聚类分析习题

姓名： 甘云冲

学号： 2101213081

1.

采用反证法，假设存在一最优划分C，但是有$\boldsymbol{x}_k \in C_i$，满足$\|\boldsymbol{x}_k - \boldsymbol{m}_j\| < \|\boldsymbol{x}_k - \boldsymbol{m}_i\|, j \neq i$。

那么可以将$\boldsymbol{x}_k$划入C_j当中：

$$\Delta J = \|\boldsymbol{x}_k - \boldsymbol{m}_i\| - \|\boldsymbol{x}_k - \boldsymbol{m}_j\| < 0$$

是一个更优的划分，与假设矛盾。

2.

(1)

$$\boldsymbol{S}_t = \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{m})(\boldsymbol{x}_i - \boldsymbol{m})^T$$

经过线性变换：

$$\tilde{\boldsymbol{S}}_t = \sum_{i=1}^{n} (\boldsymbol{z}_i - \tilde{\boldsymbol{m}})(\boldsymbol{z}_i - \tilde{\boldsymbol{m}})^T = \sum_{i=1}^{n} (\boldsymbol{A}\boldsymbol{x}_i - \boldsymbol{A}\boldsymbol{m})(\boldsymbol{A}\boldsymbol{x}_i - \boldsymbol{A}\boldsymbol{m})^T = \boldsymbol{A}\left(\sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{m})(\boldsymbol{x}_i - \boldsymbol{m})^T\right)\boldsymbol{A}^T = \boldsymbol{A}\boldsymbol{S}\boldsymbol{A}^T$$

于是有：

$$\begin{aligned}
&\min_C \sum_{k=1}^{K} \sum_{\boldsymbol{z}_i \in C_k} (\boldsymbol{A}\boldsymbol{x}_i - \boldsymbol{A}\boldsymbol{m}_k)^T {\tilde{\boldsymbol{S}}_t}^{-1} (\boldsymbol{A}\boldsymbol{x}_i - \boldsymbol{A}\boldsymbol{m}_k) \\
\Leftrightarrow &\min_C \sum_{k=1}^{K} \sum_{\boldsymbol{z}_i \in C_k} (\boldsymbol{x}_i - \boldsymbol{m}_k)^T \boldsymbol{A}^T (\boldsymbol{A}\boldsymbol{S}\boldsymbol{A}^T)^{-1} \boldsymbol{A} (\boldsymbol{x}_i - \boldsymbol{m}_k) \\
\Leftrightarrow &\min_C \sum_{k=1}^{K} \sum_{\boldsymbol{z}_i \in C_k} (\boldsymbol{x}_i - \boldsymbol{m}_k)^T \boldsymbol{A}^T {\boldsymbol{A}^T}^{-1} \boldsymbol{S}^{-1} \boldsymbol{A}^{-1} \boldsymbol{A} (\boldsymbol{x}_i - \boldsymbol{m}_k) \\
\Leftrightarrow &\min_C \sum_{k=1}^{K} \sum_{\boldsymbol{z}_i \in C_k} (\boldsymbol{x}_i - \boldsymbol{m}_k)^T \boldsymbol{S}^{-1} (\boldsymbol{x}_i - \boldsymbol{m}_k)
\end{aligned}$$

(2)

假设移动之后的均值为$\hat{\boldsymbol{m}}_i, \hat{\boldsymbol{m}}_j$：

$$\hat{\boldsymbol{m}}_i = \frac{n_i \boldsymbol{m}_i - \hat{\boldsymbol{x}}}{n_i - 1} = \boldsymbol{m}_i + \frac{\boldsymbol{m}_i - \hat{\boldsymbol{x}}}{n_i - 1} = \boldsymbol{m}_i + \Delta \boldsymbol{m}_i, \quad \hat{\boldsymbol{m}}_j = \frac{n_j \boldsymbol{m}_j + \hat{\boldsymbol{x}}}{n_i + 1} = \boldsymbol{m}_j + \frac{\hat{\boldsymbol{x}} - \boldsymbol{m}_j}{n_j + 1} = \boldsymbol{m}_j + \Delta \boldsymbol{m}_j$$

