机器学习讲义 (L19-C) 高斯混合模型

授课教师：王贝伦 / 助教：张嘉琦，黄旭，谈笑，徐浩卿

1 其他划分聚类法

之前课程所介绍的 k-means 算法是一种区别于层次聚类法的思路——划分聚类法。划分法通过预先给定确定的聚类数，不断调整聚类中心达到聚类的目的。除了 k-means 之外，其他形式的划分聚类法也有很多应用，例如 k-medoids、自组织映射、高斯混合模型等。本节简单介绍前两者，高斯混合模型将在之后章节中详细介绍。

1.1 k-medoids 算法

与 k-means 算法十分相似，k-medoids 算法确定数个样本点作为初始聚类中心。迭代中，同样将每个样本点归类至最近的聚类中心。不同的是，k-medoids 算法并不以聚类中样本的均值作为新中心，而是对每个样本点，计算若以其为中心所得到的代价函数（一般定义为样本差异的和，两个样本间的差异可事先用距离矩阵定义），最后选择使代价最小的样本点作为新聚类中心。

由于 k-medoids 算法在每步都使用样本点作为聚类中心，可以避免 k-means 算法中聚类中心脱离样本、受噪音影响大等缺陷。

1.2 自组织映射

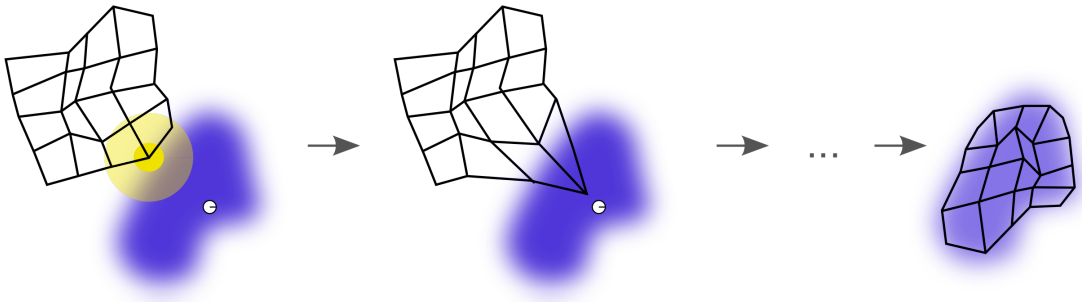

Figure 1: 自组织映射网络节点更新示意

自组织映射设想聚类中心之间呈网状连接。训练时，对一个样本点计算出距离最近的网络节点，将这个节点以及其邻居节点向该样本点移动一小段距离，如同网络被“拖动”一般。在大量样本训练后，整个聚类点网络就基本贴合样本的总体分布。

自组织映射有一个有趣的可视化应用，称音乐之岛。事先将大量音乐利用自组织映射等方法进行聚类，每个数量不同的聚类就形成高低不一的山岛。用户可以在不同岛上寻找相似的歌曲。

Figure 2: 音乐之岛

2 高斯混合模型 (Gaussian Mixture Model)

高斯混合模型 (GMM) 也是一种与 k-means 类似的划分聚类法。实际上，它描述的是样本的总体概率分布。具体来说，GMM 认为，样本是由多个高斯分布产生的样本混合而成的，而单个样本点属于某类高斯分布的概率是一定的。在这种假设下，样本的总体分布实际上就是加权高斯分布：

$$p(x) = \sum_i P(x \in C_i) \frac{1}{(2\pi)^{p/2}|\Sigma_i|^{1/2}} \exp[-\frac{1}{2}(x-\mu_i)^\top \Sigma_i^{-1}(x-\mu_i)] \tag{1}$$

