毕设课题进度更新 1

11612126 李可明

问题:

Cover Ratio Maximization (user distribution $\theta$):

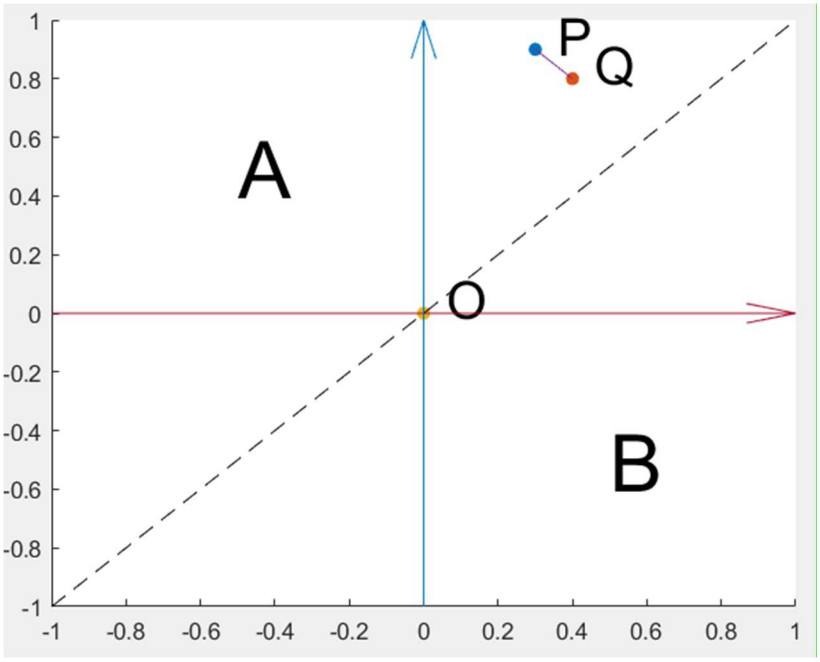
Given a dataset $D = \{d_1, d_2, \dots, d_m\}$, product group $P = \{p_1, p_2, \dots, p_m\}$, product creation budget $B$, a positive integer $k$ and a creation cost function $C$, $CRM$ introduce a new product $p$ such that $C(p) \leq B$ and $cp(p, P, k) = \frac{|\{w|\forall w \in W, \{P \cup \{p\}\} \cap TopK(w) \neq \emptyset\}|}{|W|}$ is maximized when the weight vector $w\ in\ W$ follows a distribution and is presented as probability density function.

对问题的理解

D 是 product 的一个集合， 离散分布
P 是 product 的另一个集合， 离散分布
P 中的某些 product 能在集合$D \cup P$中对于某些 w 排前 k， 一个 product 覆盖一部分 w
我们要找一个新的满足约束 C(p)<B 的 product p， 使得 P∪{p} 覆盖最多的 w

解决思路（下面会用二维的例子举例，后面会讲到如何提升到高维）

定理 1:
对于两个 non-dominated 的 product $P, Q$



如上图， 作 PQ 的垂线 $l_{\perp PQ}$ ， 垂线 $l_{\perp P}$ 将整个空间分成两部分，
对于 A 区域的所有向量都有

$$w_I \cdot P > w_I \cdot Q$$

对于 B 区域的所有向量都有

$$w_{II} \cdot P < w_{II} \cdot Q$$

本问题只关心 A、B 区域在第 I 象限的向量

具体地， 假设 $l_{\perp PQ}$ 的斜率为 $k_{\perp PQ}$，
当 weight vector 斜率在（0, $k_{\perp PQ}$）时，