计算变化量：

$$
\begin{aligned}
\Delta J_t(C) &= (\hat{\boldsymbol{x}}-\hat{\boldsymbol{m}}_j)^T\boldsymbol{S}^{-1}(\hat{\boldsymbol{x}}-\hat{\boldsymbol{m}}_j)-(\hat{\boldsymbol{x}}-\hat{\boldsymbol{m}}_i)^T\boldsymbol{S}^{-1}(\hat{\boldsymbol{x}}-\hat{\boldsymbol{m}}_i)\\
&\quad+\sum_{x_k\in C_i}\Big((\boldsymbol{x}_k-\hat{\boldsymbol{m}}_i)^T\boldsymbol{S}^{-1}(\boldsymbol{x}_k-\hat{\boldsymbol{m}}_i)-(\boldsymbol{x}_k-\boldsymbol{m}_i)^T\boldsymbol{S}^{-1}(\boldsymbol{x}_k-\boldsymbol{m}_i)\Big)\\
&\quad+\sum_{x_k\in C_j}\Big((\boldsymbol{x}_k-\hat{\boldsymbol{m}}_j)^T\boldsymbol{S}^{-1}(\boldsymbol{x}_k-\hat{\boldsymbol{m}}_j)-(\boldsymbol{x}_k-\boldsymbol{m}_j)^T\boldsymbol{S}^{-1}(\boldsymbol{x}_k-\boldsymbol{m}_j)\Big)\\
&=(\hat{\boldsymbol{x}}-\hat{\boldsymbol{m}}_j)^T\boldsymbol{S}^{-1}(\hat{\boldsymbol{x}}-\hat{\boldsymbol{m}}_j)-(\hat{\boldsymbol{x}}-\hat{\boldsymbol{m}}_i)^T\boldsymbol{S}^{-1}(\hat{\boldsymbol{x}}-\hat{\boldsymbol{m}}_i)\\
&\quad+\sum_{x_k\in C_i}\Big((-\Delta\boldsymbol{m}_i)^T\boldsymbol{S}^{-1}(\boldsymbol{x}_k-\boldsymbol{m}_i)+(\boldsymbol{x}_k-\boldsymbol{m}_i)^T\boldsymbol{S}^{-1}(-\Delta\boldsymbol{m}_i)+\Delta\boldsymbol{m}_i^T\boldsymbol{S}^{-1}\Delta\boldsymbol{m}_i\Big)\\
&\quad+\sum_{x_k\in C_j}\Big((-\Delta\boldsymbol{m}_j)^T\boldsymbol{S}^{-1}(\boldsymbol{x}_k-\boldsymbol{m}_j)+(\boldsymbol{x}_k-\boldsymbol{m}_j)^T\boldsymbol{S}^{-1}(-\Delta\boldsymbol{m}_j)+\Delta\boldsymbol{m}_j^T\boldsymbol{S}^{-1}\Delta\boldsymbol{m}_j\Big)\\
&=(\hat{\boldsymbol{x}}-\hat{\boldsymbol{m}}_j)^T\boldsymbol{S}^{-1}(\hat{\boldsymbol{x}}-\hat{\boldsymbol{m}}_j)-(\hat{\boldsymbol{x}}-\hat{\boldsymbol{m}}_i)^T\boldsymbol{S}^{-1}(\hat{\boldsymbol{x}}-\hat{\boldsymbol{m}}_i)\\
&\quad+n_i\Delta\boldsymbol{m}_i^T\boldsymbol{S}^{-1}\Delta\boldsymbol{m}_i+n_j\Delta\boldsymbol{m}_j^T\boldsymbol{S}^{-1}\Delta\boldsymbol{m}_j\\
&=\left((\frac{n_j}{n_j+1})^2+\frac{n_j}{(n_j+1)^2}\right)(\hat{\boldsymbol{x}}-\boldsymbol{m}_j)^T\boldsymbol{S}^{-1}(\boldsymbol{x}-\hat{\boldsymbol{m}}_j)-\left((\frac{n_i}{n_i-1})^2+\frac{n_i}{(n_i-1)^2}\right)(\hat{\boldsymbol{x}}-\boldsymbol{m}_i)^T\boldsymbol{S}^{-1}(\boldsymbol{x}-\hat{\boldsymbol{m}}_i)\\
&=\frac{n_j}{n_j+1}(\hat{\boldsymbol{x}}-\boldsymbol{m}_j)^T\boldsymbol{S}^{-1}(\boldsymbol{x}-\hat{\boldsymbol{m}}_j)-\frac{n_i}{n_i-1}(\hat{\boldsymbol{x}}-\boldsymbol{m}_i)^T\boldsymbol{S}^{-1}(\boldsymbol{x}-\hat{\boldsymbol{m}}_i)
\end{aligned}
$$

