ĐẠI HỌC QUỐC GIA TPHCM

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA TOÁN - TIN HỌC

Báo cáo cuối kỳ

Đề tài: Bổ đề Johnson - Lindenstrauss và ứng dụng trong khoa học dữ liệu

Môn học: Xử lý đa chiều

Sinh viên thực hiện:
Lê Thành Đạt - 20110026
Nguyễn Lộc Phúc - 20110276
B.T. Thanh Phương - 20110280

Giảng viên hướng dẫn:
Kha Tuấn Minh

Ngày 26 tháng 10 năm 2023

Mục lục

1 Tổng quan về bổ đề Johnson-Lindenstrauss

Bổ đề Johnson–Lindenstrauss là một mệnh đề về việc ánh xạ một tập hợp các điểm trong không gian Euclid nhiều chiều về không gian ít chiều. Bổ đề khẳng định rằng với mọi tập hợp điểm trong không gian Euclid, đều tồn tại ảnh của các điểm này trong không gian có số chiều nhỏ hơn rất nhiều và không phụ thuộc số chiều ban đầu, đồng thời khoảng cách giữa các điểm gần như được giữ nguyên.

Bổ đề này có nhiều ứng dụng trong *cảm biến nén (compressing sensing)*, *học đa tạp (manifold learning)*, *giảm chiều (dimensionality reduction)*, và *nhúng metric (metric embeddings)*. Trong nhiều trường hợp, dữ liệu (chẳng hạn như văn bản hay hình ảnh) có thể được xem là các điểm trong không gian nhiều chiều. Tuy nhiên các thuật toán xử lý chúng thường chậm đi nhiều khi số chiều tăng lên. Do đó một phương pháp để làm tăng tốc độ thuật toán là làm giảm số chiều của dữ liệu trong khi vẫn giữ được thông tin quan trọng trong chúng. Bổ đề Johnson–Lindenstrauss là một kết quả cổ điển về vấn đề này.

2 Phát biểu bổ đề Johnson-Lindenstrauss

Cho $0 < \epsilon < 1$, $\alpha > 0$, n và k là các số nguyên dương sao cho

$$k \geq (4 + 2\alpha)\left(\frac{\epsilon^2}{2} - \frac{\epsilon^3}{3}\right)^{-1} \ln(n)$$

Khi đó với bất kỳ tập V chứa n điểm trong R^d, tồn tại một ánh xạ $f : R^d \to R^k$ sao cho với mọi $u, v \in V$ thì

$$(1 - \epsilon)||u - v||_2^2 \leq ||f(u) - f(v)||_2^2 \leq (1 + \epsilon)||u - v||_2^2 \quad (*)$$

Bất đẳng thức (*) xảy ra với xác suất không ít hơn $1 - n^{-\alpha}$

3 Chứng minh bổ đề Jonson-Lindeanstrauss

Ý tưởng chứng minh bổ đề này là chỉ ra được một ánh xạ sao cho xác suất để mọi $u, v \in V$ thỏa mãn bất đẳng thức (*) lớn hơn 0. Có hai cách thông dụng để chỉ ra được một ánh xạ như vậy là *Gaussian Random Projection* và *Sparse Random Projection*. Tuy nhiên trong cách chứng minh này chúng ta chỉ đề cập đến *Gaussian Random Projection* thông qua bổ đề sau

Bổ đề 3.1. *Cho $A \in R^{k \times d}$ là ma trận ngẫu nhiên với $A_{1 \leq i \leq k, 1 \leq j \leq d} \sim \mathcal{N}(0, 1)$ và vector cố định $w \in R^d$. Khi đó ta có $E\left[||x||_2^2\right] = ||w||_2^2$ với $x = \frac{1}{\sqrt{k}} Aw$*

