Homework 1 排名问题

题目 1.

k 名专家对 n 件作品按从优到劣的顺序进行排序, 用 $\sigma_j^i = l$ 表示专家 i 认为作品 j 位于第 l 位. 记 $\boldsymbol{\sigma}_i = (\sigma_1^i, \sigma_2^i, ..., \sigma_n^i)$ 为专家 i 的排序向量, $\boldsymbol{\Sigma} = \{\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2, ..., \boldsymbol{\sigma}_k\}$ 为 k 名专家的排序集合. 现希望给出一种能较好地反映所有专家意见的综合排序.

对两个 n 维向量 $\boldsymbol{u} = (u_1, u_2, ..., u_n)$ 和 $\boldsymbol{v} = (v_1, v_2, ..., v_n)$, 定义它们之间的距离为 $L_1(\boldsymbol{u}, \boldsymbol{v}) = \sum_{i=1}^n |u_i - v_i|$. 排序 $\boldsymbol{\sigma}$ 与一组排序 $\boldsymbol{\Sigma} = \{\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2, ..., \boldsymbol{\sigma}_k\}$ 的综合排序定义为 $d(\boldsymbol{\sigma}, \boldsymbol{\Sigma}) = \sum_{i=1}^k L_1(\boldsymbol{\sigma}, \boldsymbol{\sigma}_i)$. 对给定的 $\boldsymbol{\Sigma}$, 与 $\boldsymbol{\Sigma}$ 综合距离最小的排序记为 $\boldsymbol{\sigma}^*$.

(1) 给定 $\boldsymbol{\Sigma} = \{\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2, ..., \boldsymbol{\sigma}_k\}$, 求 n 维向量 $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_n)$, 使得 $\sum_{i=1}^k L_1(\boldsymbol{\mu}, \boldsymbol{\sigma}_i)$ 最小. 你能否根据 $\boldsymbol{\mu}$ 给出 n 件作品的一种综合排序 $\boldsymbol{\sigma}'$, $\boldsymbol{\sigma}^*$ 是否也能从 $\boldsymbol{\mu}$ 得到, 为什么;

(2) 有人提议用 Borda 计分法给出综合排序. 首先计算作品 j 的平均得分 $\beta_j = \frac{1}{k}\sum_{i=1}^k \sigma_j^i$, 再按得分从小到大的顺序对作品进行排序 (得分相同的作品之间的顺序可任意确定), 由此给出一种综合排序 $\boldsymbol{\sigma}''$. 证明: 对任意 j, $\sum_{i=1}^k |\beta_j - \sigma_j^i| \le 2\sum_{i=1}^k |\mu_j - \sigma_j^i|$.

(3) 证明: $d(\boldsymbol{\sigma}', \boldsymbol{\Sigma}) \le 3d(\boldsymbol{\sigma}^*, \boldsymbol{\Sigma})$, 且 $d(\boldsymbol{\sigma}'', \boldsymbol{\Sigma}) \le 5d(\boldsymbol{\sigma}^*, \boldsymbol{\Sigma})$.

解答.

(1) $\sum_{i=1}^k L_1(\boldsymbol{\mu}, \boldsymbol{\sigma}_i) = \sum_{j=1}^n \left(\sum_{i=1}^k |\mu_j - \sigma_j^i|\right)$, 相当于每个 j 最小化 $\sum_{i=1}^k |\mu_j - \sigma_j^i|$.

由绝对值的几何意义, 当 μ_j 为 $\{\sigma_j\}$ 的中位数时, 绝对值的和最小. 具体而言, 我们取:

$$\mu_j = \mathrm{median}\{\sigma_j^1, \sigma_j^2, ..., \sigma_j^k\}$$

从而得到 $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_n)$.

根据 $\boldsymbol{\mu}$, 我们显然可以按照 μ_j 的值从小到大进行排序 (如果 μ_j 相等, 则顺序任意), 然后依次分配排名 $1, 2, ..., n$, 得到综合排序 $\boldsymbol{\sigma}'$.

但是综合排序 $\boldsymbol{\sigma}^*$ 不能直接从 $\boldsymbol{\mu}$ 中得到, 因为 $\boldsymbol{\mu}$ 中可能存在重复元素, 并且无法保证形成的排列是全局最优的.

(2) 注意到:

$$\begin{aligned}
\sum_{i=1}^k |\beta_j - \sigma_j^i| &\le \sum_{i=1}^k |\beta_j - \mu_j| + \sum_{i=1}^k |\mu_j - \sigma_j^i| \\
&= k\ |\frac{1}{k}\sum_{i=1}^k \sigma_j^i - \mu_j| + \sum_{i=1}^k |\mu_j - \sigma_j^i| \\
&\le k \cdot \frac{1}{k}\sum_{i=1}^k |\sigma_j^i - \mu_j| + \sum_{i=1}^k |\mu_j - \sigma_j^i| \\
&= 2\sum_{i=1}^k |\mu_j - \sigma_j^i|
\end{aligned}$$

因此命题得证.

