏感。

密度聚类题目

- - 在DBSCAN算法中，若样本x的\epsilon邻域至少包含MinPts个样本，则x是（ ）
- A. 噪声样本
- B. 核心对象
- C. 由其他核心对象密度直达的样本
- D. 密度相连的样本

密度聚类题目

- - 在DBSCAN算法中，若样本x的\epsilon邻域至少包含MinPts个样本，则x是（ B ）
- A. 噪声样本
- B. 核心对象
- C. 由其他核心对象密度直达的样本
- D. 密度相连的样本

密度聚类题目

- A. 噪声样本
- • **错误。**
- • 定义：噪声样本是指不属于任何簇的样本。若一个样本 x 的 \epsilon 邻域中包含的样本数小于 MinPts，且 x 也不是密度直达其他核心点的样本，则 x 被标记为噪声。
- • 分析：题目中明确指出 x 的 \epsilon 邻域至少包含 MinPts 个样本，这已经满足了核心点的定义条件，因此 x 不可能是噪声样本。
- • 总结：噪声样本不符合题目描述，选项 A 错误。

密度聚类题目

- B. 核心对象
- **正确。**
- 定义：核心对象是指 \epsilon 邻域中至少包含 MinPts 个样本的点。核心对象是簇形成的基础，因为核心对象可以扩展其密度范围将其他点归入簇。
- 分析：题目中说明 x 的 \epsilon 邻域至少包含 MinPts 个样本，这完全符合核心对象的定义。因此，x 是一个核心对象。
- 总结：符合题意，选项 B 正确。

密度聚类题目

- C. 由其他核心对象密度直达的样本
- • **错误。**
- • **定义：**“由其他核心对象密度直达的样本”是指边界对象（Border Object）。边界对象的 \epsilon 邻域样本数少于 MinPts，但它位于某个核心对象的 \epsilon 邻域范围内，通过密度直达关系连接到核心对象，从而被归属于该核心对象所在的簇。
- • 分析：题目中明确指出 x 的 \epsilon 邻域至少包含 MinPts 个样本，这意味着 x 是核心对象，而不是边界对象。因此，选项 C 的描述与题意不符。
- • 总结：边界对象的定义与题目条件不符，选项 C 错误

密度聚类题目

- D. 密度相连的样本
- • 错误。
- • 定义：密度相连（Density Connected）是描述两个点间关系的概念，而不是点的属性。具体而言，点 x 和点 y 被定义为密度相连，是指存在一个点链 z_1, z_2, …, z_n，使得链上相邻的点是密度直达的，且链的一端是核心对象。
- • 分析：题目中讨论的是点 x 的属性（核心对象或其他角色），而不是描述 x 与其他点的关系。因此，将“密度相连”用来描述 x 的属性是错误的。
- • 总结：密度相连是点之间的关系，不是点的分类属性，选项 D 错误。

层次聚类题目

- AGNES算法采用自顶向下的分拆策略形成聚类结构。（ ）

层次聚类题目

- 答案：错误。AGNES算法采用自底向上的聚合策略形成聚类结构。
- AGNES（Agglomerative Nesting）算法是一种基于自底向上策略的层次聚类方法，其核心思想是从每个样本开始，将每个样本初始视为一个独立的簇，然后通过迭代过程逐步合并距离最近的两个簇，直到所有样本合并为一个簇或达到预设的簇数；簇间距离的计算可以采用单链接（最小距离）、全链接（最大距离）或均链接（平均距离）等方法，不同的距离定义会影响聚类结果，AGNES 通过构建聚类树（dendrogram）直观地表示聚类过程及结果的层次结构。

层次聚类题目

- - 在AGNES算法中，当聚类簇距离由最小距离计算时，该算法被称为（ ）
- A. 单链接算法
- B. 全链接算法
- C. 均链接算法
- D. 以上都不是

层次聚类题目

- - 在AGNES算法中，当聚类簇距离由最小距离计算时，该算法被称为（ A ）
- A. 单链接算法
- B. 全链接算法
- C. 均链接算法
- D. 以上都不是