Chứng minh. Ta có

$$\begin{aligned}\mathbb{E}\left[||x||_2^2\right] &= \mathbb{E}\left[\sum_{i=1}^{k} x_i^2\right] = \sum_{i=1}^{k}\mathbb{E}\left[x_i^2\right] \\ &= \sum_{i=1}^{k}\mathbb{E}\left[\left(\sum_{j=1}^{d}\frac{1}{\sqrt{k}}A_{i,j}w_j\right)^2\right] \\ &= \frac{1}{k}\sum_{i=1}^{k}\mathbb{E}\left[\sum_{j=1}^{d}\sum_{m=1}^{d}A_{i,j}w_jA_{i,m}w_m\right] \\ &= \frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{d}\sum_{m=1}^{d}w_jw_m\mathbb{E}\left[A_{i,j}A_{i,m}\right]\end{aligned}$$

Để ý rằng

$$\mathbb{E}\left[A_{i,j}A_{i,m}\right] = \begin{cases}1 & \text{khi } j = m \\ 0 & \text{khi } j \neq m\end{cases}$$

Suy ra

$$\mathbb{E}\left[||x||_2^2\right] = \frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{d}\sum_{m=1}^{d}w_jw_m\mathbb{E}\left[A_{i,j}A_{i,m}\right] = \frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{d}w_j^2 = ||w||_2^2$$

□

Bổ đề 3.1 chỉ ra rằng *Gaussian Random Projection* bảo toàn khoảng cách giữa hai điểm bất kỳ trong không gian gốc khi chiếu xuống không gian con bất kỳ. Một cách tự nhiên, ta mong muốn phép chiếu này cũng thỏa mãn bổ đề Johnson-Lindenstrauss. Ta phát biểu lại bổ đề Johnson-Lindenstrauss theo bổ đề 3.1 như sau

Bổ đề 3.2. *Cho $0 < \epsilon < 1$, $\alpha > 0$, n và k là các số nguyên dương sao cho*

$$k \geq (4 + 2\alpha)\left(\frac{\epsilon^2}{2} - \frac{\epsilon^3}{3}\right)^{-1}\ln(n)$$

Khi đó với bất kỳ tập V chứa n điểm trong R^d, và $A \in R^{k\times d}$ là ma trận ngẫu nhiên với $A_{1\leq i\leq k, 1\leq j\leq d} \sim \mathcal{N}(0,1)$, thì xác suất để mọi $u, v \in V$ thỏa mãn

$$(1-\epsilon)||u-v||_2^2 \leq \left|\left|\frac{1}{\sqrt{k}}Au - \frac{1}{\sqrt{k}}Av\right|\right|_2^2 \leq (1+\epsilon)||u-v||_2^2 \quad (**)$$

lớn hơn 0. Hơn nữa, bất đẳng thức (**) xảy ra với xác suất không ít hơn $1 - n^{-\alpha}$

Để rút gọn bất đẳng thức (**) ta đặt $w = u - v$, $x = \frac{1}{\sqrt{k}}Aw$. Nhiệm vụ của chúng ta bây giờ là chứng minh với mọi $u, v \in V$ thì $P\left\{(1-\epsilon)\,||w||_2^2 \leq ||x||_2^2 \leq (1+\epsilon)\,||w||_2^2\right\} > 0$. Trước hết ta đi tìm xác suất để một cặp vector u, v bất kỳ thỏa mãn bất đẳng thức (**). Ta có

$$\begin{aligned}& P\left\{(1-\epsilon)\,||w||_2^2 \leq ||x||_2^2 \leq (1+\epsilon)\,||w||_2^2\right\} \\ = \;& 1 - P\left\{||x||_2^2 \notin \left[||(1-\epsilon)w||_2^2, (1+\epsilon)\,||w||_2^2\right]\right\} \\ = \;& 1 - P\left\{||x||_2^2 > (1+\epsilon)||w||_2^2\right\} - P\left\{||x||_2^2 < (1-\epsilon)||w||_2^2\right\}\end{aligned}$$

