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 $m{x}_k\in C_i$,满足 $\|m{x}_k-m{m}_j\|<\|m{x}_k-m{m}_i\|,j
eq i$ 。那么可以将 $m{x}_k$ 划入 C_i 当中:

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

是一个更优的划分,与假设矛盾。

2.

(1)

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

经过线性变换:

$$ilde{oldsymbol{S}}_t = \sum_{i=1}^n (oldsymbol{z}_i - ilde{oldsymbol{m}})(oldsymbol{z}_i - ilde{oldsymbol{m}})^T = \sum_{i=1}^n (oldsymbol{A}oldsymbol{x}_i - oldsymbol{A}oldsymbol{m})(oldsymbol{A}oldsymbol{x}_i - oldsymbol{m})^T = oldsymbol{A}oldsymbol{S}oldsymbol{A}^T = oldsymbol{A}oldsymbol{B}oldsymbol{A}^T = oldsymbol{A}oldsymbol{B}oldsymbol{A}^T = oldsymbol{A}oldsymbol{B}oldsymbol{A}^T = oldsymbol{A}oldsymbol{B}oldsymbol{A}^T = oldsymbol{A}oldsymbol{A}oldsymbol{B} + oldsymbol{A}oldsymbol{B} + oldsymbol{A}oldsymbol{B} + oldsymbol{B} + oldsymb$$

于是有:

$$egin{aligned} \min_{C} \sum_{k=1}^{K} \sum_{oldsymbol{z}_i \in C_k} (oldsymbol{A} oldsymbol{x}_i - oldsymbol{A} oldsymbol{m}_k)^T ilde{oldsymbol{S}}_t^{-1} (oldsymbol{A} oldsymbol{x}_i - oldsymbol{A} oldsymbol{m}_k) \ &\Leftrightarrow \min_{C} \sum_{k=1}^{K} \sum_{oldsymbol{z}_i \in C_k} (oldsymbol{x}_i - oldsymbol{m}_k)^T oldsymbol{A}^T oldsymbol{A}^T oldsymbol{A}^{-1} oldsymbol{A} (oldsymbol{x}_i - oldsymbol{m}_k) \ &\Leftrightarrow \min_{C} \sum_{k=1}^{K} \sum_{oldsymbol{z}_i \in C_k} (oldsymbol{x}_i - oldsymbol{m}_k)^T oldsymbol{S}^{-1} (oldsymbol{x}_i - oldsymbol{m}_k) \end{aligned}$$

(2)

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

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

计算变化量:

$$\begin{split} \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) \\ &+ \sum_{x_k \in C_i} \left((\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) \right) \\ &+ \sum_{x_k \in C_j} \left((\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) \right) \\ &= (\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}} - \boldsymbol{m}_i) \\ &+ \sum_{x_k \in C_i} \left((-\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 \right) \\ &+ \sum_{x_k \in C_j} \left((-\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 \right) \\ &= (\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) \\ &+ 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_i + 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{split}$$

于是有:

$$\hat{J}_t(C) = J_t(C) + \Delta J_t(C) = J_t(C) + \left[rac{n_j}{n_j+1}(\hat{oldsymbol{x}} - oldsymbol{m}_j)^Toldsymbol{S}^{-1}(oldsymbol{x} - \hat{oldsymbol{m}}_j) - rac{n_i}{n_i-1}(\hat{oldsymbol{x}} - oldsymbol{m}_i)^Toldsymbol{S}^{-1}(oldsymbol{x} - \hat{oldsymbol{m}}_i)
ight]$$

(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)

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

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

$$oldsymbol{q}^T oldsymbol{D} oldsymbol{q} = vol(C) + vol(ar{C})$$
 :

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

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

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

(2)

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

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

i = j:

$$(oldsymbol{Q}^Toldsymbol{D}oldsymbol{Q})_{ii} = \sum_{k=1}^n \mathbb{I}(oldsymbol{x}_k \in C_i) rac{d_k}{vol(C_i)} = rac{1}{vol(C_i)} \sum_{oldsymbol{x}_i \in C_i} d_k = 1$$

 $1 \neq j$:

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

综上所述应当有 $Q^TDQ = I_K$ 。

$$tr(oldsymbol{Q}^Toldsymbol{L}oldsymbol{Q}) = \sum_{k=1}^K rac{Cut(C_k,ar{C}_k)}{vol(C_k)}$$
 :

$$egin{aligned} tr(oldsymbol{Q}^Toldsymbol{L}oldsymbol{Q}) &= \sum_{k=1}^K oldsymbol{q}_k^Toldsymbol{L}oldsymbol{q}_k \ &= \sum_{k=1}^K rac{1}{2} \sum_{i=1}^n \sum_{j=1}^n w_{ij} (q_{ik} - q_{jk})^2 \ &= rac{1}{2} \sum_{k=1}^K \left(\sum_{i \in C_k, j \in \bar{C}_k} w_{ij} rac{1}{vol(C_k)} + \sum_{i \in \bar{C}_k, j \in C_k} w_{ij} rac{1}{vol(C_k)}
ight) \ &= \sum_{k=1}^K \sum_{i \in C_k, j \in \bar{C}_k} rac{w_{ij}}{vol(C_k)} \ &= \sum_{k=1}^K rac{Cut(C_k, \bar{C}_k)}{vol(C_k)} \end{aligned}$$