元素唯一性 (Element Distinctness) 问题

岳镝

问题 0.1 (元素唯一性问题). 设 S 是 n 个数构成的数组，判断 S 中的元素是否都是唯一的。如果唯一，则输出 "*Yes*"，否则输出 "*No*"。

定理 0.2. 以比较作为基本运算，元素唯一性问题的复杂度是 $\Theta(n \log n)$。

1 决策树

记多重集 $S = \{x_1, x_2, \ldots, x_n\}$。考虑元素唯一性问题的任意算法 $\mathcal{A}$，构造 $\mathcal{A}$ 对应的决策树 $\mathcal{T}$ 如下。每个内部节点 (i, j) 对应着算法 $\mathcal{A}$ 针对 x_i, x_j 的一次比较。按照如下规则构造 (i, j) 的子节点。

- (i, j) 的左子节点对应 $x_i < x_j$ 时算法的下一步操作。具体地，若算法结束，则将 (i, j) 的左子节点标记为叶节点；否则，将 (i, j) 的左子节点标记为下一步将比较的元素对 (k, ℓ)。
- (i, j) 的中间子节点对应 $x_i = x_j$ 时算法的下一步操作，标记为叶节点即可。
- (i, j) 的右子节点对应 $x_i > x_j$ 时算法的下一步操作。具体地，若算法结束，则将 (i, j) 的右子节点标记为叶节点；否则，将 (i, j) 的右子节点标记为下一步将比较的元素对 (k, ℓ)。

注. 上述构造得到的决策树并不是二叉树，这违背了教材的定义，但并不影响我们后续的分析。或者，你也可以不扩展出上述 "中间子节点"，这样得到的决策树还是一棵二叉树。

注. 决策树本质上是用来描述算法行为的一种工具。如果你像参考答案那样依据 $=$ 和 $\neq$ 构造决策树，则无法描述出算法 $\mathcal{A}$ 执行过程中 "比大小" 的行为，这样得到的下界可能是错误的。（回忆我们在小班课上讨论的 "两个世界"。）

2 节点性质

要确定决策树 $\mathcal{T}$ 的深度，我们需要给出 $\mathcal{T}$ 叶节点个数的下界。与排序问题不同，元素唯一性问题只有两种可能的输出 "Yes" "No"，因此叶节点个数的 $n!$ 下界远不是平凡的（平凡的下界是 2）。我们在这一节中讨论 $\mathcal{T}$ 上节点的性质，在下一节中证明 $\mathcal{T}$ 至少有 $n!$ 个叶节点。所有证明基于 [DL79]。

我们首先介绍欧氏空间中凸集的概念。直觉上，一个点集称为凸的，若其中任何两点的连线整体仍包含在这个集合中，如图 1 所示。

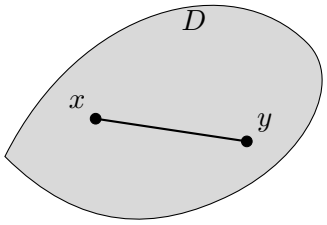
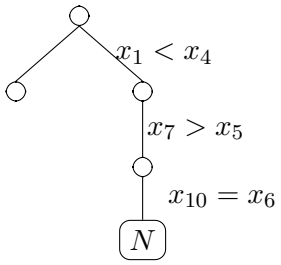


图 1: 凸集示意图。任意两点 x, y 连线 $\overline{xy}$ 仍落在 D 内部。

定义 2.1 (凸集)**.** 称 n 维欧氏空间的子集 $D \subset \mathbb{R}^n$ 为凸集，若对任意两点 $\mathbf{x}, \mathbf{y} \in D$ 及任意 $\lambda \in [0,1]$，均有 $\lambda\mathbf{x} + (1-\lambda)\mathbf{y} \in D$。

我们在小班课上讨论过，算法的输入可以视为欧氏空间 $\mathbb{R}^n$ 中的一个向量 $\mathbf{x} = (x_1, x_2, \ldots, x_n)$，而决策树上的每个节点对应着 $\mathbb{R}^n$ 的一个子集（可到达该节点的所有输入组成的集合）。以下引理表明，决策树的每个节点都是凸集。

引理 2.2. *考虑元素唯一性问题的算法 $\mathcal{A}$ 及其对应的决策树 $\mathcal{T}$。设 N 是 $\mathcal{T}$ 的任意节点，则 N 是 $\mathbb{R}^n$ 中的凸集。*



证明. 事实上，N 是由根节点到它的路径上所有比大小结果决定的（如上图所示）。假设从根节点到 N 的路径上的比大小结果分别为 $x_{i_1} \circ x_{j_1}, x_{i_2} \circ x_{j_2}, \ldots, x_{i_\ell} \circ x_{j_\ell}$，其中 $\circ \in \{<, >, =\}$。则

$$\begin{aligned} N &= \{\mathbf{x} \in \mathbb{R}^n \colon x_{i_1} \circ x_{j_1} \text{ 且 } x_{i_2} \circ x_{j_2} \text{ 且 } \ldots \text{ 且 } x_{i_\ell} \circ x_{j_\ell}\} \\ &= \bigcap_{r=1}^{\ell} \{\mathbf{x} \in \mathbb{R}^n \colon x_{i_r} \circ x_{j_r}\}. \end{aligned}$$