Như vậy ta cần đi tìm $P\left\{||x||_2^2 > (1+\epsilon)||w||_2^2\right\}$ và $P\left\{||x||_2^2 < (1-\epsilon)||w||_2^2\right\}$. Đặt $y = \frac{Aw}{||w||_2} = \frac{1}{\sqrt{k}}\frac{x}{||w||_2}$ suy ra $||y||_2^2 = \frac{1}{k}\frac{||x||_2^2}{||w||_2^2}$. Ta có

$$\begin{aligned} P\left\{||x||_2^2 > (1+\epsilon)||w||_2^2\right\} &= P\left\{\frac{||y||_2^2\,||w||_2^2}{k} > (1+\epsilon)\,||w||_2^2\right\} \\ &= P\left\{||y||_2^2 > (1+\epsilon)k\right\} \\ &= P\left\{\lambda\,||y||_2^2 > \lambda(1+\epsilon)k\right\} \\ &= P\left\{e^{\lambda||y||_2^2} > e^{\lambda(1+\epsilon)k}\right\} \end{aligned}$$

Để đánh giá đẳng thức trên ta sử dụng bổ đề sau

Bổ đề 3.3. *(Bất đẳng thức Markov) Cho X là biến ngẫu nhiên không âm và số thực dương a. Khi đó ta có*

$$P(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$$

Chứng minh. Ta chứng minh bất đẳng thức trên trong trường hợp X là biến liên tục, trường hợp X là biến rời rạc làm tương tự. Gọi f là hàm mật độ xác suất của X, ta có

$$\begin{aligned} \mathbb{E}[X] &= \int_{-\infty}^{\infty} xf(x)dx = \int_0^{\infty} xf(x)dx = \int_0^a xf(x)dx + \int_a^{\infty} xf(x)dx \\ &\geq \int_a^{\infty} af(x)dx = aP(X \geq a) \end{aligned}$$

Suy ra $P(X \leq a) \leq \frac{\mathbb{E}[X]}{a}$ □

Áp dụng bổ đề 3.3 ta có

$$P\left\{e^{\lambda||y||_2^2} > e^{\lambda(1+\epsilon)k}\right\} < \frac{\mathbb{E}\left[e^{\lambda||y||_2^2}\right]}{e^{\lambda(1+\epsilon)k}} = \frac{\mathbb{E}\left[\prod_{i=1}^{k} e^{\lambda y_i^2}\right]}{e^{\lambda(1+\epsilon)k}} = \frac{\prod_{i=1}^{k}\mathbb{E}\left[e^{\lambda y_i^2}\right]}{e^{\lambda(1+\epsilon)k}}$$

Để ý rằng $y_i = \frac{1}{||w||_2}\sum_{j=1}^{d} A_{i,j}w_j \sim \mathcal{N}\left(0, \frac{1}{||w||_2^2}\sum_{j=1}^{d} w_j^2\right) = \mathcal{N}(0,1)$ suy ra $y_i^2 \sim \chi(1)$. Tới đây ta có bổ đề quen thuộc sau

Bổ đề 3.4. *Cho X là biến ngẫu nhiên và $X \sim \chi^2(1)$. Khi đó hàm sinh moment của X có dạng như sau*

$$M_X(\lambda) = \mathbb{E}\left[e^{\lambda X}\right] = \begin{cases} (1-2\lambda)^{\frac{-1}{2}} & \text{khi } \lambda < \frac{1}{2} \\ \text{không tồn tại} & \text{khi } \lambda \geq \frac{1}{2} \end{cases}$$

Áp dụng bổ đề 3.4 ta có

$$\frac{\prod_{i=1}^{k}\mathbb{E}\left[e^{\lambda y_i^2}\right]}{e^{\lambda(1+\epsilon)k}} = \frac{(1-2\lambda)^{\frac{-k}{2}}}{e^{\lambda(1+\epsilon)k}} = g(\lambda)$$

Mong muốn của ta là cực tiểu hóa $g(\lambda)$, do đó ta sẽ tìm $0 < \lambda < \frac{1}{2}$ để $g(\lambda)$ đạt giá trị cực tiểu. Ta có

$$\frac{\partial g(\lambda)}{\partial \lambda} = \frac{\partial}{\partial \lambda} \frac{(1-2\lambda)^{\frac{-k}{2}}}{e^{\lambda(1+\epsilon)k}} = 0 \Rightarrow \lambda = \frac{\epsilon}{2(1+\epsilon)}$$

