1 绪论

1.1 常见分布的期望和方差

- 0-1 分布 (伯努利分布): $E = p, Var = p(1-p)$
- 二项分布 (n 个独立的伯努利试验中成功的次数的离散概率分布): $E = np, Var = np(1-p)$
- 泊松分布: $E = \lambda, Var = \lambda$
- 正态分布: $E = \mu, Var = \sigma^2$
- 均匀分布: $E = \frac{a+b}{2}, Var = \frac{(b-a)^2}{12}$

1.2 方差

对于独立随机变量 X_1, X_2. $Var(X_1 + X_2) = Var(X_1) + Var(X_2), Var(nX_1) = n^2 Var(X_1)$

2 尾概率不等式

2.1 Markov 不等式

定理 2.1.1. 非负随机变量 X, $a > 0$

$$P(X \geqslant a) \leqslant \frac{E(X)}{a}$$

2.2 Chebyshev 不等式

定理 2.2.1. 随机变量 X, $r > 0$

$$P(|X - E(X)| \geqslant r) \leqslant \frac{Var(X)}{r^2}$$

P45 题2 令 $X_i (i = 1, 2, \ldots, n)$ 为一组独立同分布的随机变量, 期望 $\mu = E(X_i)$ 和方差 $\delta^2 = E[(X_i - \mu)^2]$. 如果定义 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, 证明

$$P[|\overline{X} - \mu| \geqslant \epsilon|] \leqslant \frac{\delta^2}{n\epsilon^2}$$

证明$\overline{X}$ 的期望和方差为:

$$E(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X) = E(X) = \mu$$

$$Var(\overline{X}) = \frac{1}{n^2} n Var(X_i) = \frac{Var(X_i)}{n} = \frac{\delta^2}{n}$$

根据 Chebyshev 不等式,

$$P[|\overline{X} - E(\overline{X})| \geqslant \epsilon] \leqslant \frac{Var(\overline{X})}{\epsilon^2}$$

$$P[|\overline{X} - \mu| \geqslant \epsilon] \leqslant \frac{\delta^2}{n} \Big/ \epsilon^2$$

$$= \frac{\delta^2}{n\epsilon^2}$$

2.3 Chernoff 不等式

定理 2.3.1. X 二项分布

$$P(X < (1-\delta)\mu) < \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^u$$
$$P(X < (1-\delta)\mu) < \exp\left(\frac{-\mu\delta^2}{2}\right)$$
$$P(X > (1+\delta)\mu) < \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^u$$
$$P(X > (1+\delta)\mu) < \exp\left(\frac{-\mu\delta^2}{4}\right)$$

P45 题4 假设 X_i 独立随机变量满足 $P(X_i = 1) = p_i$ 和 $P(X_i = 0) = 1 - p_i$, 令 $\mu = \sum_{i=1}^{n} p_i$, 定义随机变量 $X = \sum_{i=1}^{n} X_i$. 证明以下结论:

(1) $P(X > (1+\delta)\mu) < \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu}$

(2) $P(X > (1+\delta)\mu) < \exp(-\mu\delta^2/3)$

证明

(1) 对 $t > 0$,

$$\begin{aligned} P(X > (1+\delta)\mu) &= P\big(\exp(tX) > \exp(t(1+\delta)\mu)\big) \\ &< \frac{\prod_{i=1}^{n} E(\exp(tX_i))}{\exp(t(1+\delta)\mu)} \end{aligned}$$

因为 $1 - x < e^{-x}$, 所以

$$\begin{aligned} E(\exp(tX_i)) &= p_i e^t + (1-p_i) = 1 - p_i(1-e^t) \\ &< \exp(p_i(e^t - 1)) \\ \prod_{i=1}^{n} E(\exp(tX_i)) &< \prod_{i=1}^{n} \exp(p_i(e^t-1)) \\ &= \exp(\mu(e^t-1)) \\ P(X > (1+\delta\mu)) &< \frac{\exp(\mu(e^t-1))}{\exp(t(1+\delta)\mu)} \\ &= \exp(\mu(e^t - t - t\delta - 1)) \end{aligned}$$