层次聚类题目

- A. 单链接算法

- A. 正确。单链接算法使用最小距离定义簇间距离。
- 单链接算法是一种用于层次聚类的算法，其定义簇间的距离为两个簇中所有样本之间的距离中最短的距离（即最小距离），在每次聚类步骤中优先合并最近的两个簇。它的核心思想是将两个簇视为相连的，只要簇中有两个样本之间的距离足够近。这种算法特别适合发现链状或延展性的簇结构，但由于只考虑最近的点，它对噪声点较为敏感，容易导致“链式效应”，即将原本不属于同一簇的样本通过中间样本连接起来。

层次聚类题目

- B. 全链接算法

- B. 错误。全链接算法定义为簇间距离取最大值。
- 全链接算法是一种用于层次聚类的算法，其定义簇间的距离为两个簇中所有样本之间的距离中最远的距离（即最大距离），在每次聚类步骤中优先合并距离最远边界最近的两个簇。它的核心思想是最大化簇内的紧密性，因此倾向于生成更小、更紧密的簇结构。这种算法对于链状结构的簇表现较差，但对离群点和噪声的鲁棒性较强，因为它避免了单一近距离点的影响。

层次聚类题目

- C. 均链接算法

- C. 错误。均链接算法定义为簇间平均距离。
- 均链接算法是一种用于层次聚类的算法，其定义簇间的距离为两个簇中所有样本之间的两两距离的平均值，在每次聚类步骤中合并距离平均值最小的两个簇。它的核心思想是综合考虑簇中所有点之间的关系，而不仅仅是最近点或最远点的关系。均链接算法是一种平衡方法，既能处理较为分散的簇，也能适应不同形状和大小的簇结构。

计算题

- 假设我们有三个聚类质心，mu1=[1,2]，mu2=[-3,0]和mu3=[4,2]。此外，我们还有一个训练示例x（i）=[3,1]。
- 在聚类分配步骤之后，c（i）将是哪个质心?

计算题

- 假设我们有三个聚类质心，mu1=[1,2]，mu2=[-3,0]和mu3=[4,2]。此外，我们还有一个训练示例x（i）=[3,1]。
- 在聚类分配步骤之后，c（i）将是什么?
- 解:
- 这三个质心是:
- $\mu_1 = [1,2], \mu_2 = [-3,0], \mu_3 = [4,2]$
- 任务是将x(i)=[3,1]分配给最近的质心，这是通过计算x(i)与每个质心之间的欧几里得距离来完成的。

计算题

- 步骤 1：计算从 x (i) 到每个质心的距离。

- $d(x(i),\mu_1)=\sqrt{(3-1)^2+(1-2)^2}=\sqrt{2^2+(-1)^2}=\sqrt{4+1}=\sqrt{5}\approx 2.24$ (1)

- $d(x(i),\mu_2)=\sqrt{(3-(-3))^2+(1-0)^2}=\sqrt{6^2+1^2}=\sqrt{36+1}=\sqrt{37}\approx 6.08$ (2)

- $d(x(i),\mu_3)=\sqrt{(3-4)^2+(1-2)^2}=$

  $\sqrt{(-1)^2+(-1)^2}=\sqrt{1+1}=\sqrt{2}\approx 1.41.$ (3)

计算题

- 步骤2：根据欧式距离，确定最近的质心。
- $d(x(i), \mu_1) \approx 2.24$
- $d(x(i), \mu_2) \approx 6.08$
- $d(x(i), \mu_3) \approx 1.41$
- 最近的质心是 μ_3
- c（i）=3

K-means计算题

⑩ 假设有样本集$D = \{x_1 = (1,2), x_2 = (2,1), x_3 = (3,3), x_4 = (4,2), x_5 = (5,1)\}$，使用$k$均值算法（$k = 2$）对其进行聚类，距离计算使用欧氏距离。初始均值向量随机选择为$\mu_1 = x_1 = (1,2)$，$\mu_2 = x_3 = (3,3)$。请计算第一次迭代后的簇划分和新的均值向量。