注意到每个集合 $\{\mathbf{x} \in \mathbb{R}^n \colon x_{i_r} \circ x_{j_r}\}$ 均为凸集（证明留作练习）。而有限个凸集的交集仍为凸集（证明留作练习）。所以 N 是凸集。 □

3 叶节点个数的一个下界

引理 3.1 (叶节点个数下界)**.** *考虑元素唯一性问题的算法 $\mathcal{A}$ 及其对应的决策树 $\mathcal{T}$。则 $\mathcal{T}$ 至少有 $n!$ 个叶节点。*

直觉上，我们将证明排列顺序不同的输入会最终进入不同的叶节点。严格来说，对 $[n]$ 上的置换 $\sigma \in S_n$，定义

$$A_\sigma := \{\mathbf{x} \in \mathbb{R}^n \colon\ x_{\sigma(1)} < x_{\sigma(2)} < \cdots < x_{\sigma(n)}\}.$$

显然，对于给定的 σ，A_σ 中的所有元素会进入相同的叶节点。我们以下说明，对于 $\sigma \neq \tau$，A_σ 和 A_τ 将会进入不同的叶节点。

引理 3.2. *设 L 是 $\mathcal{T}$ 的任意叶节点。则至多存在一个置换 $\sigma \in S_n$ 使得 A_σ 与 L 交集非空。*

证明. 反证法，假设存在两个不同置换 $\sigma, \tau \in S_n$，使得 $A_\sigma \cap L \neq \emptyset$ 且 $A_\tau \cap L \neq \emptyset$。取 $\mathbf{x} \in A_\sigma \cap L$, $\mathbf{y} \in A_\tau \cap L$。

注意到 $\sigma \neq \tau$，故存在一对下标 (i, j)，满足 $x_i > x_j$ 且 $y_i < y_j$。构造 $\mathbf{z} = \lambda\mathbf{x} + (1-\lambda)\mathbf{y}$，其中

$$\lambda = \frac{y_j - y_i}{x_i - x_j + y_j - y_i} \in (0, 1).$$

由引理 2.2，L 是凸集，故 $\mathbf{z} \in L$。又因为每个叶节点上有唯一输出，故算法 $\mathcal{A}$ 在 $\mathbf{z}$ 上的输出与 $\mathbf{x}, \mathbf{y}$ 相同，均为 "Yes"。

另一方面，不难验证 $z_i = \lambda x_i + (1-\lambda) y_i = \lambda x_j + (1-\lambda) y_j = z_j$，即 $\mathbf{z}$ 不满足元素唯一性。故算法 $\mathcal{A}$ 在输入 $\mathbf{z}$ 上应该输出 "No"。矛盾！ □

注. 引理 3.2 的一个更加抽象化的证明是：所有的 A_σ 是 $\mathbb{R}^n$ 中互不相交的开集，因此在 $\mathbb{R}^n$ 中互不连通。叶节点 L 是凸集，因此在 $\mathbb{R}^n$ 中连通。因此 L 不能表示成 2 个及以上的 A_σ 的并。

引理 3.1 是引理 3.2 的直接推论。

引理 3.1 的证明. 由引理 3.2，$\mathcal{T}$ 的每个叶结点至多和一个 A_σ 相交，又因为 $\mathcal{T}$ 的所有叶结点给出了 $\mathbb{R}^n$ 的一个划分，故叶节点个数 $\geq |S_n| = n!$。 □

4 元素唯一性问题的复杂度

定理 0.2 的证明. 对所有元素排序，然后检查相邻元素是否相等，即可解决元素唯一性问题。时间复杂度 $O(n \log n)$。

另一方面，对于元素唯一性问题的任意算法 $\mathcal{A}$，按第 1 节的方法构造 $\mathcal{A}$ 对应的决策树 $\mathcal{T}$。则由引理 3.1 知 $\mathcal{T}$ 的叶节点个数至少为 $n!$。从而 $\mathcal{T}$ 的深度至少为 $\log(n!) = \Omega(n \log n)$。

综上所述，元素唯一性问题的复杂度为 $\Theta(n \log n)$。 □

注. 要证明一类问题的复杂度为 $\Theta(T(n))$，必须证明两方面的结果！通过构造算法证明上界 $O(T(n))$，通过下界分析方法证明下界 $\Omega(T(n))$。

注. 元素唯一性问题的 $\Omega(n \log n)$ 下界不仅在基于比较的决策树模型下成立，甚至在一些计算能力更强的决策树模型下（例如，节点处允许做代数运算）也成立，见 [DL79, Ben83, Yao91]。

参考文献

[Ben83] Michael Ben-Or. Lower bounds for algebraic computation trees (preliminary report). In *STOC*, pages 80–86. ACM, 1983. doi:10.1145/800061.808735.

[DL79] David P. Dobkin and Richard J. Lipton. On the complexity of computations under varying sets of primitives. *J. Comput. Syst. Sci.*, 18(1):86–91, 1979. doi:10.1016/0022-0000(79)90054-0.

[Yao91] Andrew Chi-Chih Yao. Lower bounds for algebraic computation trees with integer inputs. *SIAM J. Comput.*, 20(4):655–668, 1991. doi:10.1137/0220041.