单峰子集类的分离算法由于要对概率密度函数进行估计，所以计算的工作量很大。此外，由于在进行概率估计时要选定一些参数，因此估计的结果也会受到参数的较大影响。特别是在有噪声的情况下，具有局部最大值的概率密度函数的峰点数都会发生变化，从而不能正确反映数据中的单峰子集数。在样本数较少的情况下，由于没有可能对概率密度函数进行估计，从而也就使得这种方法完全失去意义。这时分级聚类算法可能是特别有用的。

与监督模式识别相比，之所以非监督模式识别问题中存在更大的不确定性，一个主要的原因就是在非监督问题中我们没有了已知类别的样本集，甚至可能不知道类别数，可以利用的信息量大大减少了。在实际应用中，除了本章介绍的内容外，还应注意设法有效利用应用领域的专门知识，以弥补信息的不足。最终所得聚类的实际含义也往往只有依靠有关知识来解释和确定。

习　题

10.1　令 $x_1,\cdots,x_N$ 是 d 维样本，Σ 是任一非奇异 $d\times d$ 矩阵。证明使

$$\sum_{k=1}^{N}(\boldsymbol{x}_k-\boldsymbol{x})^T\Sigma^{-1}(\boldsymbol{x}_k-\boldsymbol{x})$$

最小的向量 $\boldsymbol{x}$ 是样本均值。

10.2　令 $s(\boldsymbol{x},\boldsymbol{x}')=\boldsymbol{x}^T\boldsymbol{x}'/(\|\boldsymbol{x}\|\cdot\|\boldsymbol{x}'\|)$。若 $\boldsymbol{x}$ 的 d 个特征只取 $+1$ 和 -1 二值，即当 $\boldsymbol{x}$ 具有第 i 个特征时，$x_i=1$，而当 $\boldsymbol{x}$ 没有该特征时 $x_i=-1$，说明 s 是一个相似性度量。证明对于这种情况

$$\|\boldsymbol{x}-\boldsymbol{x}'\|^2=2d(1-s(\boldsymbol{x},\boldsymbol{x}'))。$$

10.3　假使一个有 N 个样本的集合 $\mathscr{X}$ 划分为 c 个不相交的子集 $\mathscr{X}_1,\cdots,\mathscr{X}_c$，假使 $\mathscr{X}_i$ 是空集，则 $\mathscr{X}_i$ 中样本的均值 $\boldsymbol{m}_i$ 不定义。在这种情况下，误差平方和只和非空子集有关：

$$\boldsymbol{J}_e=\sum_{i}\sum_{\boldsymbol{x}\in\mathscr{X}_i}\|\boldsymbol{x}-\boldsymbol{m}_i\|^2$$

这里 i 是不包含空子集的子集标号。

假定 $N\geqslant c$，证明使 $\boldsymbol{J}_e$ 最小的划分中没有空子集。

10.4　考虑一个 $N=2k+1$ 样本的集合，其中有 k 个在 $x=-2$ 重合，有 k 个在 $x=0$ 上重合，有一个在 $x=a>0$ 上。证明若 $a^2<2(k+1)$。则使 $\boldsymbol{J}_e$ 最小的两类划分是 $x=0$ 的 k 个样本和 $x=a$ 的那个样本聚为一类。若 $a^2>2(k+1)$，则应如何聚类使 $\boldsymbol{J}_e$ 最小？

10.5　设 $\boldsymbol{x}_1=(4\quad 5)^T,\boldsymbol{x}_2=(1\quad 4)^T,\boldsymbol{x}_3=(0\quad 1)^T,\boldsymbol{x}_4=(5\quad 0)^T$。现有下列三种划分：

(1) $\mathscr{X}_1=\{\boldsymbol{x}_1,\boldsymbol{x}_2\}$，　　$\mathscr{X}_2=\{\boldsymbol{x}_3,\boldsymbol{x}_4\}$

(2) $\mathscr{X}_1=\{\boldsymbol{x}_1,\boldsymbol{x}_4\}$，　　$\mathscr{X}_2=\{\boldsymbol{x}_2,\boldsymbol{x}_3\}$

