

Johnson-Lindenstrauss 变换

张腾

2025 年 3 月 13 日

Johnson-Lindenstrauss (JL) 变换是一种保距的线性降维方法。

定理 1 (JL 变换). 给定 $\epsilon \in (0,1)$ 和 $m \in \mathbb{Z}^+$, 设 $d \in \mathbb{Z}^+$ 满足 $d \geq 4 \ln m / \epsilon^2 / 2 - \epsilon^3 / 3$, 则对 $\forall \mathcal{S} = \{\mathbf{x}_1, \dots, \mathbf{x}_m\} \subseteq \mathbb{R}^D$, 存在可在随机多项式时间内得到的线性映射 $f: \mathbb{R}^D \mapsto \mathbb{R}^d$ 使得对 $\forall \mathbf{x}_i, \mathbf{x}_j \in \mathcal{S}$ 有

$$(1 - \epsilon) \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \leq \|f(\mathbf{x}_i) - f(\mathbf{x}_j)\|_2^2 \leq (1 + \epsilon) \|\mathbf{x}_i - \mathbf{x}_j\|_2^2$$

引理 2. 设 $\mathbf{R} = (R_{ij})_{i \in [d], j \in [D]}$ 为随机矩阵, $R_{ij} \stackrel{\text{iid}}{\sim} \mathcal{N}(0,1)$, $\mathbf{u} \in \mathbb{R}^D$ 为任意固定向量, $\mathbf{v} = \mathbf{R}\mathbf{u}/\sqrt{d} \in \mathbb{R}^d$, 则 $\mathbb{P}(\|\mathbf{v}\|_2^2 \notin [(1 - \epsilon)\|\mathbf{u}\|_2^2, (1 + \epsilon)\|\mathbf{u}\|_2^2]) \leq 2/m^2$.

证明. 易知有

$$\begin{aligned} \mathbb{E}[v_i] &= \mathbb{E}\left[\frac{\sum_{j \in [D]} R_{ij} u_j}{\sqrt{d}}\right] = \frac{\sum_{j \in [D]} \mathbb{E}[R_{ij}] u_j}{\sqrt{d}} = 0 \\ \mathbb{E}[v_i^2] &= \mathbb{E}\left[\frac{1}{d} \left(\sum_{j \in [D]} R_{ij} u_j\right)^2\right] = \frac{1}{d} \mathbb{E}\left[\sum_{j \in [D]} R_{ij}^2 u_j^2 + \sum_{j \neq k} R_{ij} R_{ik} u_j u_k\right] = \frac{1}{d} \sum_{j \in [D]} \mathbb{E}[R_{ij}^2] u_j^2 = \frac{\|\mathbf{u}\|_2^2}{d} \end{aligned}$$

又 v_i 是独立正态 rv 的和, 也服从正态分布, 故 $v_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \|\mathbf{u}\|_2^2/d)$ 。

记 $\mathbf{x} = (\sqrt{d}/\|\mathbf{u}\|)\mathbf{v}$, 显然 $x_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0,1)$, 易知对 $\forall \lambda \geq 0$ 有

$$\begin{aligned} \mathbb{P}(\|\mathbf{v}\|_2^2 \geq (1 + \epsilon)\|\mathbf{u}\|_2^2) &= \mathbb{P}\left(\frac{\|\mathbf{u}\|_2^2}{d} \|\mathbf{x}\|_2^2 \geq (1 + \epsilon)\|\mathbf{u}\|_2^2\right) = \mathbb{P}(\|\mathbf{x}\|_2^2 \geq (1 + \epsilon)d) \\ &= \mathbb{P}(e^{\lambda \|\mathbf{x}\|_2^2} \geq e^{\lambda(1+\epsilon)d}) \\ &\leq \frac{\mathbb{E}[e^{\lambda \|\mathbf{x}\|_2^2}]}{e^{\lambda(1+\epsilon)d}} = \frac{\mathbb{E}[e^{\lambda \sum_{i \in [d]} x_i^2}]}{e^{\lambda(1+\epsilon)d}} = \frac{\prod_{i \in [d]} \mathbb{E}[e^{\lambda x_i^2}]}{e^{\lambda(1+\epsilon)d}} = (\sqrt{1 - 2\lambda} e^{\lambda(1+\epsilon)})^{-d} \end{aligned}$$