(3) 由 (1) 知 $d(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \le d(\boldsymbol{\sigma}^*, \boldsymbol{\Sigma})$. 下面证明: 对于任何排列 $\boldsymbol{\sigma}_0$, 都有 $L_1(\boldsymbol{\mu}, \boldsymbol{\sigma}') \le L_1(\boldsymbol{\mu}, \boldsymbol{\sigma}_0)$:

证明. 根据 $\boldsymbol{\sigma}'$ 的构造方式, 若 $\sigma_i' < \sigma_j'$, 则有 $\mu_i \le \mu_j$. 若 $\exists(i,j)$, 满足 $\sigma_{0,i} < \sigma_{0,j}$ 且 $\mu_i > \mu_j$, 则有:

$$\big(|\mu_i-\sigma_{0,i}|+|\mu_j-\sigma_{0,j}|\big)-\big(|\mu_i-\sigma_{0,j}|+|\mu_j-\sigma_{0,i}|\big)$$
$$=\big(|\mu_i-\sigma_{0,i}|-|\mu_i-\sigma_{0,j}|\big)-\big(|\mu_j-\sigma_{0,i}|-|\mu_j-\sigma_{0,j}|\big)$$

令 $f(x)=|x-\sigma_{0,i}|-|x-\sigma_{0,j}|$, 显然 $f(x)$ 在 $\mathbb{R}$ 上单调递增, 因此上式 $f(\mu_i)-f\big(\mu_j\big)\le 0$.

由此可见, 将 $\sigma_{0,i}$ 和 $\sigma_{0,j}$ 交换, 会让 $L_1(\boldsymbol{\mu},\boldsymbol{\sigma}_0)$ 的值更小.

不断进行如上交换, 由"调整法"知 $\boldsymbol{\sigma}_0$ 最终会交换成 $\boldsymbol{\sigma}'$.

于是由绝对值不等式:

$$\begin{aligned}
d(\boldsymbol{\sigma}',\boldsymbol{\Sigma})&=\sum_{i=1}^{k}L_1(\boldsymbol{\sigma}',\boldsymbol{\sigma}_i)\\
&\le\sum_{i=1}^{k}(L_1(\boldsymbol{\sigma}',\boldsymbol{\mu})+L_1(\boldsymbol{\mu},\boldsymbol{\sigma}_i))\\
&\le\sum_{i=1}^{k}L_1(\boldsymbol{\sigma}^*,\boldsymbol{\mu})+d(\boldsymbol{\mu},\boldsymbol{\Sigma})\\
&\le\sum_{i=1}^{k}(L_1(\boldsymbol{\sigma}^*,\boldsymbol{\sigma}_{\boldsymbol{i}})+L_1(\boldsymbol{\sigma}_{\boldsymbol{i}},\boldsymbol{\mu}))+d(\boldsymbol{\mu},\boldsymbol{\Sigma})\\
&=d(\boldsymbol{\sigma}^*,\boldsymbol{\Sigma})+2d(\boldsymbol{\mu},\boldsymbol{\Sigma})\\
&\le 3d(\boldsymbol{\sigma}^*,\boldsymbol{\Sigma})
\end{aligned}$$

由 (2) 知 $d(\beta,\boldsymbol{\Sigma})\le 2d(\boldsymbol{\mu},\boldsymbol{\Sigma})\le 2d(\boldsymbol{\sigma}^*,\boldsymbol{\Sigma})$, 因此同理可得:

$$\begin{aligned}
d(\boldsymbol{\sigma}'',\boldsymbol{\Sigma})&=\sum_{i=1}^{k}L_1(\boldsymbol{\sigma}'',\boldsymbol{\sigma}_i)\\
&\le\sum_{i=1}^{k}(L_1(\boldsymbol{\sigma}'',\beta)+L_1(\beta,\boldsymbol{\sigma}_{\boldsymbol{i}}))\\
&\le\sum_{i=1}^{k}L_1(\boldsymbol{\sigma}^*,\beta)+d(\beta,\boldsymbol{\Sigma})\\
&\le\sum_{i=1}^{k}(L_1(\boldsymbol{\sigma}^*,\boldsymbol{\sigma}_i)+L_1(\boldsymbol{\sigma}_i,\beta))+d(\beta,\boldsymbol{\Sigma})\\
&=d(\boldsymbol{\sigma}^*,\boldsymbol{\Sigma})+2d(\beta,\boldsymbol{\Sigma})\\
&\le 5d(\boldsymbol{\sigma}^*,\boldsymbol{\Sigma})
\end{aligned}$$

注记.

(2) 相当于对于任何实数集合, **平均值与各点的绝对差之和** 不超过 **中位数与各点的绝对差之和** 的两倍.

(3) 其实 σ' 就是 μ 离散化后的结果, 直观上很符合"L_1 距离最近"的直觉.

