1.

假设 $\delta,\theta \in (0,1)$, 给定集合 U 以及其一个子集 V, $|V|/|U| = a$

(1).

如果我们从 U 均匀采样本 S, 并确保 $S \cap V \neq \varnothing$ 的概率大于 $1-\delta$, $|S|$ 至少要多大?
考虑有放回采样:

$$\begin{aligned} P(S \cap V \neq \varnothing) &= 1 - P(S \cap V = \varnothing) \\ &= 1 - \frac{|U| - |V|}{|U|}^{|S|} \\ &= 1 - (1-a)^{|S|} \\ &> 1 - \delta \end{aligned}$$

可知, $|S|$ 应当满足 $|S| > \log_{(1-a)} \delta$

(2).

如果我们要确保 $|S \cap V| \in (1 \pm \theta)a|S|$ 的概率大于 $1-\delta$, $|S|$ 至少要多大?
显然, 随机抽样在 V 里的概率为 a, 可以考虑 $|S|$ 个独立同分布于 $B(1,a)$ 的随机变量 $X_1, \cdots, X_{|S|}$
即对 $i = 1, \cdots, |S|$ 有

$$P(X_i = k) = \begin{cases} a & ,k = 1 \\ 1-a & ,k = 0 \end{cases}$$

那么 $P(|S \cap V| = k) = P(\sum_{i=1}^{|S|} X_i = k)$, 且 $E[|S \cap V|] = a|S|$
由 Chernoff Bounds:

$$\begin{aligned} P(|S \cap V| \notin (1 \pm \theta)a|S|) &= P(|S \cap V| \notin (1 \pm \theta)E[|S \cap V|]) \\ &= P\left(\left|\sum_{i=1}^{|S|} X_i\right| \notin a|S| \pm a\theta|S|\right) \\ &\leq 2\exp(-2|S|(a\theta)^2) \end{aligned}$$

即 $P(|S \cap V| \in (1 \pm \theta)a|S|) > 1 - 2\exp(-2|S|(a\theta)^2)$
故要想$P(|S \cap V| \in (1 \pm \theta)a|S|) > 1 - \delta$, 应确保

$$1 - 2\exp(-2|S|(a\theta)^2) > 1 - \delta$$

即

$$|S| > \ln_{2(a\theta)^2} \frac{\delta}{2} = \frac{\ln\delta - \ln 2}{\ln 2 + 2\ln a\theta}$$

2.

给定高维空间 R^d 的一个点集 P, $|P| = n$, 如果我们进行 k-means 聚类,(1). 能否通过 J-L 变换来降低复杂度? (2).能否通过 PCA 来降低复杂度? 请分别给出详细的理论分析. (比如, 如果我们有一个 λ 倍近似比的 k-means 聚类算法 A, 其复杂度为 T(n,d), 那么我们用 A 对降维之后的数据运算, 并将得到的结果返回原有空间, 得到近似比是多少? 整个过程复杂度是多少?)

(1).

能否通过 **J-L** 变换来降低复杂度?

(2).

能否通过 **PCA** 来降低复杂度?