大数据算法第四次作业

2025 年 6 月 21 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.

证明. 对于 LLE 算法，我们首先要确定邻域大小的选择，即我们需要多少个邻域样本来线性表示某个样本。假设这个值为 k。我们可以通过和 KNN 一样的思想通过距离度量比如欧式距离来选择某样本的 k 个最近邻。

在寻找到某个样本的 x_i 的 k 个最近邻之后我们就需要找到找到 x_i 和这 k 个最近邻之间的线性关系，也就是要找到线性关系的权重系数。找线性关系，这显然是一个回归问题。假设我们有 m 个 n 维样本 $\{x_1, x_2, ..., x_m\}$, 我们可以用均方差作为回归问题的损失函数：即：

$$J(w) = \sum_{i=1}^{m} ||x_i - \sum_{j\in Q(i)} w_{ij}x_j||_2^2$$

其中，$Q(i)$ 表示 i 的 k 个近邻样本集合。一般我们也会对权重系数 w_{ij} 做归一化的限制，即权重系数需要满足

$$\sum_{j\in Q(i)} w_{ij} = 1$$

对于不在样本 x_i 邻域内的样本 x_j，我们令对应的 $w_{ij} = 0$，这样可以把 w 扩展到整个数据集的维度。

也就是我们需要通过上面两个式子求出我们的权重系数。一般我们可以通过矩阵和拉格朗日子乘法来求解这个最优化问题。

对于第一个式子，我们先将其矩阵化：

$$\begin{aligned}
J(W) &= \sum_{i=1}^{m} ||x_i - \sum_{j\in Q(i)} w_{ij}x_j||_2^2 && (1)\\
&= \sum_{i=1}^{m} ||\sum_{j\in Q(i)} w_{ij}x_i - \sum_{j\in Q(i)} w_{ij}x_j||_2^2 && (2)\\
&= \sum_{i=1}^{m} ||\sum_{j\in Q(i)} w_{ij}(x_i - x_j)||_2^2 && (3)\\
&= \sum_{i=1}^{m} W_i^T(x_i - x_j)(x_i - x_j)^T W_i && (4)
\end{aligned}$$

其中 $W_i=(w_{i1},w_{i2},...w_{ik})^T$。

我们令矩阵 $C=(x_i-x_j)(x_i-x_j)^T, j\in Q(i)$, 则第一个式子进一步简化为 $J(W)=\sum\limits_{i=1}^{k}W_i^TCW_i$. 对于第二个式子，我们可以矩阵化为：

$$\sum_{j\in Q(i)} w_{ij}=W_i^T1_k=1$$

其中 1_k 为 k 维全 1 向量。

现在我们将矩阵化的两个式子用拉格朗日子乘法合为一个优化目标：

$$L(W)=\sum_{i=1}^{k}W_i^TCW_i+\lambda(W_i^T1_k-1)$$

对 W 求导并令其值为 0，我们得到

$$2CW_i+\lambda 1_k=0$$

即我们的

$$W_i=\lambda' C^{-1}1_k$$

其中 $\lambda'=-\frac{1}{2}\lambda$ 为一个常数。利用 $W_i^T1_k=1$, 对 W_i 归一化，那么最终我们的权重系数 W_i 为：

$$W_i=\frac{C^{-1}1_k}{1_k^TC^{-1}1_k}$$

□

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：

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

证明. 设 X 被划分为 k 个簇 $X_1,\ldots,X_k$，每个簇大小为 n_i，中心为 c_i，有：

$$r_i^2=\frac{1}{n_i}\sum_{x\in X_i}\|x-c_i\|^2,\quad \Delta_k^2(X)=\sum_{i=1}^{k}n_ir_i^2.$$

根据 ε-separated 假设，有：

$$\Delta_k^2(X)\le\varepsilon^2\Delta_{k-1}^2(X).$$

合并任意两个簇 X_i 和 X_j：

$$\Delta_1^2(X_i\cup X_j)=\Delta_1^2(X_i)+\Delta_1^2(X_j)+\frac{n_in_j}{n_i+n_j}\|c_i-c_j\|^2.$$

因此，

$$\Delta_{k-1}^2(X)\le\Delta_k^2(X)+\frac{n_in_j}{n_i+n_j}\|c_i-c_j\|^2.$$

结合 ε-separated 不等式，

$$\Delta_k^2(X) \le \varepsilon^2 \left(\Delta_k^2(X) + \frac{n_i n_j}{n_i + n_j} \|c_i - c_j\|^2 \right).$$

整理得：

$$(1-\varepsilon^2)\Delta_k^2(X) \le \varepsilon^2 \cdot \frac{n_i n_j}{n_i + n_j} \|c_i - c_j\|^2.$$

注意 $\Delta_k^2(X) \ge n_i r_i^2$，代入可得：

$$(1-\varepsilon^2) n_i r_i^2 \le \varepsilon^2 \cdot \frac{n_i n_j}{n_i + n_j} \|c_i - c_j\|^2,$$

两边除以 n_i，得：

$$r_i^2 \le \frac{\varepsilon^2}{1-\varepsilon^2} \cdot \frac{n_j}{n_i + n_j} \|c_i - c_j\|^2 \le \frac{\varepsilon^2}{1-\varepsilon^2} \|c_i - c_j\|^2.$$

对所有 $j \ne i$ 取最小，即得证。 □

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?