Thay $\lambda = \frac{\epsilon}{2(1+\epsilon)}$ vào $g(\lambda)$ ta được

$$g(\lambda) = \left[\frac{\left(\frac{1}{1+\epsilon}\right)^{\frac{-1}{2}}}{e^{\frac{\epsilon}{2}}} \right]^k = \left[e^{\ln(1+\epsilon)-\epsilon}\right]^{\frac{k}{2}}$$

Để đánh giá đẳng thức trên ta sử dụng bổ đề sau

Bổ đề 3.5. *(Bất đẳng thức Taylor) Cho $0 < \epsilon < 1$. Khi đó ta có*

$$\ln(1+\epsilon) - \epsilon < \frac{-\epsilon^2}{2} + \frac{\epsilon^3}{3}$$

Áp dụng bổ đề 3.5 ta có

$$\left[e^{\ln(1+\epsilon)-\epsilon}\right]^{\frac{k}{2}} < \left[e^{\frac{-\epsilon^2}{2}+\frac{\epsilon^3}{3}}\right]^{\frac{k}{2}} \leq \left[e^{\frac{-(4+2\alpha)\ln(n)}{k}}\right]^{\frac{k}{2}} = n^{-(\alpha+2)}$$

Từ đây ta suy ra $P\left\{||x||_2^2 > (1+\epsilon)||w||_2^2\right\} < n^{-(\alpha+2)}$, tương tự ta có $P\left\{||x||_2^2 < (1-\epsilon)||w||_2^2\right\} < n^{-(\alpha+2)}$. Như vậy ta có

$$P\left\{||x||_2^2 \notin \left[||(1-\epsilon)w||_2^2, (1+\epsilon)\,||w||_2^2\right]\right\} < 2n^{-(\alpha+2)}$$

Mặt khác, tập V có n điểm tức là ta có tổng cộng $\delta = \frac{n(n-1)}{2}$ cặp điểm riêng biệt, vậy nên ta cần phải tính xác suất kết hợp của $\delta = \frac{n(n-1)}{2}$ cặp điểm này. Để tính được xác suất này ta cần phải dùng bổ đề sau

Bổ đề 3.6. *(Bất đẳng thức Boole) Cho tập hợp gồm n các biến cố $B_1, B_2, ..., B_n$. Khi đó ta có*

$$P\left(\bigcup_{i=1}^{n} B_i\right) \leq \sum_{i=1}^{n} P(B_i)$$

Chứng minh. Để chứng minh bổ đề này ta sử dụng nguyên lý quy nạp. Dễ thấy trường hợp $n = 1$ thì bất đẳng thức hiển nhiên đúng. Giả sử bất đẳng thức đúng trong trường hợp $n = n'$, ta cần chứng minh bất đẳng thức cũng đúng trong trường hợp $n = n' + 1$. Từ công thức $P(B_i \cup B_j) = P(B_i) + P(B_j) - P(B_i \cap B_j)$ ta có

$$\begin{aligned} P\left(\bigcup_{i=1}^{n'+1} B_i\right) &= P\left(\bigcup_{i=1}^{n'} B_i\right) + P\left(B_{n'+1}\right) - P\left(\bigcup_{i=1}^{n'} B_i \cap B_{n'+1}\right) \\ &\leq P\left(\bigcup_{i=1}^{n'} B_i\right) + P\left(B_{n'+1}\right) \\ &\leq \sum_{i=1}^{n'} P(B_i) + P\left(B_{n'+1}\right) = \sum_{i=1}^{n'+1} P(B_i) \end{aligned}$$

□

Gọi B_i $(1 \leq i \leq \delta)$ là biến cố của một cặp điểm bất kỳ trong tập V không thỏa mãn bất đẳng thức (**). Áp dụng bổ đề 3.6 ta có

$$P\left\{\bigcup_{i=1}^{\delta} B_i\right\} \leq \sum_{i=1}^{\delta} P(B_i) < \frac{n(n-1)}{2}\dot{2}n^{-(\alpha+2)} = (n-1)n^{-(\alpha+1)}$$

