中国科学技术大学
计算机科学与技术学院

“大数据算法” 作业 1
2023 年春

注: 本次作业**不**计入成绩，无需提交。

下列各题中, 我们用 [**BHK**] 指代由 Blum, Hopcroft 和 Kannan 编著的《数据科学基础》一书。

习题 1

([BHK] 中的习题 3.12) 令 $\sum_{i=1}^{r}\sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$ 为一个秩为 r 的矩阵 A 的奇异值分解 (SVD)。对于某个 $k < r$, $A_k = \sum_{i=1}^{k}\sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$ 是矩阵 A 的一个秩为 k 的近似。用奇异值 $\{\sigma_i, 1 \le i \le r\}$ 表达以下几个量。

(a) $\|A_k\|_F^2$

(b) $\|A_k\|_2^2$

(c) $\|A - A_k\|_F^2$

(d) $\|A - A_k\|_2^2$

习题 2

([BHK] 中的习题 3.18) 如果对于所有的 $\boldsymbol{x}$, 都有 $\boldsymbol{x}^\top A\boldsymbol{x} \ge 0$, 则称矩阵 A 是半正定矩阵。

(a) 令 A 为一个实矩阵。证明 $B = AA^\top$ 是半正定矩阵。

(b) 令 A 为一个图的邻接矩阵。A 的拉普拉斯矩阵为 $L = D - A$, 其中 D 是一个对角阵, 其对角线上的元素为 A 中对应行的元素之和。通过证明 $L = B^\top B$, 来证明 L 是半正定矩阵, 其中 B 是一个 $m \times n$ 的矩阵, B 的每一行对应图中的一条边, 每一列对应图中的一个顶点, 我们定义

$$b_{ei} = \begin{cases} -1 & \text{如果 } i \text{ 是 } e \text{ 中索引较小的端点} \\ 1 & \text{如果 } i \text{ 是 } e \text{ 中索引较大的端点} \\ 0 & \text{如果 } i \text{ 不是 } e \text{ 的一个端点} \end{cases}$$

习题 3

([BHK] 中的习题 3.22)

(a) 对于任意矩阵 A, 证明 $\sigma_k \le \frac{\|A\|_F}{\sqrt{k}}$。

(b) 证明存在一个秩最多为 k 的矩阵 B, 使得 $\|A - B\|_2 \le \frac{\|A\|_F}{\sqrt{k}}$。

(c) 能否将 (b) 中不等式左侧的 2 范数换成 Frobenius 范数?

. . . .

([BHK] 中的习题 3.23) 假设给定一个 $n \times d$ 的矩阵 A, 并且你可以对 A 进行预处理。然后给定一系列 d 维向量 $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m$。对于其中的每个向量, 你需要近似地找到向量 $A\boldsymbol{x}_j$。也就是说, 你需要找到一个向量 $\boldsymbol{y}_j$, 满

足 $\|\boldsymbol{y}_j - A\boldsymbol{x}_j\| \leq \varepsilon\|A\|_F\|\boldsymbol{x}_j\|$, 其中 $\varepsilon > 0$ 是一个给定的误差界。设计一个算法, 使得对于每个 $\boldsymbol{x}_j$, 以 $O\left(\frac{d+n}{\varepsilon^2}\right)$ 的时间复杂度实现上述目标, 不考虑预处理时间。

提示: 使用 [BHK] 中的习题 3.22。

习题 4

令 $A \in \mathbb{R}^{n\times d}$ 为一个数据矩阵, 其奇异值分解 (SVD) 为 $A = UDV^\top = \sum_{i=1}^r \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$, 其中 $r \leq d$。假设对于某个 $\varepsilon > 0$, 有 $\sigma_2 < (1-\varepsilon)\sigma_1$。令 $\boldsymbol{x}$ 为一个向量, 使得 $\boldsymbol{x}^\top \boldsymbol{v}_1 \geq \frac{1}{2}$。对于每个整数 $k \geq 1$, 定义向量 $\boldsymbol{b}_k = (A^\top A)^k \boldsymbol{x}$。
找到尽可能最小的 k, 使得

$$\left|\boldsymbol{b}_k^\top \cdot \boldsymbol{v}_1\right| \geq \left(1 - \varepsilon^{10}\right)\|\boldsymbol{b}_k\|,$$

并解释原因。