(3) $\mathscr{X}_1=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3\}$，　　$\mathscr{X}_2=\{\boldsymbol{x}_4\}$

证明对于平方误差和准则，第三种划分最好，而若用 $|S_w|$ 准则，前两种划分为好。

10.6　若定义下列准则函数

$$\boldsymbol{J}_T=\sum_{i=1}^{c}\sum_{\boldsymbol{x}\in\mathscr{X}_i}(\boldsymbol{x}-\boldsymbol{m}_i)^TS_T^{-1}(\boldsymbol{x}-\boldsymbol{m}_i)$$

其中 $\boldsymbol{m}_i$ 是 $\mathscr{X}_i$ 中 N_i 个样本的均值向量，S_T 是总散布矩阵，

(1) 证明 J_T 对数据的非奇异线性变换具有不变性。

(2) 证明把 $\mathscr{X}_i$ 中的样本 $\hat{\boldsymbol{x}}$ 转移到 $\mathscr{X}_j$ 中去，则使 J_T 改变为

$$J_T^* = J_T - \left[\frac{N_j}{N_j+1}(\hat{\boldsymbol{x}}-\boldsymbol{m}_j)^T S_T^{-1}(\hat{\boldsymbol{x}}-\boldsymbol{m}_j) - \frac{N_i}{N_i-1}(\hat{\boldsymbol{x}}-\boldsymbol{m}_i)^T S_T^{-1}(\hat{\boldsymbol{x}}-\boldsymbol{m}_i)\right]$$

(3) 写出使 J_T 最小化的迭代程序。

10.7 令聚类 $\mathscr{X}_i$ 包含 N_i 个样本，令 d_{ij} 是两个聚类 $\mathscr{X}_i$ 和 $\mathscr{X}_j$ 之间的距离度量。若 $\mathscr{X}_i$ 和 $\mathscr{X}_j$ 合并形成一个新的聚类 $\mathscr{X}_k$，则 $\mathscr{X}_h$ 到 $\mathscr{X}_k$ 的距离一般情况下可用下式表示：

$$d_{hk} = \alpha_i d_{hi} + \alpha_j d_{hj} + \beta d_{ij} + \gamma |d_{hi} - d_{hj}|$$

证明当用不同的距离度量时，$\alpha_i, \alpha_j, \beta, \gamma$ 分别取不同的数值：

(1) $d_{\min}$：$\alpha_i=\alpha_j=0.5, \beta=0, \gamma=-0.5$

(2) $d_{\max}$：$\alpha_i=\alpha_j=0.5, \beta=0, \gamma=0.5$

(3) d_{avg}：$\alpha_i=\dfrac{N_i}{N_i+N_j}, \alpha_j=\dfrac{N_j}{N_i+N_j}, \beta=\gamma=0$

(4) d_{mean}^2：$\alpha_i=\dfrac{N_i}{N_i+N_j}, \alpha_j=\dfrac{N_j}{N_i+N_j}, \beta=-\alpha_i\alpha_j, \gamma=0$。

10.8 证明任一对称集上的概率密度函数是单峰的。

10.9 证明对于 C-均值算法，聚类准则函数满足使算法收敛的条件。(即若 $J(\Gamma,\widetilde{K}) \leqslant J(\Gamma,K)$，则有 $J(\widetilde{\Gamma},\widetilde{K}) \leqslant J(\Gamma,\widetilde{K})$)

10.10 令 $\Delta(\boldsymbol{y},K_j)=\dfrac{1}{2}(\boldsymbol{y}-\boldsymbol{m}_j)^T\Sigma_j^{-1}(\boldsymbol{y}-\boldsymbol{m}_j)+\dfrac{1}{2}\log|\Sigma_j|$ 是点到聚类的相似性度量，式中 $\boldsymbol{m}_j$ 和 Σ_j 是聚类 Γ_j 的均值和协方差矩阵，若把一点从 Γ_i 转移到 Γ_j 中去，计算由公式(10-44)所示 J_K 的变化值。