大数据算法第二次作业

2025 年 4 月 15 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。

证明. 首先计算 A^TA 和 AA^T:

$$A^TA = \begin{bmatrix} 25 & 20 \\ 20 & 25 \end{bmatrix}$$

$$AA^T = \begin{bmatrix} 9 & 12 \\ 12 & 41 \end{bmatrix}$$

这两个矩阵有相同的特征值 $\sigma_1^2 = 45$ 和 $\sigma_2^2 = 5$。开平方后得到 $\sigma_1 = \sqrt{45}$ 和 $\sigma_2 = \sqrt{5}$。
求 A^TA 的特征向量（特征值为 45 和 5）:

$$\begin{bmatrix} 25 & 20 \\ 20 & 25 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = 45 \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} 25 & 20 \\ 20 & 25 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \end{bmatrix} = 5 \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

得右奇异向量 $v_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$，$v_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 \\ 1 \end{bmatrix}$，$u_i$ 为左奇异向量。
接下来计算 Av_1 和 Av_2，它们将分别等于 $\sigma_1 u_1 = \sqrt{45}u_1$ 和 $\sigma_2 u_2 = \sqrt{5}u_2$:

$$Av_1 = \frac{3}{\sqrt{2}} \begin{bmatrix} 1 \\ 3 \end{bmatrix} = \sqrt{45} \frac{1}{\sqrt{10}} \begin{bmatrix} 1 \\ 3 \end{bmatrix} = \sigma_1 u_1$$

$$Av_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} -3 \\ 1 \end{bmatrix} = \sqrt{5} \frac{1}{\sqrt{10}} \begin{bmatrix} -3 \\ 1 \end{bmatrix} = \sigma_2 u_2$$

通过除以 $\sqrt{10}$，u_1 和 u_2 被单位化。因此，$\sigma_1 = \sqrt{45}$ 和 $\sigma_2 = \sqrt{5}$ 符合预期。奇异值分解为 $A = U\Sigma V^T$:

$$U = \frac{1}{\sqrt{10}} \begin{bmatrix} 1 & -3 \\ 3 & 1 \end{bmatrix} \quad \Sigma = \begin{bmatrix} \sqrt{45} & 0 \\ 0 & \sqrt{5} \end{bmatrix} \quad V = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$$

□

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

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

证明. (1) 记 e_k 为单位矩阵 I_p 的第 k 列，即仅第 k 个元素为 1，其他元素都为 0 的 p 维向量。则 $X_k = e_k^T X, Y_j = p_j^T X$。于是

$$\begin{aligned} Cov(X_k, Y_j) &= Cov(e_k^T X, p_j^T X) = e_k^T \Sigma p_j \\ &= e_k^T (\Sigma p_j) = e_k^T (\lambda_j p_j) \quad (p_j是\lambda_j对应的特征向量) \\ &= \lambda_j (e_k^T p_j) = \lambda_j p_{kj}, \end{aligned}$$

$$\rho_{kj} = \rho(X_k, Y_j) = \frac{Cov(X_k, Y_j)}{\sqrt{Var(X_k) \cdot Var(Y_j)}}$$

$$= \frac{\lambda_j p_{kj}}{\sqrt{\sigma_{kk}}\lambda_j} = \frac{\sqrt{\lambda_j}}{\sqrt{\sigma_{kk}}} p_{kj}.$$

(2) 注意到

$$\Sigma = PAP^T = \sum_{j=1}^{p} \lambda_j p_j p_j^T,$$

所以

$$\sigma_{kk} = \sum_{j=1}^{p} \lambda_j p_{kj}^2,$$

于是

$$\sum_{j=1}^{p} \rho_{kj}^2 = \frac{1}{\sigma_{kk}} \sum_{j=1}^{p} \lambda_j p_{kj}^2 = 1.$$

