大数据算法第一次作业

2025 年 3 月 20 日

Problem 1. 假设 $Z_i, \ldots, Z_n$ 为 n 个独立同分布的随机变量，并且满足 $\mathbb{E}[Z_i]=0, Var(Z_i)<\infty$. 定义均值为 $\bar{Z}=\frac{1}{n}\sum_{i=1}^{n} Z_i$. 证明对于任意 $t>0$ 有:

$$\mathbb{P}(|\bar{Z}| \geq t) \to 0$$

Problem 2. 假设 $Z_i, \ldots, Z_n$ 为 n 个独立的有界随机变量，其中 $Z_i \in [a,b]$ 且 $-\infty < a \leq b < \infty$。证明

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(Z_i - \mathbb{E}[Z_i]) \geq t\right) \leq \exp\left(-\frac{2nt^2}{(b-a)^2}\right),$$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(Z_i - \mathbb{E}[Z_i]) \leq -t\right) \leq \exp\left(-\frac{2nt^2}{(b-a)^2}\right)$$

对于任意 $t \geq 0$ 成立.

Problem 3. 假设我们有一个二维数据集 $\mathrm{X} = (1,1),(1,2),(2,1),(2,2),(10,10),(10,11),(11,10),(11,11)$，我们希望将其分为 $\mathrm{k}=2$ 类

(1) 如果初始类中心选择为 $\mathrm{c}_1=(1,1), \mathrm{c}_2=(10,10)$，请执行 k-means 算法，给出迭代过程和最终的类中心

(2) 如果初始类中心选择为 $\mathrm{c}_1=(1,1), \mathrm{c}_2=(2,2)$，请执行 k-means 算法，给出迭代过程和最终的类中心

比较两种初始类中心选择，可以感受到不同类中心选择对算法执行的影响

(3) 现在使用 k-means++ 算法，并希望将数据集分为 $\mathrm{k}=3$ 类，假设第一个类中心 c_1 已经被选择为 $(1,1)$，请计算每个点被选为第二个类中心 c_2 的概率

(4) 如果第二个类中心 c_2 被选择为 $(10,10)$, 请计算每个点被选为第三个类中心 c_3 的概率

Problem 4. 回忆上课讲过的 Bicriteria approximation for k-means 算法，我们定义算法运行到第 i 步时类中心集合为 $\mathrm{S_i}$，初始化的集合为 $\mathrm{S}_0=\emptyset$。令 C_{opt} 为 k 个类的最优解，当算法运行到第 i 步时，我们将最优解导出的类 $\mathrm{A}_1, ..., \mathrm{A_k}$, 分为两种集合:

$$\mathrm{G}ood_i = \{\mathrm{A_j} | \phi_{\mathrm{A_j}}(\mathrm{S_{i-1}}) \leq 10\phi_{\mathrm{A_j}}(\mathrm{C}_{opt})\}$$

$$\mathrm{B}ad_i = \{\mathrm{A}_1, ..., \mathrm{A_k}\} \backslash \mathrm{G}ood_i$$

现在假设算法运行到第 i 步时，有 $\phi_{\mathrm{X}}(\mathrm{S_{i-1}}) \geq 10/\epsilon \phi_{\mathrm{X}}(\mathrm{C}_{opt})$, 请证明算法在该步选择的点 $\mathrm{c_i} \in \mathrm{B}ad_i$ 的概率至少为 $1-\epsilon$, 其中 $\epsilon \in (0,1)$

1 Chromatic number(染色数) of Erdos-Renyi Graph

Definition 1.1. Erdos-Renyi 随机图 $G(n,p)$ 是一个 n 点图，每条边以概率 p 独立生成。A 表示所得图的染色数，定义为染色全部顶点时，使相邻顶点颜色不同所需最少的色数。定义随机变量 $X_1,\cdots,X_{\binom{n}{2}}$，如顶点对 t 是边, 则 $X_t=1$，否则为 0. 可以证明随机变量 $Z_t=\mathbb{E}[A|X_1,...,X_t]$ 构成鞅，称为 *edge exposure martingale*。

对应的，可以定义随机变量 $Y_1,\cdots,Y_n$，使得 Y_t 代表顶点 t 的邻点集，可知有 2^{n-1} 种取值。同样可以定义 *vertex exposure martingale* 为 $Z_1,\ldots,Z_n$，$Z_t=\mathbb{E}[A|Y_1,...,Y_t]$.

Problem 5. 证明 vertex exposure martingale 是鞅.

Problem 6. 证明

$$\Pr(|A-\mathbb{E}[A]|\geq c\sqrt{n})\leq 2e^{-\frac{c^2}{2}}$$

.

Problem ex1. (本题是开放性的思考题, 不计入作业分数)K-means 算法的结果非常依赖于初始类中心的选择。虽然 K-means++ 算法通过改进初始化方式缓解了这一问题，但它仍然不能完全避免局部最优解。能否设计一种新的初始化方法，进一步减少 K-means 算法对初始类中心选择的敏感性?