其中不等号是根据 Markov's 不等式, 最后一个等号是根据

$$\mathbb{E}[e^{\lambda x^2}] = \int_{\mathbb{R}} e^{\lambda x^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(1-2\lambda)x^2} dx = \frac{1}{\sqrt{1-2\lambda}}, \quad \text{if } \lambda \in [0, 1/2)$$

为了使上界尽可能紧, 令 $\sqrt{1-2\lambda} e^{\lambda(1+\epsilon)}$ 关于 λ 的导数为零可得 $\lambda = \epsilon/2(1+\epsilon) \in [0, 1/2)$, 回代可得

$$\mathbb{P}(\|\mathbf{v}\|_2^2 \geq (1 + \epsilon)\|\mathbf{u}\|_2^2) \leq (\sqrt{1/(1+\epsilon)} e^{\epsilon/2})^{-d} = ((1 + \epsilon)e^{-\epsilon})^{d/2}$$

根据 $\ln(1+x) \leq x - x^2/2 + x^3/3$ 可得

$$\mathbb{P}(\|\mathbf{v}\|_2^2 \geq (1+\epsilon)\|\mathbf{u}\|_2^2) \leq (e^{\ln(1+\epsilon)}e^{-\epsilon})^{d/2} \leq e^{-(x^2/2-x^3/3)^{d/2}} \leq e^{-2\ln m} = 1/m^2$$

同理可得

$$\begin{aligned} \mathbb{P}(\|\mathbf{v}\|_2^2 \leq (1-\epsilon)\|\mathbf{u}\|_2^2) &= \mathbb{P}\left(\frac{\|\mathbf{u}\|_2^2}{d}\|\mathbf{x}\|_2^2 \leq (1-\epsilon)\|\mathbf{u}\|_2^2\right) = \mathbb{P}(\|\mathbf{x}\|_2^2 \leq (1-\epsilon)d) \\ &= \mathbb{P}(e^{-\lambda\|\mathbf{x}\|_2^2} \geq e^{-\lambda(1-\epsilon)d}) \\ &\leq \frac{\mathbb{E}[e^{-\lambda\|\mathbf{x}\|_2^2}]}{e^{-\lambda(1-\epsilon)d}} = \frac{\mathbb{E}[e^{-\lambda\sum_{i \in [d]} x_i^2}]}{e^{-\lambda(1-\epsilon)d}} = \frac{\prod_{i \in [d]} \mathbb{E}[e^{-\lambda x_i^2}]}{e^{-\lambda(1-\epsilon)d}} = (\sqrt{1+2\lambda}e^{-\lambda(1-\epsilon)})^{-d} \\ &\leq ((1-\epsilon)e^\epsilon)^{d/2} = (e^{\ln(1-\epsilon)}e^\epsilon)^{d/2} \quad \leftarrow \lambda = \epsilon/2(1-\epsilon) \\ &\leq (e^{-\epsilon-\epsilon^2/2}e^\epsilon)^{d/2} = e^{-d\epsilon^2/4} \quad \leftarrow \ln(1-\epsilon) \leq -\epsilon - \epsilon^2/2 \\ &\leq e^{-2\ln m} = 1/m^2 \end{aligned}$$

综上, 根据 union bound 可得

$$\mathbb{P}(\|\mathbf{v}\|_2^2 \notin [(1-\epsilon)\|\mathbf{u}\|_2^2, (1+\epsilon)\|\mathbf{u}\|_2^2]) \leq 2/m^2$$

♣

JL 变换的证明. 注意 f 是线性映射, 因此 $f(\mathbf{x}_i) - f(\mathbf{x}_j) = f(\mathbf{x}_i - \mathbf{x}_j)$, 取 $\mathbf{u} = \mathbf{x}_i - \mathbf{x}_j$, $f(\mathbf{u}) = \mathbf{R}\mathbf{u}/\sqrt{d}$, 由引理知 $\mathbb{P}(\|f(\mathbf{u})\|_2^2 \notin [(1-\epsilon)\|\mathbf{u}\|_2^2, (1+\epsilon)\|\mathbf{u}\|_2^2]) \leq 2/m^2$, 共有 $\binom{m}{2}$ 个 $(\mathbf{x}_i, \mathbf{x}_j)$ 这样的样本对, 存在一对不保距的概率

$$\mathbb{P}(\cup E_i) \leq \sum_i \mathbb{P}(E_i) \leq \binom{m}{2} \frac{2}{m^2} = 1 - \frac{1}{m}$$

即所有样本对保距的概率至少为 $1/m$, 经过 $O(m)$ 次随机投影可将成功的概率提升至任意想要的值。♣