Lecture 8: JL 变换的应用

2024.3.26

Lecturer: 丁虎 *Scribe:* 黄震

本章介绍 JL 变换的一些应用场景。

1 JL 变换结合 k-means

回顾 k-means 问题，其输入是 $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$，目标为寻找最优的 k 个类中心 $C = \{c_1, \ldots, c_k\} \subset \mathbb{R}^d$，将每个 x_i 归类于其最近的中心点 c_j，使得所有数据点到类中心的距离平方和最小，即

$$\text{cost}(x, C) = \min_{c \in C} ||x - c||^2$$
$$\text{cost}(X, C) = \sum_{x \in X} \text{cost}(x, C)$$
$$C^* = \arg \min_{|C|=k, C \subset \mathbb{R}^d} \text{cost}(X, C)$$

k-means 问题众多经典的求解方法，如 Lloyd 算法，k-means++ 算法等，均需要计算点对之间的距离，该过程和维度 d 线性相关，直接导致了总体时间复杂度中和 d 的线性依赖关系。当处理高维数据时，一个自然的想法是先对数据进行降维再来求解后续的优化问题。对于 k-means 问题，我们可以证明如下结论：

Theorem 1.1. *给定 k-means 问题的输入 $X \subset \mathbb{R}^d$，利用 JL 变换 $f : \mathbb{R}^d \to \mathbb{R}^m$ 将 X 降维成 X'，其中 $m = \Theta(\frac{\log n}{\epsilon^2})$。考虑在 X' 上的任意一个 λ-近似比的聚类结果 $\{C_1, \ldots, C_k\}$，令 $C_i^{-1} = \{f^{-1}(x) : x \in C_i\}$，那么 $\{C_1^{-1}, \ldots, C_k^{-1}\}$ 是 X 的 $\frac{1+\epsilon}{1-\epsilon}\lambda$-近似比的聚类结果。*

Remark 1.2. 对于一个最小化的优化问题 $\mathcal{P}$,对应的最优解为 X_{opt}。如果某个解满足 $\text{cost}(X) \leq \lambda \cdot \text{cost}(X_{\text{opt}})$，那么我们说 X 是问题 $\mathcal{P}$ 的一个 λ-近似比的解。

定理 1.1 为我们利用 JL 变换先做数据降维再求解 k-means 的想法提供了理论保证，它说明在降维后的数据上求解得到的聚类结果，逆变换为原空间后，同样是不错的选择。为了证明该定理，我们先介绍一个基础的结论

Claim 1.3. $\forall\, Q \subset \mathbb{R}^d$，令 $\mu(Q) = \frac{1}{|Q|}\sum_{q\in Q} q$，有

$$\sum_{q\in Q} ||q-\mu(Q)||^2 = \frac{1}{2|Q|}\sum_{q_i}\sum_{q_j} ||q_i - q_j||^2$$

Proof.

$$\begin{aligned}
&\sum_{q_j} ||q_i - q_j||^2 = \sum_{q_j} ||q_j - \mu(Q)||^2 + |Q| \cdot ||\mu(Q) - q_i||^2 \\
\Rightarrow &\sum_{q_i}\sum_{q_j} ||q_i - q_j||^2 = |Q|\sum_{q_j} ||q_j - \mu(Q)||^2 + |Q|\sum_{q_i} ||\mu(Q) - q_i||^2 \\
\Rightarrow &\sum_{q\in Q} ||q-\mu(Q)||^2 = \frac{1}{2|Q|}\sum_{q_i}\sum_{q_j} ||q_i - q_j||^2
\end{aligned}$$

□

上述结论说明点集到其重心的距离平方和，与点集内部所有点对之间的距离平方和有关。我们接下来证明定理 1.1:

Proof. (定理 1.1) 对任意点集 Q，令 $\Gamma(Q) = \frac{1}{2|Q|}\sum_{q_i}\sum_{q_j} ||q_i - q_j||^2$。记 X' 中归类于 c_i 的点的集合为 C_i，X 中归类于 $f^{-1}(c_i)$ 的点的集合为 C_i^{-1}。由 Claim 1.3 可知 $\sum_{x\in C_i^{-1}} ||x - \mu(C_i^{-1})||^2 = \Gamma(C_i^{-1})$。因此，$\mathcal{C}^{-1} = \{C_i^{-1}, \ldots, C_k^{-1}\}$ 对应的损失

