大数据算法第二次作业

2025 年 4 月 10 日

Problem 1. 求矩阵 $A = \begin{bmatrix} 3 & 0 \\ 4 & 5 \end{bmatrix}$ 的奇异值分解 U, Σ, V 矩阵。

Problem 2. 设 $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_p)^T$ 是 $\mathbf{X}$ 的主成分向量，$Var(X) = \Sigma = PAP^T$，$\mathbf{Y} = P^T\mathbf{X}$，其中 P 为主成分分析的载荷矩阵。证明：

- 原始变量 X_k 与主成分 Y_j 的相关系数为

$$\rho_{kj} = \rho(X_k, Y_j) = \frac{\sqrt{\lambda_j}}{\sqrt{\sigma_{kk}}} p_{kj},$$

  其中 $p_j = (p_1, p_2, \ldots, p_p)^T$ 是 P 的第 j 列，p_{kj} 是 P 的第 (k, j) 元素。

- 原始变量 X_k 与主成分 Y_j 的相关系数是

$$\sum_{j=1}^{p} \rho_{kj}^2 = 1, j = 1, 2, \ldots, p.$$

Problem 3. 设 $\mathcal{X} \subset \{0,1\}^d$ 为二进制向量空间，赋予汉明距离度量。定义哈希函数族

$$H = \{h_i \mid h_i(u) = u_i,\ 1 \le i \le d\}$$

则该函数族是 $(r, (1+\epsilon)r, 1 - r/d, 1 - (1+\epsilon)r/d)$-局部敏感哈希族。

Problem 4. 在度量空间 (T, d) 中，子集 K 被称为 $\epsilon-$ 分离的，当且仅当对任意不同的 $p, q \in K$，有 $d(p, q) > \epsilon$. 对于空间 T，记最大的 $\epsilon-$ 分离子集的势（大小）为 $\mathcal{N}(T, \epsilon)$，称作 T 的覆盖数。

(1) 证明：$\mathcal{N}(T, \epsilon) \le \frac{|B(\frac{\epsilon}{2}) + T|}{|B(\frac{\epsilon}{2})|}$. 其中，“+” 作用于两个集合，$A + B = \{a + b | a \in A, b \in B\}$.

（提示：考虑分离子集的每个元素，以它们为中心半径为 ϵ 的球。）

(2)（JL 变换的最优性）证明：对于任意给定的 $\epsilon \in (0, 1)$，存在 $P \subset \mathbb{R}^d, |\mathrm{P}| = n \in \mathbb{N}_+$，如果存在一个映射 $f : \mathbb{R}^d \to \mathbb{R}^k$，使得对于任意两个向量 $x, y \in P$，有

$$(1 - \epsilon)||x - y||^2 \le ||f(x) - f(y)||^2 \le (1 + \epsilon)||x - y||^2, \tag{1}$$

则必有 $k = \Omega(\log d)$.

（提示：考虑集合 $P = \{0, e_1, \cdots, e_d\}$，$e_k$ 为第 k 个标准正交基。则 $f(P)$ 是否是 $B(1)$ 的某个分离子集?）

Problem 5. 给定 N 个向量 $v_1, v_2, \ldots, v_N \in \mathbb{R}^d$，构造 jl 随机投影矩阵 $B \in \mathbb{R}^{k\times d}$，其每个元素独立采样自高斯分布 $\mathcal{N}(0, 1/k)$。令投影维度 $k > \frac{24\log N}{\epsilon^2}$。已知引理：
对于任意独立重复采样自 $\mathcal{N}(0, \frac{1}{n})$ 的向量 $w \in \mathbb{R}^n$ 和常数 $\varepsilon \in (0, 1)$ 有：

$$P\left(\left|\|w\|^2 - 1\right| \geq \varepsilon\right) \leq 2\exp\left(-\frac{\varepsilon^2 n}{8}\right).$$

要求证明至少有 $\frac{N-1}{N}$ 的概率，对于任意 $i \neq j$ 和常数 $\varepsilon \in (0, 1)$，

$$(1-\varepsilon)\|v_i - v_j\|^2 \leq \|Bv_i - Bv_j\|^2 \leq (1+\varepsilon)\|v_i - v_j\|^2.$$

1 思考题

1. 我们讲过的 JL 变换都是线性变换。如何得出非线性变换，从而更好地配合数据所处的流形？（比如为了加速 SVM 再生核的计算）

2. 如何将随机化的次线性算法改为确定性算法？