之后有若干步绝对值不等式的放缩, 其实都不难, 主要是抓住关联:

向量	对应最优排列
$\boldsymbol{\mu}$	$\boldsymbol{\sigma}'$
β	$\boldsymbol{\sigma}''$

而 $\boldsymbol{\sigma}^*$ 的性质很糟糕, 但只要注意到 $d(\boldsymbol{\mu},\boldsymbol{\Sigma})\le d(\boldsymbol{\sigma}^*,\boldsymbol{\Sigma})$, 用 $\boldsymbol{\mu}$ 当中间跳板即可规避对 $\boldsymbol{\sigma}^*$ 的刻画.

题目 2.

n 支球队进行比赛, 每场比赛在两支球队之间进行, 任意两支球队之间至多进行一场比赛, 每支球队参与比赛的场数相同. 记队 i 与队 j 比赛中, 队 i 的得分为 $p_{i,j}$, 队 j 的得分为 $p_{j,i}$, 队 i 的分差为 $q_{i,j} = p_{i,j} - p_{j,i}$. 与队 i 进行过比赛的球队集合记为 T_i. 约定 $i \in T_i$, 且 $q_{i,i} = 0$. 记 $|T_1| = |T_2| = ... = |T_n| = l$.

A-B	5-10
A-D	57-45
B-C	10-7
C-D	3-10

(1) 记 s_i 为队 i 在各场比赛中分差之和, 即 $s_i = \sum_{j \in T_i} q_{i,j}$, $\boldsymbol{S} = (s_1, s_2, ..., s_n)^T$ 称为分差向量, 可用来衡量各球队的实力. 若四支球队之间的比赛结果如表所示, 求向量 $\boldsymbol{S}$.

(2) 对任意 $j \in T_i$, 若 $k \in T_j$, 则称队 i 与队 k 之间进行了一场"二级比赛", 且在该场比赛中队 i 的分差为 $q_{i,j} + q_{j,k}$ (队 i 可与自身进行二级比赛, 队 i 与队 j 之间可以进行多场二级比赛). 记 $s_i^{(2)}$ 为队 i 在所有可能的 l^2 场二级比赛中的分差之和, $\boldsymbol{S}^{(2)} = \left(s_1^{(2)}, s_2^{(2)}, ..., s_n^{(2)}\right)^T$ 称为二级分差向量. 对表中所示的比赛结果, 求向量 $\boldsymbol{S}^{(2)}$.

(3) 定义矩阵 $\boldsymbol{M} = \left(m_{i,j}\right)_{n \times n}$, 其中 $m_{i,j} = [j \in T_i]$. 试给出由 $\boldsymbol{M}$ 和 $\boldsymbol{S}$ 计算 $\boldsymbol{S}^{(2)}$ 的公式, 并说明 $\boldsymbol{M}^2$ 中各元素的含义.

(4) 类似地, 对任意整数 r, 可定义 r 级比赛和 r 级分差向量 $\boldsymbol{S}^{(r)}$, 试给出由 $\boldsymbol{M}$ 和 $\boldsymbol{S}$ 计算 $\boldsymbol{S}^{(r)}$ 的公式.

解答.

(1) 由图知 $q_{1,2} = -5, q_{1,4} = 12, q_{2,3} = 3, q_{2,1} = 5, q_{3,4} = -7, q_{3,2} = -3, q_{4,1} = -12, q_{4,3} = 7$. 因此:

$$s_1 = q_{1,2} + q_{1,4} = 7$$

$$s_2 = q_{2,3} + q_{2,1} = 8$$

$$s_3 = q_{3,4} + q_{3,2} = -10$$

$$s_4 = q_{4,1} + q_{4,3} = -5$$

得到 $\boldsymbol{S} = (7, 8, -10, -5)^T$.

(2) 首先 $l = |T_1| = |T_2| = |T_3| = |T_4| = 3$, 我们分别计算:

$$s_1^{(2)} = 7 + (-7) + 31 = 31$$

$$s_2^{(2)} = 8 + (-1) + 22 = 29$$

$$s_3^{(2)} = (-10) + (-26) + (-1) = -37$$

$$s_4^{(2)} = (-5) + (-29) + 11 = -23$$

得到 $\boldsymbol{S}^{(2)} = (31, 29, -37, -23)^T$.

(3) $\boldsymbol{S}^{(2)} = \boldsymbol{M}\boldsymbol{S} + l\boldsymbol{S}$. $\boldsymbol{M}_{i,j}^2$ 的含义为队 i 与队 j 进行"二级比赛"的场数.

(4) 显然 $\boldsymbol{S}^{(r)} = \boldsymbol{M}\boldsymbol{S}^{(r-1)} + l^{r-1}S$, 化简得 $\boldsymbol{S}^{(r)} = \sum_{k=0}^{r-1} l^{r-1-k}\boldsymbol{M}^k\boldsymbol{S}$.

注记.

- 浙江大学 2015-2016 学年秋冬学期《数学建模》课程期末考试