K-means计算题

- 计算样本到均值向量的距离并分配簇
- 计算x_1到μ_1和μ_2的距离：
- $d_{11} = \| x_1 - \mu_1 \|_2 = \sqrt{(1-1)^2 + (2-2)^2} = 0$
- $d_{12} = \| x_1 - \mu_2 \|_2 = \sqrt{(1-3)^2 + (2-3)^2} = \sqrt{5}$
- 计算x_2到μ_1和μ_2的距离：
- $d_{21} = \| x_2 - \mu_1 \|_2 = \sqrt{(2-1)^2 + (1-2)^2} = \sqrt{2}$

K-means计算题

- 计算样本到均值向量的距离并分配簇
- 计算x_3到μ_1和μ_2的距离：
- $d_{31} = \| x_3 - \mu_1 \|_2 = \sqrt{(3-1)^2 + (3-2)^2} = \sqrt{5}$
- $d_{32} = \| x_3 - \mu_2 \|_2 = \sqrt{(3-3)^2 + (3-3)^2} = 0$
- 计算x_4到μ_1和μ_2的距离：
- $d_{41} = \| x_4 - \mu_1 \|_2 = \sqrt{(4-1)^2 + (2-2)^2} = 3$

K-means计算题

- 计算样本到均值向量的距离并分配簇
- 计算x_5到μ_1和μ_2的距离：
- $d_{51} = \| x_5 - \mu_1 \|_2 = \sqrt{(5-1)^2 + (1-2)^2} = \sqrt{17}$
- $d_{52} = \| x_5 - \mu_2 \|_2 = \sqrt{(5-3)^2 + (1-3)^2} = 2\sqrt{2}$

K-means计算题

- 根据距离分配簇：
- x_1距离μ_1更近，所以x_1划入C_1。
- x_2距离μ_1更近，所以x_2划入C_1。
- x_3距离μ_2更近，所以x_3划入C_2。
- x_4距离μ_2更近，所以x_4划入C_2。
- x_5距离μ_2更近，所以x_5划入C_2。
- 第一次迭代后的簇划分为$C_1 = \{x_1, x_2\}$，$C_2 =$

K-means计算题

- 计算新的均值向量
- 计算C_1的新均值向量μ_1'：
- $\mu_1' = \frac{1}{|C_1|}\sum_{x \in C_1} x = \frac{1}{2}((1,2)+(2,1)) = \left(\frac{3}{2},\frac{3}{2}\right)$
- 计算C_2的新均值向量μ_2'：
- $\mu_2' = \frac{1}{|C_2|}\sum_{x \in C_2} x = \frac{1}{3}((3,3)+(4,2)+(5,1)) = \left(\frac{12}{3},\frac{6}{3}\right) = (4,2)$

K-means计算题

⑩ 综上，第一次迭代后的簇划分为$C_1 = \{x_1, x_2\}$，$C_2 = \{x_3, x_4, x_5\}$，新的均值向量为$\mu_1' = \left(\frac{3}{2}, \frac{3}{2}\right)$，$\mu_2' = (4,2$

GMM计算题

- 题目：

  假设样本集 $D = \{x_1 = (1,2), x_2 = (2,1), x_3 = (3,3), x_4 = (4,2), x_5 = (5,1)\}$，使用高斯混合模型（$k = 2$）进行聚类，初始化参数如下：

- $\alpha_1 = \alpha_2 = 0.5$, $\mu_1 = (1,2)$, $\mu_2 = (3,3)$, $\Sigma_1 = \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. (1)

- 请执行一次 EM 算法迭代，并更新模型参数。

GMM计算题

- 1. E步：
- 我们需要计算每个样本 x_i 属于每个簇的后验概率（责任度）：
- $$\gamma_{ik} = \frac{\alpha_k \cdot \mathcal{N}(x_i|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \alpha_j \cdot \mathcal{N}(x_i|\mu_j, \Sigma_j)}, \qquad (2)$$
- 其中 $\mathcal{N}(x_i|\mu_k, \Sigma_k)$ 是二维高斯分布概率密度函，

$$\mathcal{N}(x|\mu, \Sigma) = \frac{1}{2\pi|\Sigma|^{0.5}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right) \qquad (3)$$

