Problem Set2:Solutions

MF20330007 陈明远

2020 年 11 月 17 日

Problem1:
Solution:
Let$Y = \sum_{i=1}^{s} Y_i$,defineY_i:

$$Y_i = \begin{cases} 1 & \text{the output of the Algorithm lies in correct ranges} \\ 0 & \text{others} \end{cases}$$

Let $\mu = \sum_{i=1}^{s} Y_i$,so we can get:$\mathbb{E}[Y_i] = \frac{3}{4}$,$\mu = \mathbb{E}[\sum_{i=1}^{s} Y_i] = \frac{3}{4}s$(since the linearity of expectation.),when the $\sum_{i=1}^{s} Y_i \geq \frac{s}{2}$, According to the nature of the median,the X is correct,so we gan get:

$$\begin{aligned} Pr[Y \leq \frac{s}{2}] &= Pr[Y \leq \frac{2}{3}\mu](s = \frac{4}{3}\mu) \\ &= Pr[Y \leq (1 - \frac{1}{3})\mu] \end{aligned}$$

Using Chernoff bound:

$$\begin{aligned} &\leq exp\left[-\frac{\mu}{2} * (\frac{1}{3})^2\right] \\ &= e^{-\frac{1}{18}\mu} \\ &= e^{-\frac{1}{24}s} \\ &\leq \delta \end{aligned}$$

so we can get:

$$s \geq -24ln\delta$$

s can be as small as $\lceil -24ln\delta \rceil$

Problem2:
Solution:

- (i):Since s=0, it means that each i is mapped to greater than $\frac{1}{2T}$,so we can get:

$$Pr_h[s=0] = Pr_h[\forall i_{1\le i\le N}, h(i) \ge \frac{1}{2T}]$$
$$= \prod_{i=1}^{N} Pr_h[h(i) \ge \frac{1}{2T}]$$
$$\le \prod_{i=1}^{N} Pr_h[h(i) \ge \frac{1}{||x||_0}](\text{Since } T < \frac{1}{2}||x||_0)$$
$$\le \prod_{i=1}^{N} Pr_h[h(i) \ge \frac{1}{N}](||x||_0 \le N)$$
$$= \prod_{i=1}^{N} 1 - \frac{1}{N+1}$$
$$= (1 - \frac{1}{N+1})^N$$
$$< \frac{1}{e} \approx 0.37$$

- (ii):same as above,we can get:

$$Pr_h[s=0] = Pr_h[\forall i_{1\le i\le N}, h(i) \ge \frac{1}{2T}]$$
$$= \prod_{i=1}^{N} Pr_h[h(i) \ge \frac{1}{2T}]$$
$$\ge \prod_{i=1}^{N} Pr_h[h(i) \ge \frac{1}{4N}](\text{Since } T > 2N \ge 2||x||_0)$$
$$= \prod_{i=1}^{N} 1 - \frac{1}{4(N+1)}$$
$$> e^{-\frac{1}{4}}$$
$$> 2^{-\frac{1}{2}} > 0.5$$

- Algorithm:

1.scan the input sequence$x_1, x_2, ..., x_n \in 1, ..., N$ in a single pass to compute:
2.$s_j(1 \leq j \leq k) = \sum_{i \in 1,...,N: h(i) < \frac{1}{2T}} x_i$
3.$S = \sum_{j=1}^{k} s_i$
4.if:$S = 0$ T is **HIGH**

- Algorithm:

Problem3:
Solution:

- **Proof.** The original formula can be transformed into:

$$1 - sim(A, B) + 1 - sim(B, C) \geq 1 - sim(A, C)$$
$$\Rightarrow 1 \geq sim(A, B) + sim(B, C) - sim(A, C)$$

Recall that:

$$sim(A, C) = Pr_{h \in \mathcal{F}}[h(A) = h(C)]$$
$$= Pr_{h \in \mathcal{F}}[h(A) = h(B) \cap h(B) = h(C)]$$
$$= Pr_{h \in \mathcal{F}}[h(A) = h(B)] * Pr_{h \in \mathcal{F}}[h(B) = h(C)]$$
$$= sim(A, B) * sim(B, C)$$

so we only need to prove:$1 \geq sim(A, B) + sim(B, C) - sim(A, B) * sim(B, C)$,using the mean inequality,we only need to prove:

$$1 \geq sim(A, B) + sim(B, C) - \frac{(sim(A, B) + sim(B, C))^2}{4}$$

among them: $sim(A, B) \in [0, 1], sim(B, C) \in [0, 1]$.
Let $f(x) = x - \frac{x^2}{4}, x \in [0, 2]$,we can get:

$$f(x)_{max} = f(2) = 1$$

so the original inequality is correct. so the distance function satisfies triangle inequality.

