第十二章（聚类分析）作业

12.1 假设我们的样本集 $D=\{\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_n\}$ 来自于 k 个高斯分布，即 $p_M(\boldsymbol{x})=\sum_{j=1}^{k}\alpha_j\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu_j},\boldsymbol{\Sigma_j})$，其中 $\alpha_j,\boldsymbol{\mu_j},\boldsymbol{\Sigma_j}$ 是混合高斯分布第j个成分的比例、均值以及协方差矩阵，试推导采用 EM 算法求解混合高斯模型参数的过程。

（1） 根据贝叶斯公式，求后验期望，写出需要最大化的函数形式。

（2） 最大化（1）中的函数，写出模型参数 α_j、$\boldsymbol{\mu_j}$、$\boldsymbol{\Sigma_j}$ 的更新公式。

答案：

（1） 对于第 i 个样本 $\mathbf{x}_i$，由贝叶斯公式可得：

$$\gamma_{ij}=p_M(c_i=j\,|\,\mathbf{x}_i)=\frac{P(c_i=j)p_M(\mathbf{x}_i\,|\,y_i=j)}{p_M(\mathbf{x}_i)}=\frac{\alpha_j\,\mathrm{N}(\mathbf{x}_i\,|\,\mu_j,\Sigma_j)}{\sum_{j=1}^{k}\alpha_j\,\mathrm{N}(\mathbf{x}_i\,|\,\mu_j,\Sigma_j)}$$

对数似然函数为：

$$l(\mathrm{D})=\ln(\prod_{i=1}^{n}p_M(\mathbf{x}_i))=\sum_{i=1}^{n}\ln(\sum_{j=1}^{k}\alpha_j\,\mathrm{N}(\mathbf{x}_i\,|\,\mu_j,\Sigma_j))$$

且约束为 $\alpha_i\geq 0;\ \sum_{i=1}^{k}\alpha_i=1$

（2） 使用拉格朗日法，得到

$$L(\mathrm{D})=\mathrm{l(D)}+\lambda(\sum_{i=1}^{k}\alpha_i-1)$$

对概率求偏导：

$$\frac{\partial L}{\partial\alpha_j}=\sum_{i=1}^{n}\frac{\alpha_j\,\mathrm{N}(\mathbf{x}_i\,|\,\mu_j,\Sigma_j)}{\sum_{j=1}^{k}\alpha_j\,\mathrm{N}(\mathbf{x}_i\,|\,\mu_j,\Sigma_j)}+\lambda=\sum_{i=1}^{n}\frac{\gamma_{ij}}{\alpha_j}+\lambda$$

令 $\frac{\partial L}{\partial\alpha_j}=0$，得 $\alpha_j=-\frac{1}{\lambda}\sum_{i=1}^{n}\gamma_{ij}$

等式两边同时对 j 进行加和，得 $\lambda=-n$

因此，$\alpha_j=\frac{1}{n}\sum_{i=1}^{n}\gamma_{ij}$

对均值求偏导：

$$\frac{\partial l(D)}{\partial \mu_j} = -\frac{1}{2}\sum_{i=1}^{n}\gamma_{ij}(x_i^T\Sigma_j^{-1}x_i - x_i^T\Sigma_j^{-1}\mu_j - \mu_j^T\Sigma_j^{-1}x_i + \mu_j^T\Sigma_j^{-1}\mu_j)$$

$$= \frac{1}{2}\sum_{i=1}^{n}\gamma_{ij}((x_i^T\Sigma_j^{-1})^T + \Sigma_j^{-1}x_i - ((\Sigma_j^{-1})^T + \Sigma_j^{-1})\mu_j)$$

$$= \sum_{i=1}^{n}\gamma_{ij}(\Sigma_j^{-1}x_i - \Sigma_j^{-1}\mu_j)$$

令 $\frac{\partial l(D)}{\partial \mu_j} = 0$，得 $\mu_j = \frac{\sum_{i=1}^{n}\gamma_{ij}x_i}{\sum_{i=1}^{n}\gamma_{ij}}$

对方差求偏导：

$$\frac{\partial l(D)}{\partial \Sigma_j} = \frac{1}{2}\sum_{i=1}^{n}\gamma_{ij}(\Sigma_j - (x_i - \mu_j)(x_i - \mu_j)^T)$$

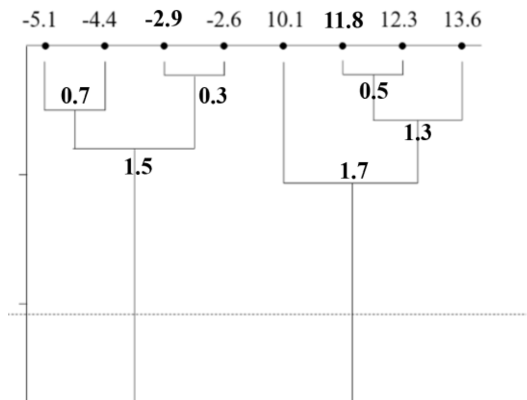
令 $\frac{\partial l(D)}{\partial \Sigma_j} = 0$，得 $\Sigma_j = \frac{\sum_{i=1}^{n}\gamma_{ij}(x_i - \mu_j)(x_i - \mu_j)^T}{\sum_{i=1}^{n}\gamma_{ij}}$

