大数据算法-2025 春

Lecture 9: JL 变换的应用：线性回归

2025.4.12

Lecturer: 丁虎　　*Scribe:* 沈俊杰，王运韬，王向禄

lecture 8 讨论了 JL 变换应用于聚类问题时的情况。本节讨论 JL 变换应用于线性变换问题时的细节。

1 线性回归 Linear Regression

给定:

$$A = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d-1} & 1 \\ x_{21} & x_{22} & \cdots & x_{2d-1} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd-1} & 1 \end{pmatrix}, \quad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

目标是:

$$\min_{\beta \in \mathbb{R}^d} ||A \cdot \beta - y||_2^2.$$

Remark 1.1. 样本数量 n 一般远大于特征数 d。对于一些特定问题, 我们可以认为 β 是 Sparse 的，即大部分 β_i 都是 0，这是稀疏线性回归（Sparse Linear Regression）。

对于线性回归问题，一种理解是将 A 看作一个线性变换，将 n 维的数据空间映射到 d 维的特征空间。是一种投影。对应的，矩阵 $A \cdot \beta = A(A^TA)^{-1}A^Ty = H \cdot y$，$H$ 是投影矩阵。

线性回归的解析解是:

$$\beta_{\text{opt}} = (A^TA)^{-1}A^Ty, \quad \text{其中 } A^TA \text{ 必须是可逆的。}$$

2 结合 JL 变换

首先我们需要明白，通过 JL 变换之后必然会产生误差，我们具体需要保证的误差是哪一种误差？关于此有许多种不同定义，此处我们采用如下的定义，即保证变换后求解出的 $\widetilde{\beta}_{\text{opt}}$ 能在原问题中也保持较好的效果。

Theorem 2.1. 给定 $A \in \mathbb{R}^{n\times d}$ 和 $y \in \mathbb{R}^n$，$A' = S \cdot A$, $y' = S \cdot y$ 是压缩后的数据，则对于 $\epsilon \in (0,1)$，有:

$$||A \cdot \beta_{opt} - y||_2^2 \leq (1+\theta(\epsilon))||A \cdot \widetilde{\beta}_{opt} - y||_2^2,$$

其中 β_{opt} 是原始线性回归的解析解，$\widetilde{\beta}_{opt}$ 是 A' 的线性回归的解析解，S 是我们的 *JL* 变换

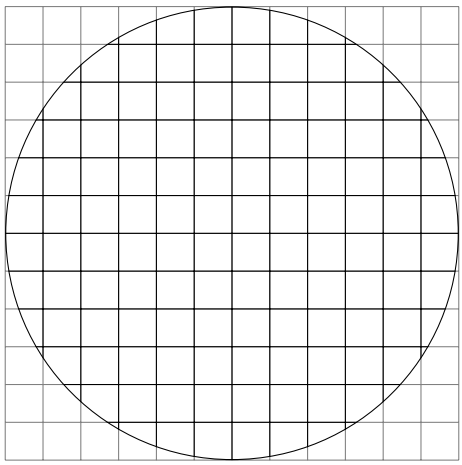
想要应用 JL 变换，第一个问题就是如何将应用于有限个点的 JL 变换拓展到无穷个点上，特殊的，这无穷个点在一个 k 维度子空间下（k 为某个常数）。

我们考虑一种简化的特殊情况，即这无穷个点在一个单位球面（此处是 l_2 单位球，即下图中的圆）上。我们便可以使用离散化的方法，通过有限个点的 JL 变换，再加上空间距离的性质（此处是欧几里得空间），可以得到对于无穷个点仍然成立的 JL 变换。下面给出证明。

Remark 2.2. 事实上，只要保证 $\mathbb{F} = \mathbb{R}^d$ 中的单位 l_2 球上的点 x 均满足，就可以推广到所有点 $y \in \mathbb{F}$ 上均成立。通过线性性可以很简单的转化为单位球上的问题。

对于一个 k 维单位球（此处单位球指 l_∞ 球，二维空间上，即为正方形），我们对其离散化，有一些相关的几何性质可以利用:

- cell 边长为 $\frac{\epsilon}{k}$(网格步长)
- cell 个数为 $(\frac{k}{\epsilon})^k$(体积相除)
- $\forall$ cell 内两点的距离 $d \leq \frac{\epsilon}{\sqrt{k}}$



考虑 l_∞ 球内的格点（即图中所有格点），我们对其进行 JL 变换，根据之前的 JL 变换结论，降维之后的结果为:

$$\mathbb{R}^n \longrightarrow \mathbb{R}^{\frac{k}{\epsilon^2} log(\frac{k}{\epsilon})}$$

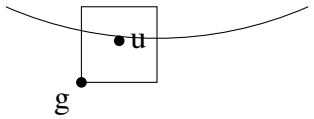
Remark 2.3. 这些在 n 维空间的 k 维子空间中的顶点，通过 JL 变换，映射到了一个低维空间。

现在我们需要证明的是，这样的 JL 变换，对于我们关注的那无穷个点（l_2 单位球上的点），是否满足 JL 定理类似的结果。

Proof. 对于 l_2 单位球上的点 u，我们关注其最近的格点 g，我们可以将 u 看作 g 附加一个扰动：

$$u = g + \sum_{i=1}^{k} \epsilon_i \cdot e_i, \quad \epsilon_i \in [0, \frac{\epsilon}{k}]$$