对 $\mu(e^t - t - t\delta - 1)$ 关于 t 求导, 并令其为 0. 当 $t = \ln(1+\delta)$ 时, $\mu(e^t - t - t\delta - 1)$ 为最小值. 所以

$$P(X > (1+\delta)\mu) < \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu}$$

(2)

$$\begin{aligned}(1+\delta)\ln(1+\delta) &= (1+\delta)\left(\sum_{i=1}^{\infty}(-1)^{i-1}\frac{\delta^i}{i}\right)\\ &> \delta+\frac{\delta^2}{3}-\frac{\delta^2}{6}\\ (1+\delta)^{(1+\delta)} &> \exp(\delta+\frac{\delta^2}{3})\end{aligned}$$

所以,

$$\begin{aligned}P(X>(1+\delta)\mu) &< \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu}\\ &< \left(\frac{e^{\delta}}{\exp(\delta+\frac{\delta^2}{3})}\right)^{\mu}\\ &= \exp\left(\frac{-\mu\delta^2}{3}\right)\end{aligned}$$

2.4 Morris 算法 (计数器)

- 优化过程?

1. 初始化计数器 $X=0$
2. 以 $\frac{1}{2^X}$ 的概率 ++
3. 最终估计值为 2^X-1

2.4.1 Morris+

运行多个 Morris, 结果取平均值

2.4.2 Morris++

运行多个 Morris++, 结果取中位数

2.5 例题

P45 题3 假设抛一枚均匀的硬币 n 次, 随机变量 X 定义为正面朝上的次数.

(1) 运用 Chebyshev 不等式给出事件 $X<\frac{n}{4}$ 的概率上界;

(2) 运用 Chernoff 不等式给出事件 $X<\frac{n}{4}$ 的概率上界.

解 根据题意得 $E(X)=\frac{n}{2}, Var(X)=\frac{n}{4}$

(1) 根据 Chebyshev 不等式:

$$\begin{aligned} P(X < \frac{n}{4}) &= P(X > \frac{3n}{4}) \\ &= P(X - \frac{n}{2} < \frac{n}{4}) \\ &< P(|X - \frac{n}{2}| < \frac{n}{4}) \\ &< \frac{n}{4} \Big/ \frac{n^2}{4} = \frac{4}{n} \end{aligned}$$

(2) 根据 Chernoff 不等式:

$$\begin{aligned} P(X < \frac{n}{4}) &= P(X < (1 - \frac{1}{2})\frac{n}{2}) \\ &< \exp(\frac{-\frac{n}{2}\frac{1}{2}^2}{2}) = \exp(-\frac{n}{16}) \end{aligned}$$

3 哈希

3.1 布隆过滤器

- `insert(v)`: 使用 k 个 `hash`, 在 `hash(v)` 对应的位置标志 1
- `exists(v)`: 使用 k 个 `hash`, `hash(v)` 对应的位置标志全部为 1

性质 3.1.1.

$$(1-a)^b \approx e^{-ab}, a \in [0,1), b > 0$$

3.1.1 误判率

插入 n 个元素后, `exists(v)` 的错误率:

假设每个 hash 函数有 m 个可能结果，且平均分布.

$$\begin{aligned}
\varepsilon &= \overbrace{\prod_{i=1}^{k} P(X_i = 1)}^{\text{所有标志均为 1}} = P(X_i = 1)^k \\
&= (1 - P(X_i = 0))^k \\
&= (1 - \overbrace{(\underbrace{1 - \frac{1}{m}}_{\text{没有被单个 hash 选中}})^{kn}}^{\text{没有被 n 个元素的 k 个 hash 选中}})^k \\
&\approx (1 - e^{-\frac{kn}{m}})^k
\end{aligned}$$

最优解: $k = \ln 2 \cdot \frac{m}{n}$, $m \geqslant 1.44 n \log_2 \frac{1}{\varepsilon}$

3.2 局部敏感哈希 (LSH)

以 Jaccard 距离为例: $Jaccard(A,B) = \frac{|A \cap B|}{|A \cup B|}$

3.2.1 k-Shingling

定义 3.2.1. 文本中 k 个 token 组成的序列. 例如 $D = abcab$, 2-Shingling 为 $\{ab, bc, ca\}$

3.2.2 Min-Hashing

定义 3.2.2. $h_\pi(C) = \min_\pi(C)$: 随机排列 Shingles, 取第一个匹配的 Shingle.