Suy ra

$$P\left\{\bigcap_{i=1}^{\delta} \overline{B_i}\right\} = P\left\{\overline{\bigcup_{i=1}^{\delta} B_i}\right\} = 1 - P\left\{\bigcup_{i=1}^{\delta} B_i\right\} > 1 - (n-1)n^{-(\alpha+1)} = 1 - n^{-\alpha} + n^{-(\alpha+1)} > 1 - n^{-\alpha} > 0$$

Mặt khác xác suất để mọi $u, v \in V$ thỏa mãn bất đẳng thức (**) chính là $P\left\{\bigcap_{i=1}^{\delta} \overline{B_i}\right\}$, vậy nên ta có điều phải chứng minh

4 Một phát biểu khác của bổ đề Johnson-Lindenstrauss

Cho $0 < \epsilon < 1$, $c > 0$, đặt $V = \left\{x_i : i = \overline{1...M}; i \in \mathbb{N}\right\} \subset \mathbb{R}^m$ là tập hợp các điểm nằm trong $\mathbb{R}^m$. Nếu $n \geq \frac{c}{\epsilon^2} \ln M$ thì tồn tại một ánh xạ tuyến tính $A : \mathbb{R}^m \to \mathbb{R}^n$ sao cho với mọi i khác j thì

$$1 - \epsilon \leq \frac{||A(x_i) - A(x_j)||}{||x_i - x_j||} \leq 1 + \epsilon \quad (*)$$

Chứng minh. Tương tự như chứng minh ở mục **3**, ta cũng sử dụng *Gaussian Random Projection*. Tuy nhiên ta sẽ sử dụng thêm một bổ đề mới là *Gaussian Concentration*. Ý tưởng chứng minh như sau: Đầu tiên chúng ta xây dựng một ma trận ngẫu nhiên kích thước $m \times n$ trong đó các phần tử là các biến ngẫu nhiên thuộc phân phối Gaussian, sau đó ta sử dụng *random projection* đối với hiệu của hai điểm bất kỳ thuộc V và ảnh sẽ thu được là vector ngẫu nhiên theo phân phối Graussian. Sau đó ta sử dụng bổ đề *Gaussian Concentration* trên các hàm Lipschitz để chỉ ra được xác xuất của hai điểm bất kì thuộc V thỏa mãn bất đẳng thức (*), xác suất này sẽ phụ thuộc vào các biến n, ϵ và M, Bằng cách chọn n hợp lí, xác suất này sẽ lớn hơn 0, điều đó chứng tỏ sự tồn tại của một ánh xạ thỏa mãn bổ đề. Để hiểu rõ hơn ta đi sẽ vào chứng minh chi tiết hơn

Giả sử G là một ma trận ngẫu nhiên với các giá trị g_{ij} thuộc phân phối Gaussian. Đặt $y = Gx$, với y là ảnh của ánh xa tuyến tính G lên điểm $x \in V$, khi đó ta có

$$y = \begin{bmatrix} g_{1,1} & \cdots & a_{1,m} \\ \vdots & \ddots & \vdots \\ a_{n,1} & \cdots & a_{n,m} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} = \begin{bmatrix} g_{1,1}x_1 + & \cdots + & g_{1,m}x_m \\ \vdots & \vdots & \ddots \\ g_{n,1}x_1 + & \cdots + & g_{n,m}x_m \end{bmatrix}$$

Nếu chúng ta áp dụng ánh xạ tuyến tính ngẫu nhiên G cho 2 điểm $x^p, x^q \in V$, thì để chứng minh bổ đề Johnson Lindenstrauss Lemma ta cần có

$$\mathbb{P}(\forall x^p, x^q \in V : (1-\epsilon)||x^p - x^q|| \leq ||y^p - y^q|| \leq (1+\epsilon)||x^p - x^q||) > 0 \quad (1)$$

