保序回归问题

本文作为一个总结性质的东西，不涉及具体的证明，详细请参阅参考资料部分的论文。

序列上的保序回归问题

对给定的数列 $a_1, a_2, \ldots, a_n$ 找到一个非降序列 $b_1, b_2, \ldots, b_n$ 使得

$$ \text{minimize } \sum _{i=1}^n w_i |a_i - b_i|^p, \ p \in \mathbb{N}_+ $$

其中 p 是一个定值，并称其为 L_p 问题。

对问题的一个感性理解：将位置 i 的数增加或减少 Δx 所需代价是 $w_i(\Delta x)^p$ 求让整个数列非降的最小代价。

一些规定：

> **定义 1** 对序列的一个子区间 $[l, r]$ 使 $\sum _{i=l}^r w_i |a_i - k|^p$ 最小的 k 称为该区间的 L_p 均值。

这是说当 $[l, r]$ 区间都只能取同一个值时，将 b_i 全部设为 L_p 均值是最优的。

L_1 问题

> **引理 1** 对于 $w_i \equiv 1$ 的情况 L_1 均值是区间的中位数。

经典结论，在数轴上画一下可以得到。

> **引理 2** 若区间中 $a_l \leq a_{l+1} \geq \cdots \leq a_r$ 则取 $b_i = a_i$ 是最优的。
>
> **引理 3** 若区间中 $a_l \geq a_{l+1} \geq \cdots \geq a_r$ 则取 $b_i = L_1$ 均值是最优的。

因此，序列可以划分为多个非升子序列，将每个子序列都修改为区间中位数即可。若此时仍然存在两个相邻区间不满足非降，则可以将这两个区间合并为一个区间，并将其中位数设为该区间新的 L_1 均值赋值，直到所有区间都满足非降，用栈就能做。

可以证明这样的结果是最优的。通常的实现过程是用大根堆维护区间的一半向上取整的元素，堆顶即为中位数，这样操作的正确性在此略去。如果你想要更直观的正确性，可以用平衡树 + 启发式合并，但会多一个 $O(\log n)$

L_2 问题

> **引理 4** L_2 均值是区间的加权平均数。

对 L_2 的式子求导即可得到。

> **引理 5** 若区间中 $a_l > a_{l+1} > \cdots > a_r$ 则最优解满足 $b_l = b_{l+1} = \cdots = b_r$

证明见论文。这指引我们可以像 L_1 问题一样，当某两个区间的 L_2 均值不满足非降时，就合并两个区间，同时重新计算 L_2 均值赋值。直到所有区间都满足非降。

例子

BOI2004 Sequence

L_1 问题模板题，注意要求 $x_{i+1} > x_i$ 则令 $x'_i = x_i - i$ 就变成了 $x'_{i+1} \geq x_i$ 的条件。

Petrozavodsk Winter 2020. Day 5. Jagiellonian U Contest C - Bookface

题面与题解 见此。要求 $x_{i+1} - x_i \geq d$ 则令 $x'_i = x_i - i \cdot d$ 就变成了 $x'_{i+1} \geq x_i$ 的条件。

参考资料

1. 《浅谈保序回归问题》 - 高睿泉，国家集训队 2018 论文集
2. 《左偏树的特点及其应用》 - 黄源河，国家集训队 2005 论文集

From:
https://wiki.cvbbacm.com/ - **CVBB ACM Team**

Permanent link:
https://wiki.cvbbacm.com/doku.php?id=2020-2021:teams:i_dont_know_png:nikkukun:isotonic_regression

Last update: **2020/08/07 18:11**