其中 $P(x \in C_i)$ 表示样本点 x 属于类别 C_i 的概率，μ_i, Σ_i 分别为类别 C_i 对应正态分布的期望、协方差。

下图展示了由高斯分布组成 GMM 的大致形态：

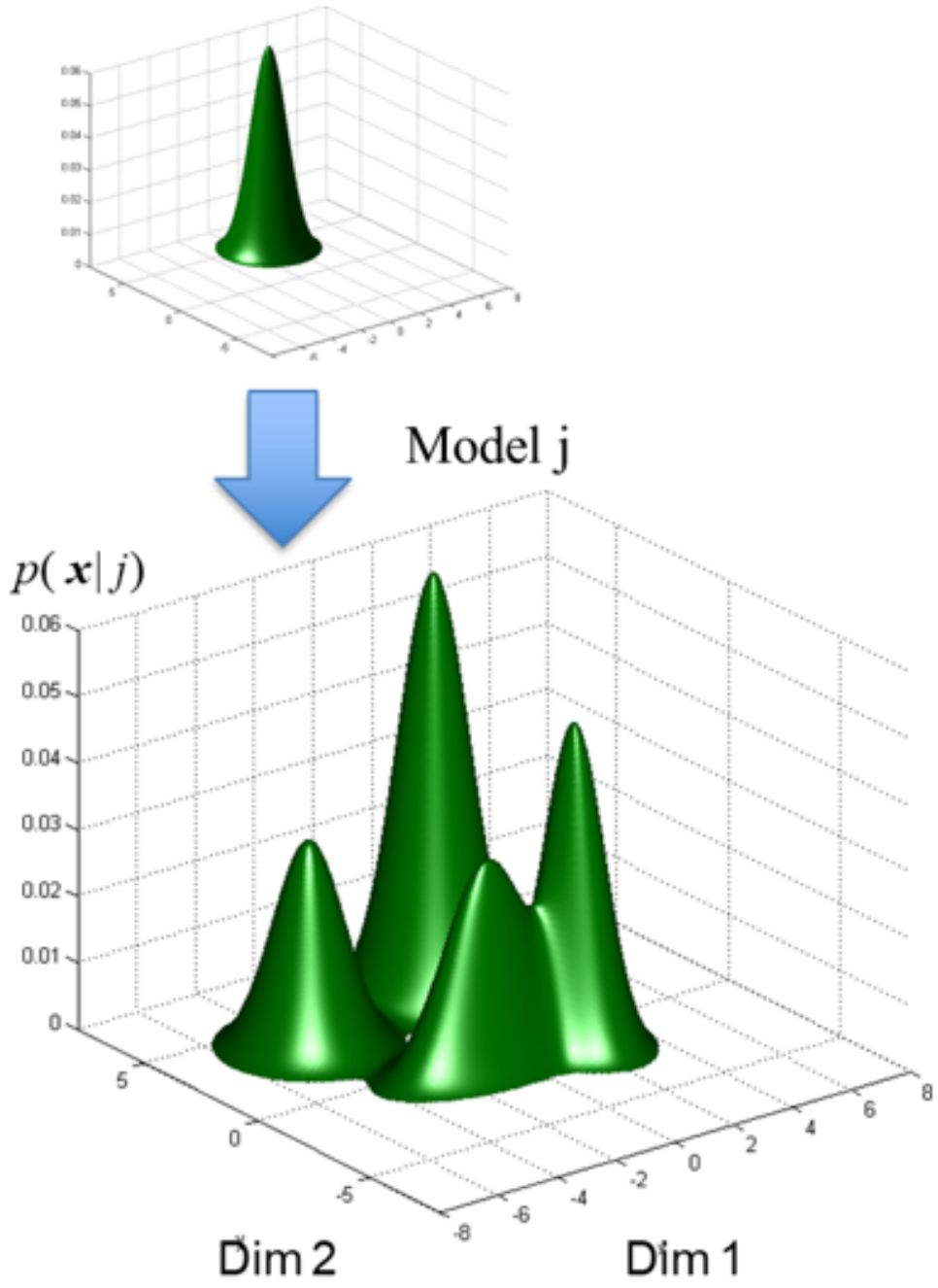

Figure 3: GMM 示意

GMM 应用于聚类问题上时，首先需要给出对各子分布参数的一些假设和初始化。一般情况下，各类初始均值、协方差随机设定。之后使用后文中介绍的 EM 算法进行逐步迭代优化，直到收敛。下图展示了一个含有 2 个聚类的 GMM 的迭代过程。