- **Proof.** (i)If there is a hash function family satisfies Dice's coefficient, it must satisfy the following formula:

$$sim_{Dice}(A, B) + sim_{Dice}(B, C) - sim_{Dice}(A, C) \leq 1$$

Because of locality sensitive hash function family must satisfy the triangle inequality (according to question1),but $\frac{2|A \cap B|}{|A|+|B|} + \frac{2|B \cap C|}{|B|+|C|} - \frac{2|A \cap C|}{|A|+|C|}$ not always $\leq 1(|A \cap C| = 0)$,so there is no locality sensitive hash function family corresponding to Dice's coefficient.
(ii) Same as above,

$$\begin{aligned} & sim_{ovl}(A, B) + sim_{ovl}(B, C) - sim_{ovl}(A, C) \\ & = \frac{|A \cap B|}{min(|A|, |B|)} + \frac{|B \cap C|}{min(|B|, |C|)} - \frac{|A \cap C|}{min(|A|, |C|)} \end{aligned}$$

it not always satisfy the triangle inequality(eg.$A = \{a, b, c\}, B = \{b, d\}, C = \{d\}, |A \cap C| = 0$),so so there is no locality sensitive hash function family corresponding to Overlap coefficient.

- **Proof.** from this problem,we can compute by the following formula:

$$\begin{aligned} & Pr_{h \in \mathcal{F}'}[h'(A) = h'(B)] \\ & = Pr_{h \in \mathcal{F}}[h(A) = h(B)] * Pr_{f \in \mathcal{B}}[f(x) = f(y)|x = y] \\ & + Pr_{h \in \mathcal{F}}[h(A) \neq h(B)] * Pr_{f \in \mathcal{B}}[f(x) = f(y)|x \neq y] \\ & = sim(A, B) * 1 + (1 - sim(A, B)) * \frac{1}{2} \\ & = \frac{1 + sim(A, B)}{2} \end{aligned}$$

,so the locality sensitive hash function family $\mathcal{F}'$corresponding to the similarity function $\frac{1+sim(A,B)}{2}$.

Problem4:

Solution:

1.Algorithm:

Input:an undirected graph G(V,E),

with an with an arbitrary order of vertices $V = \{v_1, v_2, ..., v_n\}$

---

initially $S_{i:1\leq i\leq k} = \emptyset$

for $i = 1, 2, ..., n$

  v_i joins one of $S_1, S_2, ..., S_k$ to maximize the current $W = \sum_{\exists i\neq j:u\in S_i,v\in S_j} w_{uv}$

return W

Proof.

Let OPT_G denote the weighted of the max k-cut

  $OPT_G = maxW$

Let SOL_G denote the W returned by the GreedyAlgorithm

  $SOL_G = \sum_{j=1}^{n} \underset{1\leq i\leq k}{max}\, w(S_{ji}, v_{ji})$

2.The blank should be filled:

$$\sum_{u\in S_{i-1},v\in S_i} w(v,u) \leq \sum_{u\in S_i,v\in S_{i-1}} w(v,u)$$

Proof.

Problem5.
Solution:

(a):**Proof.**If the original formula holds,wo only need to prove

$$\sum_{i=1}^{k}\sum_{j\in C_i}(\mathbf{x}_j-\mu_i)^T(\mathbf{x}_j-\mu_i)=cost(\mathbf{x}_1,...,\mathbf{x}_n,C_1,...,C_k)$$

we can make some transformations to the left formula:

$$\begin{aligned}&\sum_{j\in C_i}(\mathbf{x}_j-\mu_i)^T(\mathbf{x}_j-\mu_i)\\&=\sum_{j\in C_i}(\mathbf{x}_j^T\mathbf{x}_j-2\mu_i\mathbf{x}_j^T+\mu_i^T\mu_i)\\&=\sum_{j\in C_i}[\mathbf{x}_j^T\mathbf{x}_j-\frac{2}{|C_i|}\mathbf{x}_j^T\sum_{j\in C_i}\mathbf{x}_j+\frac{1}{|C_i|^2}\sum_{j\in C_i}\mathbf{x}_j^T\sum_{k\in C_i}\mathbf{x}_k]\end{aligned}$$

we also get:

$$\begin{aligned}&\sum_{i=1}^{k}\frac{1}{|C_i|}\sum_{j,l\in C_i,j<l}||\mathbf{x}_j-\mathbf{x}_l||_2^2\\&=\sum_{i=1}^{k}\frac{1}{|C_i|}\sum_{j,l\in C_i,j<l}[\mathbf{x}_j^T\mathbf{x}_j-2\mathbf{x}_j\mathbf{x}_l^T+\mathbf{x}_l^T\mathbf{x}_l]\end{aligned}$$

So they are equal

(b)**Proof.**Let $\mathbf{cost}(\hat{\mathbf{X}},\mathbf{C})$denote $\mathbf{cost}(\hat{\mathbf{x}}_1,...,\hat{\mathbf{x}}_n,C_1,...,C_k)$ Using the conclusion of (a),if we want to prove:

$$\mathbf{cost}(\hat{\mathbf{X}},\mathbf{C})\le\mathbf{cost}(\mathbf{X},\mathbf{C})$$

This can be directly derived from (a):

$$\sum_{i=1}^{k}\frac{1}{|C_i|}\sum_{j,l\in C_i,j<l}||\hat{\mathbf{x}}_j-\hat{\mathbf{x}}_l||_2^2\le\sum_{i=1}^{k}\frac{1}{|C_i|}(1+\epsilon)\sum_{j,l\in C_i,j<l}||\mathbf{x}_j-\mathbf{x}_l||_2^2$$

the left is similar.