12.2 现有 8 个一维样本点如下：

$\{-5.1, \quad -4.4, \quad -2.9, \quad -2.6, \quad 10.1, \quad 11.8, \quad 12.3, \quad 13.6\}$

请尝试采用分级聚类的方法对上述样本进行聚类，其中类别之间的距离定义为 $d_{\min}(\mathrm{D}_i, \mathrm{D}_j)$，$d_{\min}(\mathrm{D}_i, \mathrm{D}_j) = \min_{\substack{x \in D_i \\ x' \in D_j}} \left\| x - x' \right\|$。画出聚类树，注明横纵坐标值，并说明你倾向于将这些数据聚为几类，为什么。

参考答案：



从图中可以看出，数据大致分为两类。

12.3 计算机小实验：MNSIT 数据集聚类

在本问题中，我们将利用聚类算法对 MNIST 训练集的“0”，“3”，“7”三类样本集进行聚类分析，并观察各聚类算法的性能差异。

（1） 请利用 K-means 算法，在 1 到 10 的类别数下对样本集进行聚类，分别使用误差平方和J_e和轮廓系数 SC 确定聚类数目，并结合真实的类别数进行分析。

（2） 请利用 K-means 算法对样本集进行聚类（类别数为 3），选用任意一种初始化方法，并采用 NMI 评估你的聚类结果（多次运行取平均值）

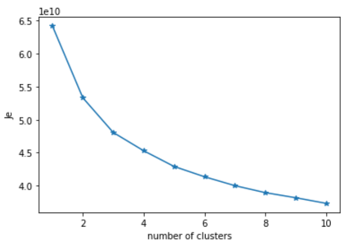
（3） 【选做】结合（2）中的聚类结果，编写一致聚类算法对样本重新聚类，采用 NMI 评估你的聚类结果，观察在此问题中，一致聚类算法是否对聚类效果有所提升。

提示：

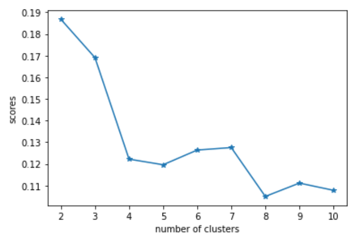
若原始 MNIST 数据集过大，可从中随机抽取一部分作为该题的样本集。

答案：

（1） Je



SC



从评价指标可以看出， Je 在类别数为 3 处出现了拐点，这与实际是相符的，SC 在类别数 2 和 3 上都有着较好的表现，类别数 2 的效果要略高于 3。Je 与实际有偏差的原因可能是由于轮廓系数更加偏好于凸的簇结构，从而导致一定的偏差。在我们无法确定类别数目时，我们应当在表现较优的几个类别数目下都进行尝试。

（2）

初始化方法	mean
random	0.7405
k-means++	0.7409
ndarray	0.7409

表格中，random 为随机初始化；k-means++的初始化方法为从输入的数据点集合（要求有 k 个聚类）中随机选择一个点作为第一个聚类中心，然后对于数据集中的每一个点 x，计算它与最近聚类中心(指已选择的聚类中心)的距离 D(x)，

在此基础上选择一个新的数据点作为新的聚类中心，选择的原则是：D(x)较大的点，被选取作为聚类中心的概率较大，从而选出 k 个聚类中心；ndarray 指的是选取前 k 个样本作为聚类中心。

通过三种初始方法的比较，通过 k-means++的方法和指定前 k 个样本的初始化方法在该数据集上得到了较好的效果，但相比较而言，前 k 个样本的初始化方法对数据集较为依赖，k-means++的方法较为普遍。总体而言，无论是随机初始化，还是通过某种方法初始化，都得到了较好的聚类效果。

（3）

由于一致聚类对内存的占用较大，因此在进行一致聚类时，从原始训练集中抽取 1000 个样本作为样本集。使用 K-means 算法作为内层算法，分级聚类作为外层算法，进行 5 次采样，并将一致聚类的结果和内层的 K-means 算法的 NMI 作比较，结果如下：

	1	2	3	4	5
Kmeans	0.7011	0.7540	0.7406	0.7099	0.7154
一致聚类	0.7218				

从图结果可以看出，每次采样运行的 K-means 算法结果较为波动，一致聚类的结果基本取得了多次 K-means 算法的平均结果。