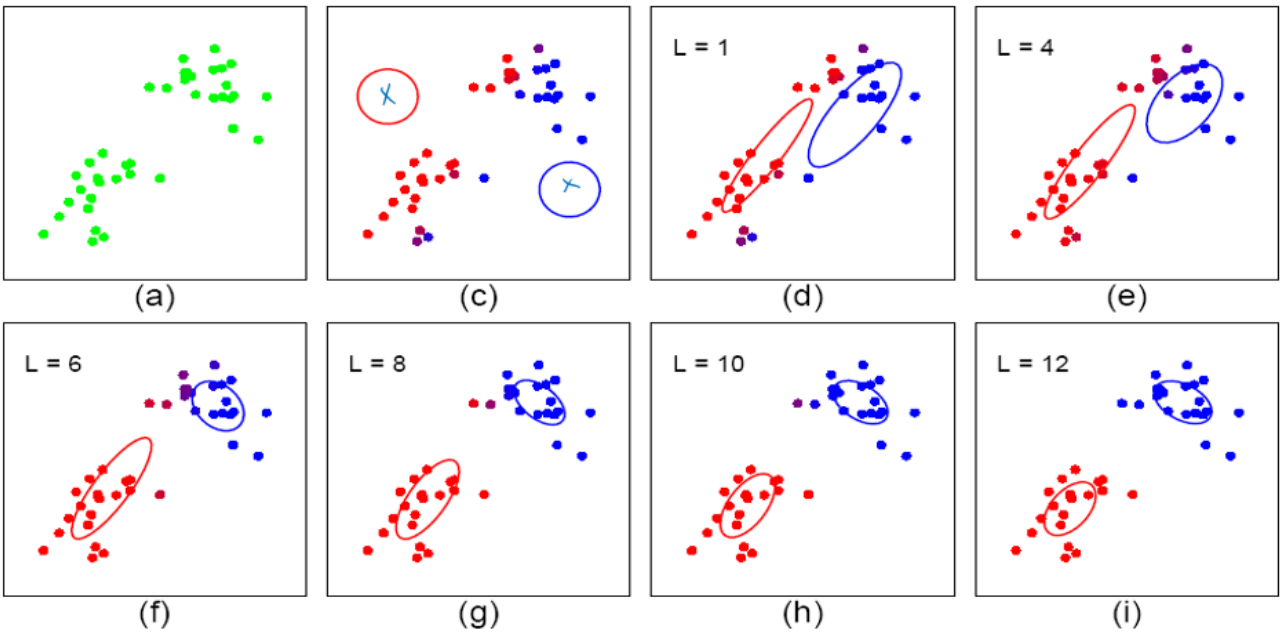

Figure 4: GMM 迭代优化过程示意

3 高斯混合模型的基本算法

3.1 高斯混合模型的假设

不同复杂度的高斯混合模型会对类的均值 μ_i 和协方差矩阵 Σ_i 做出不同的假设。下面给出几种不同的假设条件。

- 假设 1：每个类的均值为 μ_i，协方差矩阵相同且为数量矩阵，即 $\Sigma_i = \sigma^2 I$。这是高斯混合模型最简单的假设条件，模型的概率分布图如图5所示。