定理 3.2.3. $P(h_\pi(C_1) = h_\pi(C_2)) = Jaccard(C_1, C_2)$

3.2.2.1 最小哈希签名 (矩阵)

定义 3.2.4. 最小哈希签名: 使用 n 个随机排列进行最小哈希。随机排列可以用 `hash(index)` 模拟。

表 3.1: 最小哈希签名矩阵

	d_1	d_2	d_3
π_1	1	3	4
π_2	2	2	1
π_3	5	5	1

3.2.2.2 基于最小哈希的局部敏感哈希

定义 3.2.5. 将最小哈希矩阵分为 b 组 (bands), 每组 r 行. 每组有多个向量对应多个文档, 如果相同代表文档很可能相似. 假设两个集合的 Jaccard 相似度为 s, 没有被映射到同一个桶的概率为 $(1 - s^r)^b$

定义 3.2.6. 相似度阈值 t: $(1 - (1 - t^r))^b = \frac{1}{2}$.

4 抽样算法

4.1 系统抽样

定义 4.1.1. 系统抽样: 先随机抽一个样本, 然后按照一定规律选后面的样本.

定义 4.1.2. 直线 (圆形) 等距抽样: 假设在 N 个中抽 n 个, 在前 $k = \frac{N}{n}$ 中随机抽一个, 然后每隔 k 抽取. $\frac{N}{n}$ 不是整数时可以用圆形列表.

4.2 分层抽样

4.3 水库抽样

定义 4.3.1. 水库抽样: 从数据流中取 k 个数据

1. 先取 k 个进水库
2. 接下来的第 i 个元素以 $\frac{k}{i}$ 的概率替换水库中的任意一个

4.4 例题

P28 题6 从词库中随机等概率抽取 100 个单词，请写出水库抽样的主要步骤。若被告知词库容量为 10000，请设计合理的抽样方法抽取由 100 个单词组成的样本。

解

(a) 水库抽样的主要步骤:

1. 将数据流中的前 100 条记录保存下来
2. 对于第 m 条记录 ($m > k$), 以 $\frac{k}{m}$ 的概率决定是否替换水库中的一条概率
3. 重复执行第 2 步, 直至数据流终止

(b) 若知道词库容量为 10000, 可以使用直线等距抽样。

$$k = \frac{10000}{100} = 100$$

在 $1 \sim 100$ 中随机抽取一个变化 r。取第 $r + 100k$ 个数据, 其中 $1 \leqslant r + 100k \leqslant 10000$。

P28 题15 在水库抽样算法中，当第 i 个元素到达时，以 $\frac{1}{i}$ 的概率替换水库中选定的某个元素，直到最后一个元素到达为止。试证明每个元素被选中的概率相等，即为 $\frac{1}{n}$。

解

(a) 当输入第 1 个元素时, 该元素必被选中 ($\frac{1}{i} = 1$)。

(b) 当输入第 i 个元素时, 该元素被选中的概率为 $\frac{1}{i}$。

(c) 当输入第 $i+1$ 个元素，只有当第 $i+1$ 个元素不被选中时, 第 i 个元素才会被保留。因此第 i 个元素仍保留在水库中的概率为:

$$\frac{1}{i}(1-\frac{1}{i+1})=\frac{1}{i+1}$$

因此根据归纳假设，对于长度 n 的数据，水库抽样能保证每条记录以 $\frac{1}{n}$ 的概率保留在水库中，即每个元素被选中的概率相等。

5 Sketch

5.1 数据流模型

根据对实际数组的影响不同:

1. 时间序列模型: 每次来一个新数据
2. 收银机模型: 每次输入一个增量
3. 十字转盘模型: 每次输入一个 diff

根据元素重要度:

1. 界标模型: 某一时间 t 到现在
2. 滑动窗口模型: W 窗口大小
3. 衰减窗口模型: 新到达的重要程度较高

5.2 Misra Gries

使用 k 个计数器, 若计数器满了, 所有计数器减 1

5.3 Count Sketch

5.3.1 简单抽样算法

按照概率 p 存入元素, 最终概率为 $\frac{c_i}{p}$, c_i 为存入的元元素个数.