$$\mathrm{cost}(\mathcal{C}^{-1}) = \sum_{i=1}^{k} \Gamma(C_i^{-1})$$

同时，由于 X' 是由 X 经过 JL 变换而来，因此

$$\begin{aligned}
\Gamma(C_i^{-1}) &\in (\frac{1}{1+\epsilon}, \frac{1}{1-\epsilon}) \cdot \Gamma(C_i) \\
\mathrm{cost}(\mathcal{C}^{-1}) &\in (\frac{1}{1+\epsilon}, \frac{1}{1-\epsilon}) \cdot \sum_{i=1}^{k} \Gamma(C_i)
\end{aligned}$$

假设 X 上的最优聚类划分为 $\mathcal{U} = \{U_1, \ldots, U_k\}$，令 $U_i' = \{f(x) : x \in U_i\}$，则

$$\sum_{i=1}^{k} \Gamma(U_i') \in (1 \pm \epsilon) \sum_{i=1}^{k} \Gamma(U_i)$$

由于 $\mathcal{C} = \{C_1, \ldots, C_k\}$ 是 X' 上的 λ-近似比的聚类结果，所以

$$\mathrm{cost}(\mathcal{C}^{-1}) \le \frac{1}{1-\epsilon}\sum_{i=1}^{k}\Gamma(C_i) \le \frac{\lambda}{1-\epsilon}\sum_{i=1}^{k}\Gamma(U_i') \le \frac{1+\epsilon}{1-\epsilon}\lambda\sum_{i=1}^{k}\Gamma(U_i)$$

□

在上述证明中，JL 变换后的维度为 $\Theta(\frac{\log n}{\epsilon^2})$，所以我们可以以很大概率保证所有点对的距离都变化不大，从而完成整体推导。但实际上，k-means 的需求是弱于这个前提的，它仅需要每个类内部的点对平方和变化不大。从该角度来看，我们指定投影后的维度为 $\Theta(\frac{\log n}{\epsilon^2})$ 其实是稍微有点强了，这个维度直觉上可以更低。

2 降维视角下的 k-means

Definition 2.1 (聚类指示矩阵)**.** 给定数据输入 $X \in \mathbb{R}^{n\times d}$，考虑其一个聚类划分 $\mathcal{C}$，定义聚类指示矩阵（cluster indicator matrix）为 $I_{\mathcal{C}} \in \mathbb{R}^{n\times d}$，满足

$$I_{\mathcal{C}}(i,j) = \begin{cases} \frac{1}{\sqrt{|C_j|}} & x_i \in C_j \\ 0 & x_i \notin C_j \end{cases}$$

Example 2.2. *考虑* $X = \{x_1, x_2, x_3, x_4, x_5, x_6\}$*，它们分别属于聚类 2，1，1，3，2，1，那么*

$$I_{\mathcal{C}} = \begin{pmatrix} 0 & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{3}} & 0 & 0 \\ \frac{1}{\sqrt{3}} & 0 & 0 \\ 0 & 0 & 1 \\ 0 & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{3}} & 0 & 0 \end{pmatrix}$$

根据 $I_{\mathcal{C}}$ 的定义，我们容易发现它的 k 个列向量是单位正交的。更进一步，我们有如下关系：

Claim 2.3. *令* $\mathrm{cost}(\mathcal{C})$ *表示聚类划分* $\mathcal{C}$ *对应的 k-means 损失，那么*

$$||X - I_{\mathcal{C}} I_{\mathcal{C}}^T X||_F^2 = \mathrm{cost}(\mathcal{C})$$

Proof.

$$I_{\mathcal{C}}^T X = \begin{pmatrix} \vdots \\ \frac{1}{\sqrt{|C_i|}} \sum_{x \in C_i} x \\ \vdots \end{pmatrix}_{k \times d} = \begin{pmatrix} \vdots \\ \sqrt{|C_i|}\mu(C_i) \\ \vdots \end{pmatrix}_{k \times d}$$

考虑 $x \in C_i$，令 $c(x) = \mu(C_i)$，则

$$I_{\mathcal{C}} I_{\mathcal{C}}^T X = \begin{pmatrix} c(x_1) \\ \vdots \\ c(x_n) \end{pmatrix}_{n \times d}$$

$$||X - I_{\mathcal{C}} I_{\mathcal{C}}^T X||_F^2 = \sum_{i=1}^{n} ||x_i - c(x_i)||^2 = \text{cost}(\mathcal{C})$$