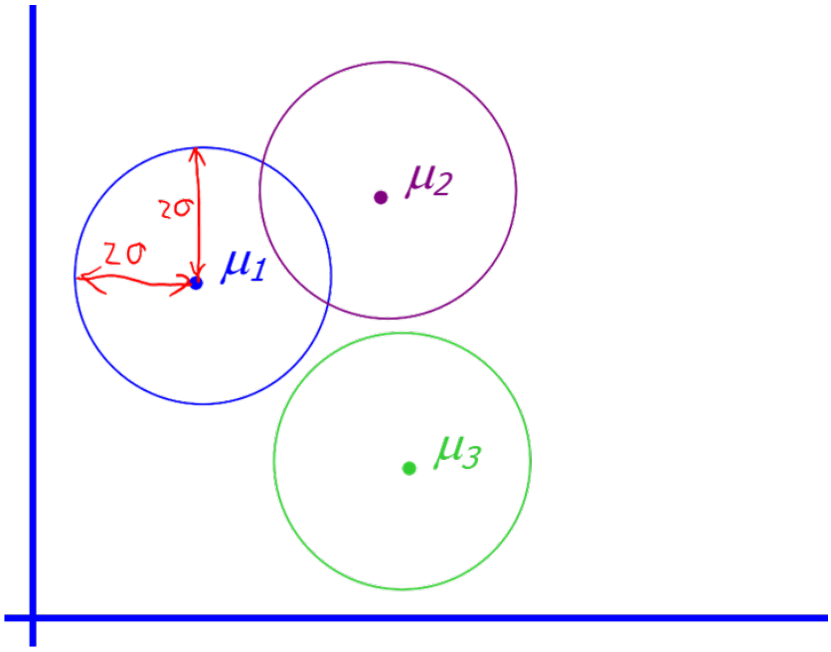

Figure 5: 假设 1 模型的概率分布

- 假设 2：每个类的均值为 μ_i，协方差矩阵相同且为对角矩阵。模型的概率分布图如图6所示。

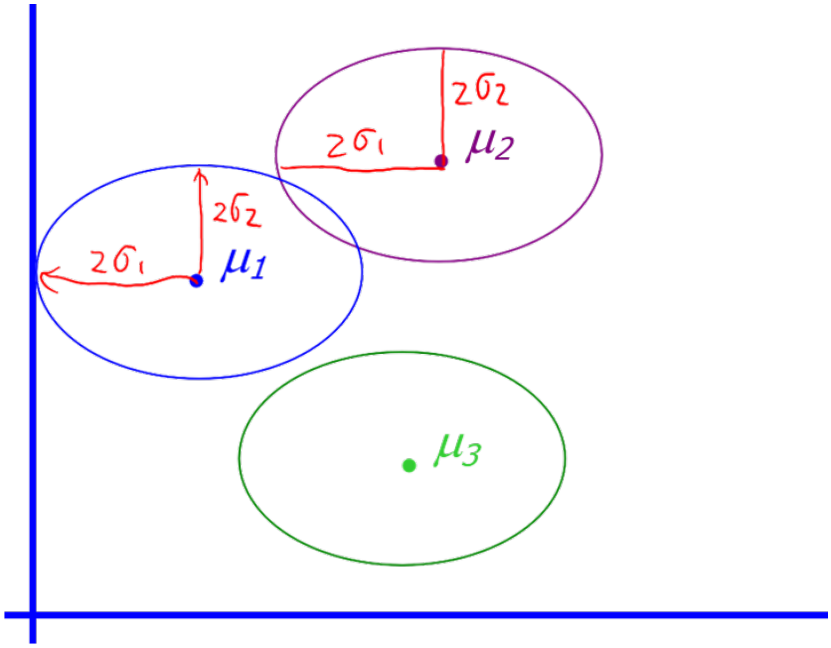

Figure 6: 假设 2 模型的概率分布

- 假设 3：每个类的均值为 μ_i，类与类的协方差矩阵不同，但都为数量矩阵，即 $\Sigma_i = \sigma_i^2 I$。模型的概率分布图如图7所示。