5.3.2 Basic Count Sketch

两个 hash 函数, 一个控制位置, 另一个控制 +1 还是-1

5.3.3 Count Sketch

多个 Basic Count Sketch 取中位数

5.3.3.1 例题

例 5.3.1 证明 $\bar{f}_a = \text{median}_{1\leqslant i\leqslant t} g_i(a)C[i][h_i(a)]$ 中 $t = O(log(1/\delta))$

解

$E(\bar{f}_a) = f_a, Var(\bar{f}_a) = \frac{||f_{-a}||_2^2}{k}$

$P(|\bar{f}_a - f_a| \geqslant \epsilon||f||_2) \leqslant P(|\bar{f}_a - f_a| \geqslant \epsilon||f_{-a}||_2) \leqslant \frac{Var(\bar{f}_a)}{\epsilon^2||f_{-a}||_2^2} = \frac{1}{k\epsilon^2}$

令 $\frac{1}{k\epsilon^2}=\frac{1}{3}$, 定义:

$$Y_i=\begin{cases}1 & |\bar{f}_a-f_a|\geqslant\epsilon||f||_2\\ 0\end{cases}$$

则有 $P(Y_i=1)\leqslant\frac{1}{3}$, 记 $\mu=E(\sum_{i=1}^{t}Y_i)\leqslant\frac{t}{3}$

$P(\sum_{i=1}^{t}Y_i>\frac{t}{2})\leqslant P(\sum_{i=1}^{t}Y_i>(1+\frac{1}{2})\mu)\leqslant exp(-\frac{\mu}{16})<\delta$

$exp(-\frac{t}{48})\leqslant exp(-\frac{\mu}{16})<\delta$

例 5.3.2 $\bar{f}_a=\frac{g_i(a)C[i][h_i(a)]}{t}$, 分析性能

解

$Var(\bar{f}_a)=\frac{||f_{-a}||_2^2}{tk}$

$P(|\bar{f}_a-f_a|\geqslant\epsilon||f||_2)\leqslant P(|\bar{f}_a-f_a|\geqslant\epsilon||f_{-a}||_2)\leqslant\frac{Var(\bar{f}_a)}{\epsilon^2||f_{-a}||_2^2}=\frac{1}{tk\epsilon^2}<\delta$

因此 $t=O(\frac{1}{\epsilon^2\delta})$

5.3.4 Count-min Sketch

Count Sketch 基础上, 改为取最小值, 只增. 计数器个数: $O(\log(1/\delta)/\epsilon)$

6 随机游走

- 马尔科夫性: 只与上一时刻的状态有关
- 可约的 (irreducible): 强连通
- 状态 x 的周期: $\{n|P^n_{x,x} > 0\}$ 的公约数
- 反周期 (aperiodic): 所有状态周期为 1
- PageRank: $pr^T = pr^T \cdot P$ 直至平稳
  - PageRank 的改进: $P' = \beta P + (1-\beta)[\frac{1}{n}]_{n\times n}$

7 整数规划

7.1 表示约束

7.1.1 或

例 7.1.1 $x+y \leqslant 3$ or $x-y \geqslant 4$

解 引入足够大的数 M 和 $\omega \in \{0,1\}$.

$$\begin{cases} x+y \leqslant 3+M\omega \\ x-y \geqslant 4-M(1-\omega) \end{cases}$$

7.1.2 分段函数

P218 题3 将下面的非线性规划问题重写成整数规划问题。

$$f(x)=\begin{cases} 10x, & \text{如果} 0 \leqslant x \leqslant 50 \\ 500, & \text{如果} 51 \leqslant x \leqslant 100 \\ 5x, & \text{如果} x \geqslant 101 \end{cases}$$