于是有：

$$
\hat{J}_t(C)=J_t(C)+\Delta J_t(C)=J_t(C)+\left[\frac{n_j}{n_j+1}(\hat{\boldsymbol{x}}-\boldsymbol{m}_j)^T\boldsymbol{S}^{-1}(\boldsymbol{x}-\hat{\boldsymbol{m}}_j)-\frac{n_i}{n_i-1}(\hat{\boldsymbol{x}}-\boldsymbol{m}_i)^T\boldsymbol{S}^{-1}(\boldsymbol{x}-\hat{\boldsymbol{m}}_i)\right]
$$

(3)

1. 随机分配所有的点到K个类上，计算K个类的中心
2. 随机选择一个点，把它移出所属类
3. 对所有类（包括移出的）计算赋予之后$J_t(C)$的变化量，并将其赋予$\Delta J_t(C)$最小的对应类
4. 重新计算移出和移入类的中心
5. 重复步骤2～4，直到达到收敛或者某个停止阈值

3.

(1)

$\boldsymbol{q}^T\boldsymbol{D}\boldsymbol{e}=0$：

$$
\boldsymbol{q}^T\boldsymbol{D}\boldsymbol{e}=\sum_{i=1}^n d_iq_i=\sum_{\boldsymbol{x}_i\in C}d_iq_i+\sum_{\boldsymbol{x}_i\in\bar{C}}d_iq_i=\sqrt{\frac{vol(\bar{C})}{vol(C)}}\sum_{\boldsymbol{x}_i\in C}d_i-\sqrt{\frac{vol(C)}{vol(\bar{C})}}\sum_{\boldsymbol{x}_i\in\bar{C}}d_i=\sqrt{\frac{vol(\bar{C})}{vol(C)}}vol(C)-\sqrt{\frac{vol(C)}{vol(\bar{C})}}vol(\bar{C})=0
$$

$\boldsymbol{q}^T\boldsymbol{D}\boldsymbol{q}=vol(C)+vol(\bar{C})$：

$$
\boldsymbol{q}^T\boldsymbol{D}\boldsymbol{q}=\sum_{i=1}^n d_iq_i^2=\sum_{\boldsymbol{x}_i\in C}d_iq_i^2+\sum_{\boldsymbol{x}_i\in\bar{C}}d_iq_i^2=\frac{vol(\bar{C})}{vol(C)}\sum_{\boldsymbol{x}_i\in C}d_i+\frac{vol(C)}{vol(\bar{C})}\sum_{\boldsymbol{x}_i\in\bar{C}}d_i=\frac{vol(\bar{C})}{vol(C)}vol(C)+\frac{vol(C)}{vol(\bar{C})}vol(\bar{C})=vol(C)+vol(\bar{C})
$$