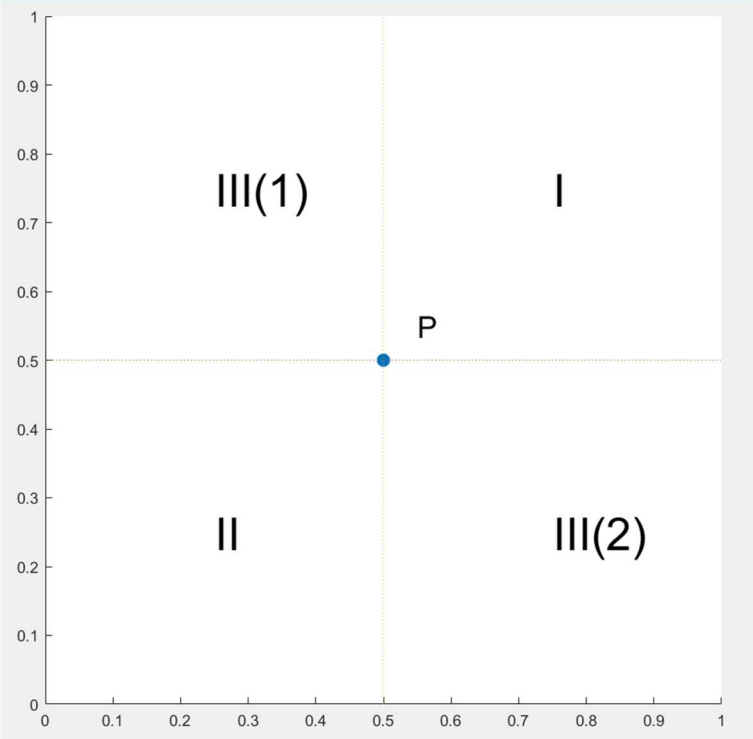
$$w_I \cdot P < w_I \cdot Q$$

当 weight vector 斜率在（$k_{\perp PQ}$，$+\infty$）时，

$$w_{II} \cdot P > w_{II} \cdot Q$$

定义: **两个点的 score 等分线**
两个点的连线的经过原点的垂线
Score 等分线像上图 $l_{\perp PQ}$ *那样将w分成两部分， 一部分P的score更高，另一部分Q的socre更高*

如何判断一个 product 是否存在 weight vector w 使得它排 TopK




如上图， 其他点与点 p 的关系有三种

I: Dominate p

这里的任意点 q 对于任意第一象限的 weight vector w 也就是我们问题要讨论的都有 $w \cdot q > w \cdot p$

II: Dominated by p

这里的任意点 q 对于任意第一象限的 weight vector w 也就是我们问题要讨论的都有 $w \cdot q < w \cdot p$

III: Non-dominated

这里的任意点 q 与 p 都有自己具有优势的 weight vector w, 由 **score 等分线**分割

即对于某些 w， $w \cdot q > w \cdot p$

对另一些 w, $w \cdot q < w \cdot p$

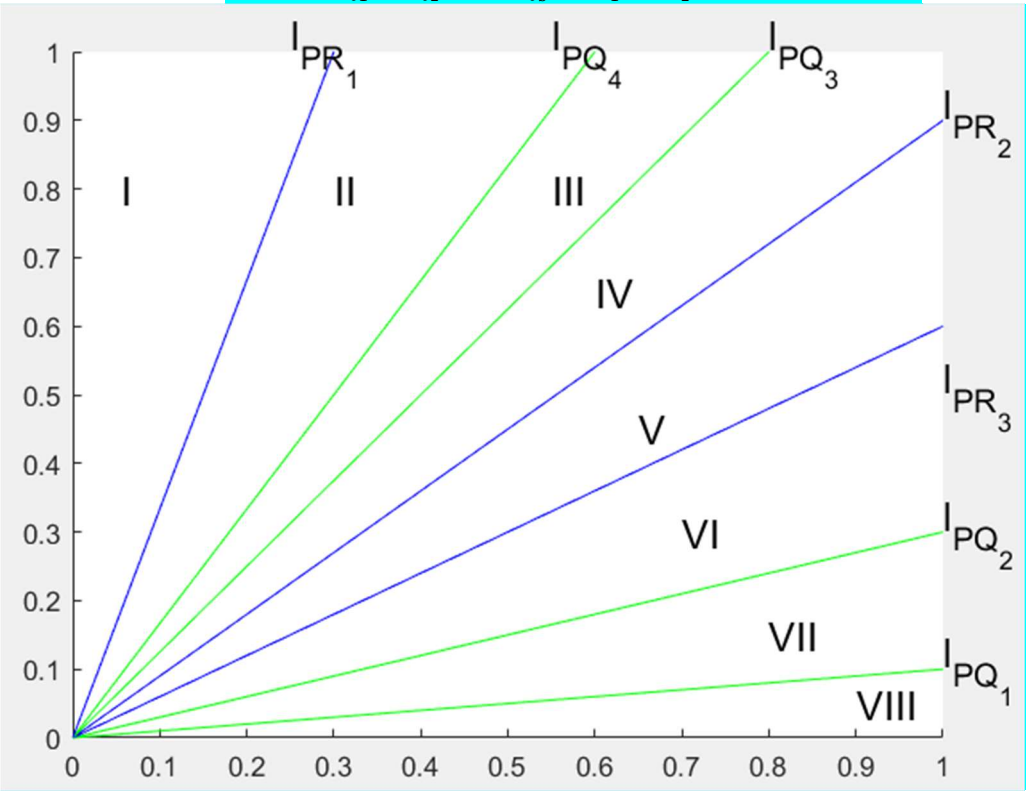
根据三种区域的点与 p 的关系， p 如果要排前 k:

根据定理一， 假设区域 I 有 a 个点，

区域 III(1)有 b 个点, 分别为$Q_1, \dots Q_b$, $w$斜率分别在$(0, k_{\perp PQ_1})$,…, $(0, k_{\perp PQ_b})$时$P$的$score$更高

区域 III(2) c 个点， 分别为$R_1, \dots, R_c$, $w$斜率分别在 $(k_{\perp PR_1}, +\infty)$ ,…, $(k_{\perp PR_c}, +\infty)$ 时$P$的$score$更高

将直线$l_{\perp PQ_1}, l_{\perp PQ_2}, \ldots, l_{\perp PQ_b}, l_{\perp PR_1}, l_{\perp PR_2}, \ldots, l_{\perp PR_b}$画在图上




当$w$位于区域 I, P 的 score 大于点$R_1, R_2, R_3$
当$w$位于区域 II, P 的 score 大于点 $R_2, R_3$
当$w$位于区域 III, P 的 score 大于点$Q_4, R_2, R_3$
当$w$位于区域 IV, P 的 score 大于点$Q_4, Q_3, R_2, R_3$
当$w$位于区域 V, P 的 score 大于点$Q_4, Q_3, R_3$
当$w$位于区域 VI, P 的 score 大于点$Q_4, Q_3$
当$w$位于区域 VII, P 的 score 大于点$Q_4, Q_3, Q_2$
当$w$位于区域 VIII, P 的 score 大于点$Q_4, Q_3, Q_2, Q_1$

定义: 一个 product P 的 non-dominated set $N_P$
一个点的 non-dominated set $N_P$是指所有的与该点具有 non-dominated 关系的点

假设 a = 2(即 dominate P 的点有两个), topK 的 k =6, P 要排前 k, 那么在 w 应该所在的区域要使得 P 在$N_P \cup \{P\}$中排前 4, 只需要 P 在这片区域的 score 能大于$|N_P|$-3 个点即可

如上图$|N_P|$=7, 区域 IV, VIII 满足条件使得点 P 在所有点的集合中排前 k(k=6)

定义: product 一个点 P 的 non-dominated topK region:
能使得点 P 在$N_P \cup \{P\}$排前 k 的 w 的范围

讨论到这里时可以看到我们的方法是不需要知道 w 的分布的，我们找到的是 w 在哪些范围时一个点能排前 k

算法——二维平面, 给定一个 product P, 找到哪些范围 w 能使它排前 k
1. 数出 dominate P 的点的数量, 记为 a
2. 找出所有与 P non-dominated 的点

3. 找到点 P 的 non-dominated top (k-a) region， 这个结果就是能使得 P 排前 k 的 w 的范围

算法——找到一个点 P 的 non-dominated top (k-a) region
讨论：在二维情况下是非常容易找到的，根据斜率将垂线都排好序后， 我们能用时间复杂度非常小的方法完成

从二维到高维——找到一个点 P 的 non-dominated top (k-a) region
二维的例子中，分割一个点 P 和另外一个点 Q(Q 与 P non-dominated)的 score 等分线是直线。而在三维中是一个面，这里用 cell tree 可以找得到超过二维的 P 的 non-dominated top k region 不过此时算法更加暴力无法像二维那样对直线排序然后有一些比较快的方法， 高维的优化只能从剪枝做，剪枝这里应该可以参考之前的工作

尚未解决的：
**如何在约束 C(p)<B 上挑选使得$cp(p,P,k)=\frac{|\{w|\forall w\in W,\{P\cup\{p\}\}\cap TopK(w)\neq\emptyset\}|}{|W|}$最大化的点？？**

现在的情况是能解决 给定一个点能返回这个点覆盖哪些 w
加上给定 w 的概率密度分布 能返回一个 product P 的 hit-probability

但是

个人一直卡在红字的问题， 这个问题看起来像姚新老师或者 hisao 老师做的问题，用遗传之类的算法可以大约解决。这里想了比较长时间没找到能确定结果的方法(能确定自己找到的就是最优的)，个人觉得需要老师的一些帮助
这个问题的难点主要在于
1. 虽然我们能找到某个 product 覆盖哪些范围的 weight vector w，但不同的范围的分布是概率密度函数， 哪些区域更有”覆盖的价值”是个难题
2. 虽然我们能找到某个 product 覆盖哪些范围的 weight vector w，但约束条件 C(p)<B 划定的边界是连续的，有无穷个 product