解 引入 0-1 变量 w_1, w_2, w_3, 整数变量 x_1, x_2, x_3, 满足:

$$w_1=\begin{cases} 1 & 0 \leqslant x \leqslant 50 \\ 0 & \text{其他} \end{cases} \quad x_1=\begin{cases} x & 0 \leqslant x \leqslant 50 \\ 0 & \text{其他} \end{cases}$$

$$w_2=\begin{cases} 1 & 51 \leqslant x \leqslant 100 \\ 0 & \text{其他} \end{cases} \quad x_2=\begin{cases} x & 51 \leqslant x \leqslant 100 \\ 0 & \text{其他} \end{cases}$$

$$w_3=\begin{cases} 1 & 101 \leqslant x \\ 0 & \text{其他} \end{cases} \quad x_3=\begin{cases} x & 101 \leqslant x \\ 0 & \text{其他} \end{cases}$$

对应约束条件可改写为:

$$
\begin{aligned}
f(x) =& 10x_1 + 500w_2 + 5x_3 \\
& 0 \leqslant x_1 \leqslant 50w_1 \\
& 51w_2 \leqslant x_2 \leqslant 100w_2 \\
& 101w_3 \leqslant x_3 \\
& w_1, w_2, w_3 \in \{0, 1\} \\
& x_1 + x_2 + x_3 = x \\
& x_1, x_2, x_3 \in Z
\end{aligned}
$$

7.2 线性规划

7.2.1 标准型

所有的约束都是不等式

例 7.2.1

$$
\begin{aligned}
\min \ & -2x_1 + 3x_2 \\
\text{约束条件 } & x_1 + x_2 = 7 \\
& x_1 - 2x_2 <= 4 \\
& x_1 >= 0
\end{aligned}
$$

解

1. 目标函数可能是最小化，而不是最大化

$$
\min -2x_1 + 3x_2 \rightarrow \max 2x_1 + 3x_2
$$

2. 可能有变量不具有非负约束. 将 x_2 改为 $x_2' - x_2''$

$$
\begin{aligned}
\text{约束条件 } & x_1 + x_2' - x_2'' = 7 \\
& x_1 - x_2' - x_2'' <= 4 \\
& x_1, x_2', x_2'' >= 0
\end{aligned}
$$

3. 可能有等式约束. 转化成一组不等式

$$
x_1 + x_2' - x_2'' = 7 \rightarrow x_1 + x_2' - x_2'' \geqslant 7 x_1 + x_2' - x_2'' \leqslant 7
$$

4. 可能有不等式约束，但不是小于等于号，而是大于等于号

最终结果:

$$\begin{aligned}&\max\ 2x_1-3(x_2'-x_2'')\\ \text{约束条件}&=x_1+x_2'-x_2''\geqslant 7\\ &=x_1+x_2'-x_2''\leqslant 7\\ &x_1,x_2',x_2''>=0\end{aligned}$$

7.2.2 松弛型

约束都是等式（除了要求变量非负的约束）. 为每一个标准型中的不等式引入一个变量. 例如:

$$x+y\geqslant 3\rightarrow z=x+y-3,z\geqslant 0$$

7.2.3 单纯型法

See wikipedia.

7.3 例题

P218 题6 给定下面的整数规划问题:

$$\begin{aligned}&IP(1)\\ &\max\ 10x_1+4x_2+9x_3\\ \text{约束条件:}\ &5x_1+4x_2+3x_3\leqslant 9\\ &0\leqslant x_i\leqslant 1,x_i\in Z,1\leqslant i\leqslant 3\end{aligned}$$

1. 写出关于 $IP(1)$ 的线性规划松弛。
2. 如果 $x_1=1$, 试找出对应整数规划问题的上界。
3. 使用分支定界方法求解 $IP(1)$

解

1. 写出关于 $IP(1)$ 的线性规划松弛。

$$\begin{aligned}&\max\ 10x_1+4x_2+9x_3\\ \text{约束条件:}\ &5x_1+4x_2+3x_3\leqslant 9\\ &0\leqslant x_i\leqslant 1,1\leqslant i\leqslant 3\end{aligned}$$