□

本质上，$I_{\mathcal{C}} I_{\mathcal{C}}^T$ 可以视为对 X 进行一个投影操作，而 k-means 问题就相当于寻找最能维持原结构的投影。

Theorem 2.4. 给定 $X \subset \mathbb{R}^{n \times d}$，令 $R \in \mathbb{R}^{\kappa \times d}$ 为一个 $\mathbb{R}^d \to \mathbb{R}^\kappa$ 的 *JL* 变换矩阵，其中 $\kappa = \Theta(\frac{\log k}{\epsilon^2})$。$X$ 经过 R 变换后为 $\tilde{X}$。假设 P^* 是 X 的最优 k-*means* 投影，$\tilde{P}$ 是 $\tilde{X}$ 的 λ-近似比的 k-*means* 投影，那么有

$$||X - \tilde{P}X||_F^2 \leq (9 + \Theta(\epsilon))\lambda ||X - P^* X||_F^2$$

Proof. 令 $B = P^* X, \bar{B} = (I - P^*)X$，则 $X = B + \bar{B}$。于是有，

$$\begin{aligned} ||X - \tilde{P}X||_F &= ||B + \bar{B} - \tilde{P}(B + \bar{B})||_F \\ &\leq ||B - \tilde{P}B||_F + ||\bar{B} - \tilde{P}\bar{B}||_F \quad (1) \\ &\leq ||B - \tilde{P}B||_F + ||\bar{B}||_F \quad (2) \end{aligned}$$

式 1 是由 Schwarz's 不等式所得。式 2 是因为 $I - \tilde{P}$ 仍然是一个投影矩阵，投影后的 Frobenius 范数不超过原始值。

由于 B 和 $\tilde{P}B$ 中实际只有 k 个向量，因此根据 JL 变换的性质，我们以高概率保证下式成立：

$$||B - \tilde{P}B||_F^2 \leq (1 + \epsilon) \cdot ||(B - \tilde{P}B)R^T||_F^2$$

因此有

$$||X - \tilde{P}X||_F \le \sqrt{1+\epsilon} \cdot ||BR^T - \tilde{P}BR^T||_F + ||\bar{B}||_F$$

由于 $(B + \bar{B})R^T = \tilde{X}$，于是

$$\begin{aligned}
||X - \tilde{P}X||_F &\le \sqrt{1+\epsilon} \cdot ||(\tilde{X} - \bar{B}R^T) - \tilde{P}(\tilde{X} - \bar{B}R^T)||_F + ||\bar{B}||_F \\
&\le \sqrt{1+\epsilon} \cdot ||\tilde{X} - \tilde{P}\tilde{X}||_F + \sqrt{1+\epsilon} \cdot ||(I - \tilde{P})\bar{B}R^T||_F + ||\bar{B}||_F \\
&\le \sqrt{1+\epsilon} \cdot ||\tilde{X} - \tilde{P}\tilde{X}||_F + \sqrt{1+\epsilon} \cdot ||\bar{B}R^T||_F + ||\bar{B}||_F \\
&\le \sqrt{(1+\epsilon)\lambda}||\tilde{X} - P^*\tilde{X}|| + (1+\epsilon)||\bar{B}||_F + ||\bar{B}||_F \\
&\le (1+\epsilon)\sqrt{\lambda}||X - P^*X||_F + (2+\epsilon)||\bar{B}||_F \\
&\le (3 + \Theta(\epsilon))\sqrt{\lambda}||X - P^*X||_F
\end{aligned}$$

因此

$$||X - \tilde{P}X||_F^2 \le (9 + \Theta(\epsilon))\lambda||X - P^*X||_F^2$$

□

定理 2.4 告诉我们当投影维度为 $\Theta(\frac{\log k}{\epsilon^2})$ 时仍然可以保持 k-means 解的损失，仅需要引入额外 $9 + \Theta(\epsilon)$ 的乘性误差。在 2019 年，该结论被进一步改进，维度仍然是 $\Theta(\frac{\log k}{\epsilon^2})$，额外的乘性误差改进为 $1 + \Theta(\epsilon)$（参考 [1]）。

References

[1] K. Makarychev, Y. Makarychev, and I. Razenshteyn. Performance of johnson-lindenstrauss transform for k-means and k-medians clustering. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 1027–1038, 2019.