大数据算法第四次作业

2025 年 6 月 15 日

Problem 1. 证明: 局部线性嵌入中, 权重矩阵 W 满足

$$W_i = \frac{C^{-1}1}{1^T C^{-1}1},$$

其中 $C_{jk} = (x_i - x_j)^T (x_i - x_k)$ 为局部协方差矩阵.

Problem 2. 设将输入数据集 X 通过最优的 k-均值聚类划分为 $X_1 \cup \cdots \cup X_k = X$，其中每个簇 X_i 的质心记为 c_i。则有 $\Delta_k^2(X) = \sum_{i=1}^k \sum_{x \in X_i} \|x - c_i\|^2 = \sum_{i=1}^k \Delta_1^2(X_i)$。记 $n_i = |X_i|$，$n = |X|$，并定义 $r_i^2 = \frac{\Delta_1^2(X_i)}{n_i}$。我们假设聚类误差满足如下 ε-separated 条件：$\Delta_k^2(X) \le \epsilon^2 \Delta_{k-1}^2(X)$。请证明，对于每个簇 i，均有如下不等式成立：

$$r_i^2 \le \frac{\epsilon^2}{1-\epsilon^2} \cdot \min_{j \ne i} \|c_i - c_j\|^2$$

Problem 3. 在课堂上我们学习了针对 k-means 问题（$k = 2$ 时）的 Lloyd-Type 方法（Beyond Worst-Case Analysis, BWCA）。请写出当 k 为一般情形时，该算法中的采样步骤。

Problem 4. 设 $X \in \mathbb{R}^{n \times d}$ 为中心化数据矩阵（即 $\sum_{i=1}^n x_i = 0$），其协方差矩阵为 $C = \frac{1}{n} X^T X$。经典多维尺度分析（MDS）以欧氏距离矩阵 D（其中 $D_{ij} = \|x_i - x_j\|_2$）为输入，输出低维嵌入 $Y \in \mathbb{R}^{n \times k}$；主成分分析（PCA）以 X 为输入，输出降维数据 $Z \in \mathbb{R}^{n \times k}$。证明：经典 MDS 与 PCA 在数学上等价，即满足 $Y = Z$（忽略符号和排列顺序的差异）。

Problem 5. 假设有以下 4 个点在二维空间中的坐标：

$$X = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{bmatrix}$$

1. 计算这些点之间的欧氏距离矩阵 D。
2. 使用 $k = 2$ 近邻构建邻域图，并计算最短路径距离矩阵 $\hat{D}$（假设邻域内的边权重为欧氏距离）。
3. 对 $\hat{D}$ 应用经典 MDS 算法，计算二维嵌入表示 Y（只需写出双中心化矩阵 B 的表达式，无需完全计算）。

Problem 思考题. 1. 我们在局部线性嵌入中讨论了降维的情形. 如果让 $k > n$, 这时的求解有什么困难? 如何解决?