这样，ρ_{kj}^2 可以看成是在原始变量 X_k 的方差中，第 j 主成分 Y_j 能够解释的方差比例。 □

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。

证明. 考虑任意两个向量 $u, v \in \{0,1\}^d$ 满足 $\|u-v\|_1 \le r$。由汉明距离的定义可知，这两个向量最多有 r 个坐标不同。因此，对于随机选取的哈希函数 h_i，有

$$\mathbb{P}[h_i(u) \ne h_i(v)] = \frac{\|u-v\|_0}{d} \le \frac{r}{d}$$

其补事件概率为

$$\mathbb{P}[h_i(u) = h_i(v)] \ge 1 - \frac{r}{d}$$

即得 $p_1 = 1 - r/d$。同理可证，当 $\|u-v\|_1 \ge (1+\epsilon)r$ 时，碰撞概率 $p_2 = 1-(1+\epsilon)r/d$，证毕。 □

Problem 5. 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 $\frac{\epsilon}{2}$ 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?）

证明.（1）提示中的集合被包含于 $B(\frac{\epsilon}{2}) + T$ 中，比较两者大小即得。

（2）$f(P)$ 是 $B(1)$ 的 $\sqrt{1-\epsilon}-$ 分离子集。从而有

$$d+1 \le \mathcal{N}(B(1), \sqrt{1-\epsilon}) \le \frac{|B(\frac{3}{2})|}{|B(\frac{1}{4})|} = 2^{\Theta(k)}.$$

从而 $k = \Omega(\log d)$. □

Problem 6. 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：

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

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

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

证明. 直接运用引理，令 $w = B\frac{v_i - v_j}{||v_i - v_j||}$，得到一个概率界，然后使用 union bound 即可证明
首先，如果 $w \in \mathbb{R}^d$ 是一个单位向量，而 $B \in \mathbb{R}^{k\times d}$ 独立重复采样自 $\mathcal{N}(0, 1/k)$，那么 Bw 的每个分量都独立地服从 $\mathcal{N}(0, 1/k)$。根据定义，每个分量

$$(Bw)_i = \sum_j B_{i,j} w_j,$$

由于 $B_{i,j}$ 相互独立，所以 $(Bw)_i$ 相互独立，并且由于 $B_{i,j} \sim \mathcal{N}(0, 1/k)$，正态随机变量和的分布依然是正态分布，因此 $(Bw)_i$ 服从正态分布，其均值为

$$\sum_j w_j \times 0 = 0,$$

其方差为

$$\sum_j w_j^2 \times \frac{1}{k} = \frac{1}{k}.$$

所以，Bw 相当于从 $\mathcal{N}(0, 1/k)$ 独立重复采样出来的 k 维向量。现在代入 $w = \frac{v_i - v_j}{\|v_i - v_j\|}$，利用引理，得到

$$P\left(\left|\left\|\frac{B(v_i - v_j)}{\|v_i - v_j\|}\right\|^2 - 1\right| \geq \varepsilon\right) \leq 2\exp\left(-\frac{\varepsilon^2 k}{8}\right).$$

此结果对于任意 $i \neq j$ 都成立。遍历所有 $i \neq j$ 的组合，至少有一项不满足条件的概率不超过

$$P\left(\exists(i,j) : \left|\left\|\frac{A(v_i - v_j)}{\|v_i - v_j\|}\right\|^2 - 1\right| \geq \varepsilon\right) \leq 2\binom{N}{2}\exp\left(-\frac{\varepsilon^2 k}{8}\right).$$

反过来，对于任意 $i \neq j$，都成立

$$\left|\left\|\frac{A(v_i - v_j)}{\|v_i - v_j\|}\right\|^2 - 1\right| \leq \varepsilon$$

的概率不小于

$$1 - 2\binom{N}{2}\exp\left(-\frac{\varepsilon^2 k}{8}\right) = 1 - N(N-1)\exp\left(-\frac{\varepsilon^2 k}{8}\right).$$

代入 $k > \frac{24\log N}{\varepsilon^2}$，可以得到

$$1 - N(N-1)\exp\left(-\frac{\varepsilon^2 k}{8}\right) \geq 1 - N(N-1)N^{-3} \geq 1 - N^{-1}.$$

□

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？