2. 如果 $x_1 = 1$, 试找出对应整数规划问题的上界。

   当 $x_1 = 1$ 时, $IP(1)$ 可化为:

$$\begin{aligned} &\max\ 10 + 4x_2 + 9x_3 \\ &\text{约束条件: } 4x_2 + 3x_3 \leqslant 4 \\ &\qquad 0 \leqslant x_i \leqslant 1, x_i \in Z, 2 \leqslant i \leqslant 3 \end{aligned}$$

   此时 (x_2, x_3) 的可行域为 $\{(0,0),(1,0),(0,1)\}$。易得当 $(x_2, x_3) = (0,1)$ 时 $10 + 4x_2 + 9x_3$ 取最大值 19。

3. 使用分支定界方法求解 $IP(1)$

   令 $Z_I = -\infty$

   $LP(1)$ 的最优解为:

$$x_1 = 1, x_2 = \frac{1}{4}, x_3 = 1, Z_{LP(1)} = 20$$

   由于 $Z_{LP(1)} > Z_I$ 且非整数解，选择 $x_1 = 1$ 时的节点 $IP(2)$ 和 $x_1 = 0$ 时的节点 $IP(3)$。

   由题 2 可得, $IP(2)$ 的最优解为:

$$x_1 = 1, x_2 = 0, x_3 = 1, Z_{IP(2)} = 19$$

   $Z_{IP(2)} > Z_I$, 将 Z_I 设置为 19。

   求 $IP(3)$, $IP(3)$ 可表示为:

$$\begin{aligned} &\max\ 4x_2 + 9x_3 \\ &\text{约束条件: } 4x_2 + 3x_3 \leqslant 9 \\ &\qquad 0 \leqslant x_i \leqslant 1, x_i \in Z, 2 \leqslant i \leqslant 3 \end{aligned}$$

   此时 (x_2, x_3) 的可行域为 $\{(0,0),(1,0),(0,1),(1,1)\}$。易得当 $(x_2, x_3) = (1,1)$ 时 $4x_2 + 9x_3$ 取最大值 13，即 $Z_{IP(3)} = 13$，小于当前 Z_I。

   算法终止，$IP(1)$ 的最优解为:

$$x_1 = 1, x_2 = 0, x_3 = 1, Z_{IP(1)} = 19$$

8 子模函数及其应用

8.1 子模函数

$$
\begin{aligned}
& f(S)+f(T) \geqslant f(S \cup T)+f(S \cap T) \\
\Leftrightarrow\ & f(S \cup\{v\])-f(S) \geqslant f(T \cup\{v\})-f(T), S \subseteq T \subseteq V, v \in V \backslash T \\
\Leftrightarrow\ & f(S \cup C)-f(S) \geqslant f(T \cup C)-f(T), S \subseteq T \subseteq V, C \subseteq V \backslash T
\end{aligned}
$$

8.2 例题

P232 题12 假设全集 $U=\{1,2,3,4,5\}$ 和子集族

$$S=\{\{1,2\},\{1,2,4\},\{2,3\},\{4,5\},\{3\},\{1,5\},\}$$

当 $k=2$ 时，运用爬山算法求解最大覆盖问题，可能的解有哪些?

解 第 1 次迭代，显然 S_2 的边界覆盖最大，为 3。$S=\{1,2,3\}$, $f(S)=3$。

第 2 次迭代，可选子集为 S_3, S_4, S_5, S_6, 边界贡献均为 1。

子集	$f(S)$	$f(S \cup S_i)$	$\delta(S_i)$
S_1	3	3	0
S_3	3	4	1
S_4	3	4	1
S_5	3	4	1
S_6	3	4	1

表 8.1: 第二轮迭代

综上，可能的解有 $\{S_2,S_3\},\{S_2,S_4\},\{S_2,S_5\},\{S_2,S_6\}$。

P232 题2 给定下面的全集和子集族, 当 $k=3$ 时, 使用爬山算法求解最大覆盖问题.

$$
\begin{aligned}
\text{全集}: & \{a,b,c,d,e,f,g,h,i,j,k,l\} \\
\text{子集族}: & A_1=\{b,c,d\}, A_2=\{e,f,g\}, A_3=\{i,j,k,l\} \\
& A_4=\{a,e,i\}, A_5=\{i,b,g\}, A_6=\{c,d,g,h,k,l\} \\
& A_7=\{a,l\}, A_8=\{a,e,i\}
\end{aligned}
$$

解 第 1 轮迭代, 显然 A_6 的边际覆盖最大, 为 6. $S=\{c,d,g,h,k,l\}$, $f(S)=6$.