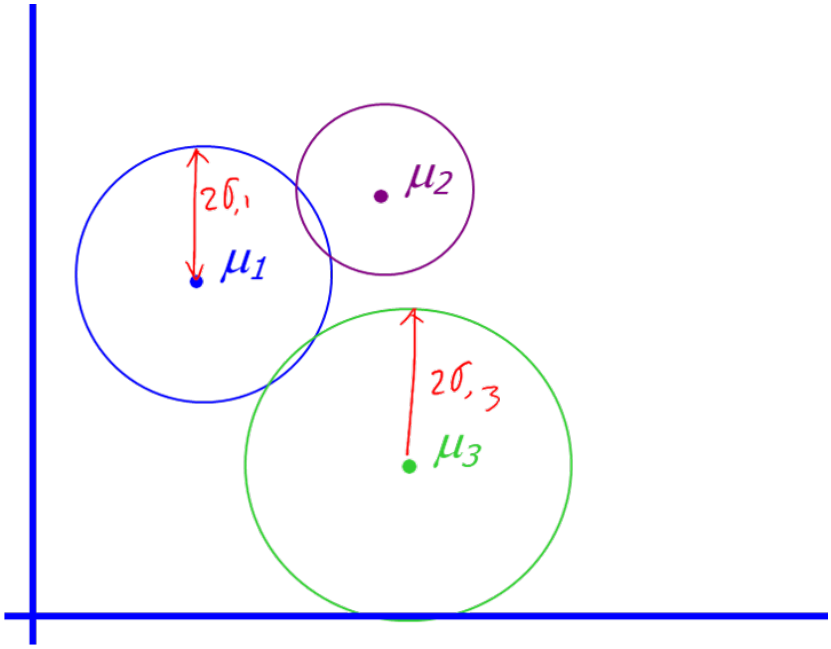

Figure 7: 假设 3 模型的概率分布

- 假设 4：每个类的均值为 μ_i，协方差矩阵相同。这个假设比之前的更加泛化。模型的概率分布图如图8所示。

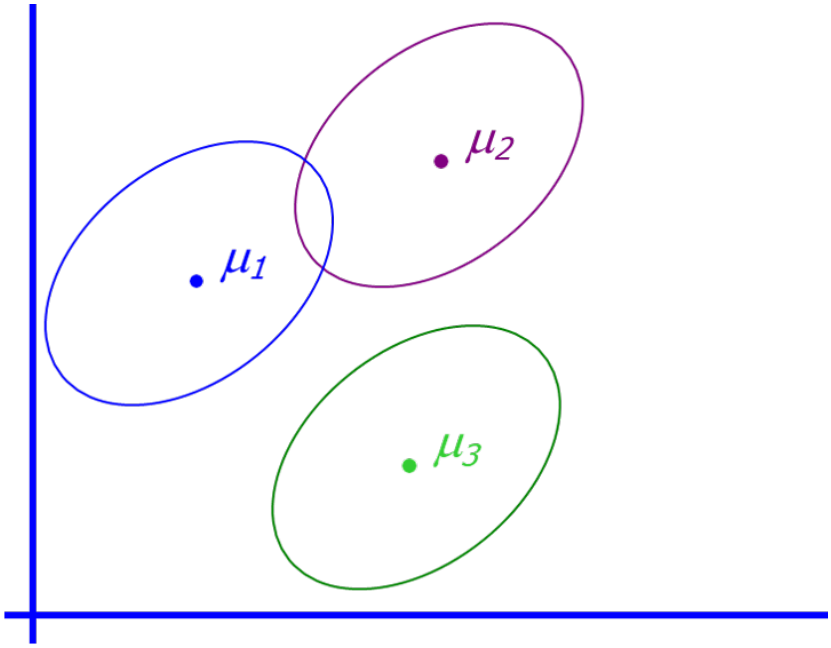

Figure 8: 假设 4 模型的概率分布

- 假设 5：每个类的均值为 μ_i，协方差矩阵为 Σ_i。这个便是泛化的假设，每个类只需满足高斯分布。模型的概率分布图如图9所示。

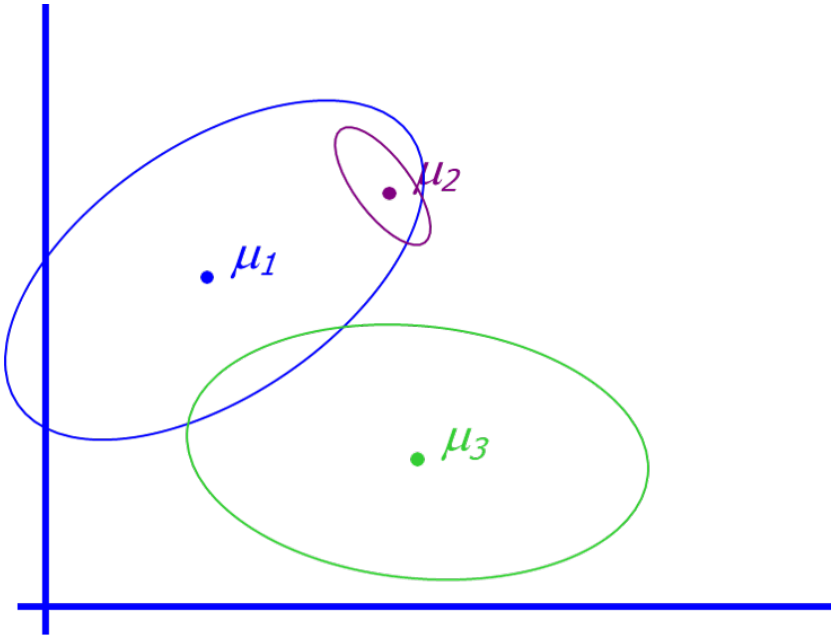

Figure 9: 假设 5 模型的概率分布

3.2 迭代公式

高斯混合模型的迭代公式推导使用的是 EM 算法（Expectation Maximization algorithm）。EM 算法基于极大似然估计，适用于对包含隐变量或缺失数据（如不知道样本的类别）的概率模型进行参数估计。算法的每次迭代过程有期望步（E-step）和最大化步（M-step）交替组成。假设每个类的协方差矩阵都相同，记为 Σ，类别 i 的均值为 μ_i，高斯分布的概率密度函数为

$$N(x|\mu,\Sigma) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)} \tag{2}$$