□



根据离散化的性质，我们可以做一些推导：

$$\begin{aligned}
\|S \cdot u\|_2 &= \|S \cdot g + S \cdot \sum_{i=1}^{k} \epsilon_i \cdot e_i\|_2 \\
&\leq \|S \cdot g\|_2 + \|S \cdot \sum_{i=1}^{k} \epsilon_i \cdot e_i\|_2 \\
&\leq \|S \cdot g\|_2 + \sum_{i=1}^{k} \epsilon_i \cdot \|S \cdot e_i\|_2 \\
&\leq \sqrt{1+\epsilon}\|g\|_2 + \epsilon \cdot \sqrt{1+\epsilon} \qquad \text{(由 JL 变换的性质)}
\end{aligned}$$

结合 $\|g\|_2 \leq \|u\|_2 + \|u - g\|_2 \leq 1 + \epsilon/\sqrt{k}$

$$\|S \cdot u\|_2 \leq 1 + \Theta(\epsilon)$$

此处的复杂度分析略去，注意是在 $\epsilon \to 0$ 时的讨论，事实上可以写成不带根号的形式，只是这样写方便推导。我们有：

$$\|S \cdot u\|_2 \leq 1 + \Theta(\epsilon)\|u\|_2$$

另一边的推导同理，我们可以得到：

$$\|S \cdot u\|_2 \geq 1 - \Theta(\epsilon)\|u\|_2$$

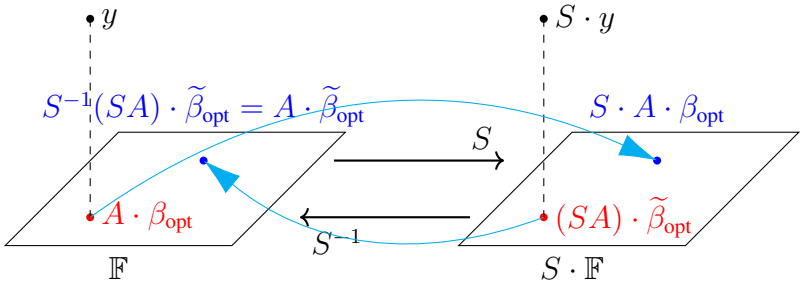
一般的，我们对于 JL 变换，有以下的推广：

Theorem 2.4. *JL 变换的推广*

$\mathbb{F}$ 为 $\mathbb{R}^n$ 上的一个 k 维子空间，假设 f 是一个 *JL* 变换，$f : \mathbb{R}^n \to \mathbb{R}^{\theta(\frac{k}{\epsilon^2}\log\frac{k}{\epsilon})}$ 则对 $\forall q \in \mathbb{F}$, $\|S \cdot q\| \in (1 \pm \epsilon)\|q\|$

回到 Linear Regression 的问题，我们将 $\{A_1, A_2, \cdots, A_{d-1}, y\}$ 看作一个 d 维子空间（参考开头，为了方便，我们认为 x 数据只有 d-1 个维度），那么我们可以应用推广后的 JL 变换。我们考虑一些和目标函数相关的量，并观察他们在变换前后的误差。

$$a_1 = A \cdot \beta_{\text{opt}} - y \qquad a_2 = A \cdot \widetilde{\beta}_{\text{opt}} - y$$
$$a_3 = S \cdot a_1 \qquad a_4 = S \cdot A \cdot \widetilde{\beta}_{\text{opt}} - S \cdot y$$



Remark 2.5. 蓝色点是经过变换的点，红色点是通过解 Linear Regression 得到的点。我们将 $\mathbb{R}^n$ 中 $\mathbb{R}^d$ 子空间中的点，通过 JL 变换到一个低维空间，即 $\mathbb{R}^{\theta(\frac{d}{\epsilon^2}\log\frac{d}{\epsilon})}$。

通过解的最优性，和 JL 变换的性质，我们可以得到：

$$\begin{cases} a_3^2 \leq (1+\theta(\epsilon))a_1^2 \\ a_4^2 \geq (1-\theta(\epsilon))a_2^2 \\ a_4^2 \leq a_3^2 \end{cases} \implies a_2^2 \leq \frac{1+\theta(\epsilon)}{1-\theta(\epsilon)}a_1^2 = (1+\theta(\epsilon))a_1^2$$

这已经完成了一边的证明。另一边的证明同理。所以我们可以得到 $A\widetilde{\beta}_{\text{opt}} - y$ 是一个 $1+\theta(\epsilon)$ 近似比的解。

3 时间复杂度分析

原先计算 LR 的时间复杂度包括矩阵求逆和矩阵乘法，分别为 $O(d^3)$ 和 $O(nd^2)$，当 $n \gg d$ 时，复杂度为 $O(nd^2)$。

改进后的 LR 复杂度包含 JL 变换和计算 LR 的复杂度，如果选择 Fast-JL 变换，那么分别是 $\widetilde{\theta}(d(n+\frac{1}{\epsilon}^2))$ 和 $\widetilde{\theta}(d^2 \cdot \frac{d}{\epsilon^2})$，当 $n \gg d$ 时，复杂度为 $\widetilde{\theta}(nd)$。可以看到有明显的改进。

4 关于别的 JL 变换应用

一个基础的问题，当一个高维的 k-Sparse 向量，我们需要多少维度来保证其误差？这个问题也同样可以用本节中的推广版的 JL 变换解决。

类似之前的 LR 问题，不过此时我们约束原始数据是 k-Sparse 的，即有 k 个非零元素。若此时要求 $\|x\| = 1$，那么 x 会落在 $\binom{n}{k}$ 个可能的单位球上。

那么根据之前的推广 JL 变换，我们可以得到降维结果为：

$$y = S \cdot x \in \mathbb{R}^{\theta(\frac{1}{\epsilon^2}\log\left[\binom{n}{k}\cdot(\frac{k}{\epsilon})^k\right])} \Longrightarrow y \in \mathbb{R}^m, \quad m = \theta\left(\frac{k\log n + k\log\frac{k}{\epsilon}}{\epsilon^2}\right)$$