GMM计算题

- 1. E步：
- 计算每个样本的高斯密度值：
- 对于簇 1 和簇 2，分别计算每个样本的密度值：

样本 x_i	$\mathcal{N}(x_i\|\mu_1,\Sigma_1)$	$\mathcal{N}(x_i\|\mu_2,\Sigma_2)$
$x_1=(1,2)$	0.159	0.027
$x_2=(2,1)$	0.183	0.005
$x_3=(3,3)$	0.027	0.159
$x_4=(4,2)$	0.001	0.105
$x_5=(5,1)$	0.00003	0.027

GMM计算题

- 1. E步：

- 计算 γ_{ik} ：

- 对于簇 1 和簇 2，分别计算每个样本的密度值：

$$\gamma_{ik} = \frac{\alpha_k \cdot \mathcal{N}(x_i|\mu_k \Sigma_k')}{\sum_{j=1}^{K} \alpha_j \cdot \mathcal{N}(x_i|\mu_j \Sigma_j')} \quad (4)$$

GMM计算题

- 1. E步：

$$\gamma_{ik} = \frac{a_k \cdot \mathcal{N}(x_i|\mu_k \Sigma_k')}{\sum_{j=1}^{K} a_j \cdot \mathcal{N}(x_i|\mu_j \Sigma_j')}$$

- 计算结果如下：$\gamma = \begin{bmatrix} 0.9241 & 0.0759 \\ 0.8176 & 0.1824 \\ 0.0759 & 0.9241 \\ 0.0293 & 0.9707 \\ 0.0110 & 0.9890 \end{bmatrix}$ (5)

- 每行表示样本在两个簇后验概率，列分别对应簇 1 和簇

GMM计算题

- 2. M步：更新参数

- (1) 更新混合系数 $\boldsymbol{\alpha_k}$：

- 混合系数的更新公式：

- $\alpha_k = \frac{1}{N}\sum_{i=1}^{N} \gamma_{ik}$. (6)

- 计算结果：

- $\alpha_1 = 0.372, \quad \alpha_2 = 0.628$ (7)

GMM计算题

- 2. M步：更新参数

- (2) 更新均值 $\boldsymbol{\mu_k}$：
- 均值的更新公式：

$$\mu_k = \frac{\sum_{i=1}^{N} \gamma_{ik}\, x_i}{\sum_{i=1}^{N} \gamma_{ik}}. \qquad (8)$$

- 计算结果：
- $\mu_1 = [1.593, 1.595], \quad \mu_2 = [3.832, 1.921]$ (9

GMM计算题

- 2. M步：更新参数

- (2) 更新均值 $\boldsymbol{\mu_k}$：
- 均值的更新公式：

$$\mu_k = \frac{\sum_{i=1}^{N} \gamma_{ik} x_i}{\sum_{i=1}^{N} \gamma_{ik}}. \qquad (8)$$

- 计算结果：

$$\mu_1 = [1.593, 1.595], \quad \mu_2 = [3.832, 1.921] \qquad (9)$$

GMM计算题

⑩ 2. M步：更新参数

- (3) 更新协方差矩阵 $\boldsymbol{\Sigma_k}$：
- 协方差矩阵的更新公式：

$$\Sigma_k = \frac{\sum_{i=1}^{N} \gamma_{ik} (x_i - \mu_k)(x_i - \mu_k)^T}{\sum_{i=1}^{N} \gamma_{ik}}. \qquad (10)$$

- 计算结果：

$$\Sigma_1 = \begin{bmatrix} 0.489 & -0.142 \\ -0.142 & 0.323 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 1.030 & -0.506 \end{bmatrix} \qquad (11)$$

End