第 2 轮迭代, A_4 和 A_8 的边际贡献最大. 选择 A_4, $S = \{a, c, d, e, g, h, i, k, l\}$, $f(S) = 9$.

子集	$f(S)$	$f(S \cup A_i)$	$\Delta(A_i)$
A_1	6	7	1
A_2	6	8	2
A_3	6	8	2
A_4	6	9	3
A_5	6	8	2
A_7	6	7	1
A_8	6	9	3

第 3 轮迭代, A_1, A_2, A_3, A_5 的边际贡献最大. 选择 A_1, $S = \{a, b, c, d, e, g, h, i, k, l\}$, $f(S) = 10$.

子集	$f(S)$	$f(S \cup A_i)$	$\Delta(A_i)$
A_1	9	10	1
A_2	9	10	1
A_3	9	10	1
A_5	9	10	1
A_7	9	9	0
A_8	9	9	0

最终输出为 A_6, A_4, A_1.

P232 题6 令 $F = \{1, \ldots, n\}$ 和 $D = \{1, \ldots, m\}$ 分别表示 n 个设备和 m 个客服组成的集合. 设备 i 提供服务价值 v_{ij} 给客服 j. 假定 $S \subset F$, 验证

$$f(S) = \sum_{j \in D} \max_{i \in S} v_{ij} \tag{8.1}$$

是一个子模函数.

证 令 $\forall S \subseteq T \subseteq F$, $\forall w \in F \backslash T$.

$$\begin{aligned}
f(S) &= \sum_{j \in D} \max_{i \in S} v_{ij} \\
f(S \cup \{w\}) &= \sum_{j \in D} \max_{i \in S \cup \{w\}} v_{ij} \\
f(S \cup \{w\}) - f(S) &= \sum_{j \in D} \begin{cases} v_{wj} - \max_{i \in S} v_{ij} & v_{wj} > \max_{i \in S} v_{ij} \\ 0 & else \end{cases}
\end{aligned}$$

$f(T \cup \{w\})$ 类似, 由上式可得:

$$(f(S\cup\{w\})-f(S))-(f(T\cup\{w\})-f(T))=$$
$$\sum_{j\in D}\left(\begin{cases}v_{wj}-\max_{i\in S}v_{ij} & v_{wj}>\max_{i\in S}v_{ij}\\ 0 & else\end{cases}-\begin{cases}v_{wj}-\max_{i\in T}v_{ij} & v_{wj}>\max_{i\in T}v_{ij}\\ 0 & else\end{cases}\right) \tag{8.2}$$

对于 $\forall j\in D$:

由假设可得:

$$\max_{i\in S}v_{ij}\leqslant\max_{i\in T}v_{ij} \tag{8.3}$$

当 $v_{wj}>\max_{i\in S}v_{ij}$, $v_{wj}>\max_{i\in T}v_{ij}$ 时: $(v_{wj}-\max_{i\in S}v_{ij})-(v_{wj}-\max_{i\in T}v_{ij})\geqslant 0$.

当 $v_{wj}>\max_{i\in S}v_{ij}$, $v_{wj}\leqslant\max_{i\in T}v_{ij}$ 时: $(v_{wj}-\max_{i\in S}v_{ij})-0>0$.

当 $v_{wj}\leqslant\max_{i\in S}v_{ij}$, $v_{wj}>\max_{i\in T}v_{ij}$ 时: 与8.3冲突.

当 $v_{wj}\leqslant\max_{i\in S}v_{ij}$, $v_{wj}\leqslant\max_{i\in T}v_{ij}$ 时: $0-0=0$

综上所述, 式8.2总是大于等于 0. 即 $f(S\cup\{w\})-f(S)\geqslant f(T\cup\{w\})-f(T)$

因此, 式8.1是一个子模函数.