$\boldsymbol{q}^T\boldsymbol{L}\boldsymbol{q}=\Big(vol(\bar{C})+val(C)\Big)NCut(C,\bar{C})$：

$$
\begin{aligned}
\boldsymbol{q}^T\boldsymbol{L}\boldsymbol{q} &= \sum_{i=1}^n d_i q_i^2 - \sum_{i=1}^n \sum_{j=1}^n w_{ij} q_i q_j \\
&= \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n w_{ij} (q_i^2 - 2q_i q_j + q_j^2) \\
&= \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n w_{ij} (q_i - q_j)^2 \\
&= \frac{1}{2} \left(\sum_{i\in C, j\in \bar{C}} w_{ij} \left(\sqrt{\frac{vol(\bar{C})}{vol(C)}} + \sqrt{\frac{vol(C)}{vol(\bar{C})}} \right)^2 + \sum_{i\in \bar{C}, j\in C} w_{ij} \left(\sqrt{\frac{vol(\bar{C})}{vol(C)}} + \sqrt{\frac{vol(C)}{vol(\bar{C})}} \right)^2 \right) \\
&= \sum_{i\in C, j\in \bar{C}} w_{ij} \left(\frac{vol(\bar{C})}{vol(C)} + \frac{vol(C)}{vol(\bar{C})} + 2 \right) \\
&= \sum_{i\in C, j\in \bar{C}} w_{ij} \left(\frac{vol(\bar{C}) + val(C)}{vol(C)} + \frac{vol(\bar{C}) + vol(C)}{vol(\bar{C})} \right) \\
&= \Big(vol(\bar{C}) + val(C) \Big) \left(\frac{\sum_{i\in C, j\in \bar{C}} w_{ij}}{vol(C)} + \frac{\sum_{i\in C, j\in \bar{C}} w_{ij}}{vol(\bar{C})} \right) \\
&= \Big(vol(\bar{C}) + val(C) \Big) NCut(C, \bar{C})
\end{aligned}
$$

(2)

$\boldsymbol{Q}^T\boldsymbol{D}\boldsymbol{Q} = \boldsymbol{I}_K$:

$$
(\boldsymbol{Q}^T\boldsymbol{D}\boldsymbol{Q})_{ij} = \sum_{k=1}^n d_k q_{ki} q_{kj} = \sum_{k=1}^n \mathbb{I}(\boldsymbol{x}_k \in C_i)\mathbb{I}(\boldsymbol{x}_k \in C_j) d_k \sqrt{\frac{1}{vol(C_i)vol(C_j)}}
$$

$i = j$:

$$
(\boldsymbol{Q}^T\boldsymbol{D}\boldsymbol{Q})_{ii} = \sum_{k=1}^n \mathbb{I}(\boldsymbol{x}_k \in C_i) \frac{d_k}{vol(C_i)} = \frac{1}{vol(C_i)} \sum_{\boldsymbol{x}_k \in C_i} d_k = 1
$$

$1 \neq j$:

$$
\mathbb{I}(\boldsymbol{x}_k \in C_i)\mathbb{I}(\boldsymbol{x}_k \in C_j) = 0 \Rightarrow (\boldsymbol{Q}^T\boldsymbol{D}\boldsymbol{Q})_{ij} = 0
$$

综上所述应当有$\boldsymbol{Q}^T\boldsymbol{D}\boldsymbol{Q} = \boldsymbol{I}_K$。

$tr(\boldsymbol{Q}^T\boldsymbol{L}\boldsymbol{Q}) = \sum_{k=1}^K \frac{Cut(C_k,\bar{C}_k)}{vol(C_k)}$:

$$
\begin{aligned}
tr(\boldsymbol{Q}^T\boldsymbol{L}\boldsymbol{Q}) &= \sum_{k=1}^K \boldsymbol{q}_k^T \boldsymbol{L} \boldsymbol{q}_k \\
&= \sum_{k=1}^K \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n w_{ij} (q_{ik} - q_{jk})^2 \\
&= \frac{1}{2} \sum_{k=1}^K \left(\sum_{i\in C_k, j\in \bar{C}_k} w_{ij} \frac{1}{vol(C_k)} + \sum_{i\in \bar{C}_k, j\in C_k} w_{ij} \frac{1}{vol(C_k)} \right) \\
&= \sum_{k=1}^K \sum_{i\in C_k, j\in \bar{C}_k} \frac{w_{ij}}{vol(C_k)} \\
&= \sum_{k=1}^K \frac{Cut(C_k, \bar{C}_k)}{vol(C_k)}
\end{aligned}
$$