大数据算法作业一：2021 年 5 月 3 日下午上课交纸质版给助教
（之前请将作业拍成照片或扫描，发到助教邮箱）

1. 假设δ，$\theta \in (0, 1)$，给定集合 U 以及它的一个子集 V, |V|/|U|=a,（1）如果我们从 U 均匀采集样本 S, 并确保 S∩ $V \neq \phi$的概率大于 1-δ，|S|至少要多大？（2）如果我们要确保|$S \cap V$| ∈ （$1 \pm \theta$）$a|S|$的概率大于 1-δ，|S|至少要多大？（10+15 分）
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 JL-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，得到的近似比是多少？整个过程的复杂度是多少？）（10+15 分）
3. 请构造一个反例，证明 Gonzalez 算法不能代替 k-means++算法用于 k-means 聚类的初始点选择。（25 分）
4. 请计算平面上（1）矩形,（2）半圆形，以及（3）凸多边形的 VC dimension 。（7+7+11 分）