P233 题10 设 $f_i(A)$ 为子模函数, 而且对 $i=1,2,\ldots,n$ 都有 $a_i\geqslant 0$, 试证明 $\sum_{i=1}^{n}a_if_i(A)$ 也是一个子模函数.

证 令 $g(A)=\sum_{i=1}^{n}a_if_i(A)$.

$$g(S)+g(T)=\sum_{i=1}^{n}a_i(f_i(S)+f_i(T))$$
$$g(S\cup T)+g(S\cap T)=\sum_{i=1}^{n}a_i(f_i(S\cup T)+f_i(S\cap T))$$

对任意 $i=1,2,\ldots,n$, $f_i(S)+f_i(T)\geqslant f_i(S\cup T)+f_i(S\cap T)$ 成立, 并且 $a_i\geqslant 0$. 所以 $a_i(f_i(S)+f_i(T))\geqslant a_i(f_i(S\cup T)+f_i(S\cap T))$.

因此 $g(S)+g(T)\geqslant g(S\cup T)+g(S\cap T)$, $g(A)$, 即 $\sum_{i=1}^{n}a_if_i(A)$ 是一个子模函数.

9 社区发现

9.1 模块度

9.2 例题

P254 题4 设有权图 G 的权重矩阵为

$$W=\begin{pmatrix}0&2&3\\2&0&0\\3&0&0\end{pmatrix}$$

(1) 如果两个社区分别为 $C_1=\{1,3\}$ 和 $C_2=\{2\}$, 试使用

$$Q=\frac{1}{2m}\sum_{s\in C}\sum_{i\in c}\sum_{j\in c}\left(W_{ij}-\frac{k_ik_j}{2m}\right)$$

计算模块度。

(2) 社区结构如 (1) 相同，试使用

$$Q=\sum_{c\in C}\left[\frac{\sum_{in}^{c}}{2m}-\left(\frac{\sum_{tot}^{c}}{2m}\right)^2\right]$$

计算模块度。

(3) 社区结构如 (1) 相同，试使用

$$Q=\frac{1}{4m}s^TBs$$

计算模块度。其中 $B_{ij}=W_{ij}-\frac{k_ik_j}{2m}$。

解

(1)

$$\begin{aligned}Q&=\frac{1}{2m}\sum_{s\in C}\sum_{i\in c}\sum_{j\in c}\left(W_{ij}-\frac{k_ik_j}{2m}\right)\\&=\frac{1}{2\times5}\left[\left(0-\frac{5\times5}{2\times5}\right)+2\left(3-\frac{5\times3}{2\times5}\right)+\left(0-\frac{3\times3}{2\times5}\right)+\left(0-\frac{2\times2}{2\times5}\right)\right]\\&=-0.08\end{aligned}$$

(2)

$$
\begin{aligned}
Q &= \sum_{c \in C} \left[\frac{\sum_{in}^{c}}{2m} - \left(\frac{\sum_{tot}^{c}}{2m} \right)^2 \right] \\
&= \left[\frac{3+3}{2 \times 5} - \left(\frac{3+5}{2 \times 5} \right)^2 \right] + \left[\frac{0}{2 \times 5} - \left(\frac{2}{2 \times 5} \right)^2 \right] \\
&= -0.08
\end{aligned}
$$

(3)

$$
s = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}, B = \begin{pmatrix} -2.5 & 1 & 1.5 \\ 1 & -0.4 & -0.6 \\ 1.5 & -0.6 & -0.9 \end{pmatrix}
$$

$$
\begin{aligned}
Q &= \frac{1}{4m} s^T B s \\
&= \frac{1}{20} \times \begin{pmatrix} 1 & -1 & 1 \end{pmatrix} \begin{pmatrix} -2.5 & 1 & 1.5 \\ 1 & -0.4 & -0.6 \\ 1.5 & -0.6 & -0.9 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} \\
&= -0.08
\end{aligned}
$$

9.3 社区发现算法

1. 谱方法
2. Louvain