由之前的推导可得，高斯混合模型的分布为

$$p(x = x_i) = \sum_j p(\mu = \mu_j) N(x_i|\mu_j, \Sigma) \tag{3}$$

在高斯混合模型中的隐变量便是每个样本的类别，我们用 z_{ij} 表示样本 x_i 是否属于类别 j，如果属于则 $z_{ij} = 1$，否则为 0。

E-step 期望步根据初始参数值或者上一次迭代得到的参数值，计算隐变量的期望，可以简单理解为根据参数来计算每个样本分别属于某个类的概率。那么 z_{ij} 的期望为

$$\begin{aligned} E[z_{ij}] &= p(\mu = \mu_j | x = x_i) \\ &= \frac{p(x = x_i | \mu = \mu_j) p(\mu = \mu_j)}{p(x = x_i)} \\ &= \frac{N(x_i | \mu_j, \Sigma) p(\mu = \mu_j)}{\sum_{s=1}^{k} N(x_i | \mu_s, \Sigma) p(\mu = \mu_s)} \end{aligned} \tag{4}$$

M-step 最大化步根据上一步得到的期望，使用极大似然估计更新参数的值。更新方式如下：

$$\mu_j \leftarrow \frac{1}{\sum_{i=1}^{n} E[z_{ij}]} \sum_{i=1}^{n} E[z_{ij}] x_i \tag{5}$$

$$p(\mu = \mu_j) \leftarrow \frac{1}{n} \sum_{i=1}^{n} E[z_{ij}] \tag{6}$$

如果每个类的协方差矩阵不同，协方差矩阵的值也需要迭代更新

$$\Sigma_j \leftarrow \frac{\sum_{i=1}^{n} E[z_{ij}](x_i - \mu_j)(x_i - \mu_j)^T}{\sum_{i=1}^{n} E[z_{ij}]} \tag{7}$$

通过循环期望步和最大化步，迭代更新参数值直到收敛，最终可以学习到较准确的高斯混合模型。

4 高斯混合模型与 K-means 算法的联系

上一章讲到的 K-means 算法其实也可以看作是 EM 算法的特例。K-means 算法的目标函数为

$$\arg\min_{\mu_k, M} \sum_{k=1}^{K} \sum_{i=1}^{n} M_{ik} \|x_i - \mu_k\|_2^2 \tag{8}$$

其中，M_{ik} 可以看作隐变量。在 E-step 中，根据 μ_k 的值更新 M 的值

$$M_{ik} = \begin{cases} 1, & k = \arg\min_k \|x_i - \mu_k\|_2^2 \\ 0, & otherwise \end{cases} \tag{9}$$

在 M-step 中，根据 M 的值更新每个类的均值

$$\mu_k = \frac{\sum_{i=1}^{n} M_{ik} x_i}{\sum_{i=1}^{n} M_{ik}} \tag{10}$$

高斯混合模型中的 EM 算法像是 K-means 算法的“松弛版本”。在 K-means 算法的 E-step 中，M_{ik} 被直接赋值为 0 或 1，而不是期望。在 K-means 算法的 M-step 中，同样是用 0 或 1 作为权重来更新均值。K-means 算法只能学习到圆形分布的类别，而高斯混合模型可以学习到椭圆形分布的类别。

5 高斯混合模型与 K-means 算法的不足

非监督学习并不像看起来这么困难，一个任务有时候会很简单，但有时候也似乎是不可能完成的。高斯混合模型和 K-means 算法都有一些不足之处：

(1) 两种方法都需要计算类别的中心和显式度量距离的方法。当使用特殊的距离度量方式时，类别的中心可能会很难计算。

(2) 两种方法都致力于寻找紧凑的聚类结构，而在一些情况下，连通的聚类结构是更合理的，这时使用基于图的聚类方法会更好。

引用

[1] https://blog.csdn.net/lin_limin/article/details/81048411