Để chứng minh bất đẳng thức (1), trước hết ta chỉ chứng minh xác suất của hai điểm bất kỳ x^p và x^q lớn hơn 0. Từ cách mà ta định nghĩa y, ta suy ra y là vector ngẫu nhiên có kích thước n với các phần tử thuộc phân phối Gaussian, có giá trị mean bằng 0 và variance bằng $\sum_{i=1}^{m} x_i^2$, tức là

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad \text{với } y_i = \sum_{j=1}^{m} g_{i,j} x_j \sim \mathcal{N}\left(0, \sum_{j=1}^{m} x_j^2\right)$$

Để giải thích rõ hơn, ta có $g_{i,j} \sim \mathcal{N}(0,1)$ suy ra $g_{i,j}x_j \sim \mathcal{N}(0, x_j^2)$, nên $y_i = \sum_{j=1}^{m} g_{i,j}x_j \sim \mathcal{N}\left(0, \sum_{j=1}^{m} x_j^2\right)$.

Tới đây ta có thể nói $y = ||x||z$ với $z = (z_1, ..., z_n) \sim \mathcal{N}(0,1)$. Áp dụng đẳng thức này với hai điểm bất kỳ x^p và x^q ta được $y^p - y^q = ||x^p - x^q||z$, suy ra $||y^p - y^q|| = ||x^p - x^q||||z||$. Từ đây ta cần phải chứng minh

$$\mathbb{P}\left((1-\epsilon)||x^p - x^q|| \le ||x^p - x^q||||z|| \le (1+\epsilon)||x^p - x^q||\right) > 0 \quad (2)$$

Để chứng minh bất đẳng thức (2) ta sử dụng bổ đề *Gaussian Concentration*

Bổ đề 4.1. *Xét ánh xạ Lipschitz $F : \mathbb{R}^m \to \mathbb{R}$, tồn tại số thực dương $L > 0$ và*

$$||F(x) - F(y)|| \le L||x - y|| \quad \forall x, y \in \mathbb{R}^m$$

Đặt $g = (g_i)_{i \le m} \in \mathbb{R}^m$ và $g_i \sim \mathcal{N}(0,1)$. Khi đó với bất kì $t > 0$, ta có

$$\mathbb{P}\left(||F(g) - \mathbb{E}[F(g)]|| \ge t\right) \le 2exp\left(-\frac{t^2}{4L^2}\right)$$

Để dử dụng bổ đề *Gaussian Concentration* ta cần một ánh xạ Lipschitz. Hai bổ đề sau sẽ cho chúng ta thấy điều đó

Bổ đề 4.2.

$$||v|| = sup\left\{\sum_{i}^{m} \alpha_i v_i : ||\alpha|| = 1\right\} \quad \textit{với mọi } v \in \mathbb{R}^m$$

Bổ đề 4.3. *Cho A là tập bị chặn bới a trong $\mathbb{R}^m$. Xét ánh xạ $F : \mathbb{R}^m \to \mathbb{R}$ được định nghĩa bởi:*

$$F(x) = \sup_{a \in A} \langle a, x \rangle$$

Khi đó F được gọi là ánh xạ Lipschitz

Chứng minh. Ta có

$$\begin{aligned}
||F(x) - F(y)|| &= \left|\left|\sup_{a\in A}\langle a, x\rangle - \sup_{a\in A}\langle a, y\rangle\right|\right| \\
&= \left|\left|\sup_{a\in A}(a_1x_1 + ... + a_mx_m) - \sup_{a\in A}(a_1y_1 + ... + a_my_m)\right|\right| \\
&\le \left|\left|\sup_{a\in A} a_1(x_1 - y_1) + ... + a_m(x_m - y_m)\right|\right| \\
&= \sup_{a\in A} ||a_1(x_1 - y_1) + ... + a_m(x_m - y_m)|| \\
&\le \sup_{a\in A} ||a||||x - y||
\end{aligned}$$

□

Từ bổ đề 4.2 và 4.3 ta có thể chọn $F(z) = ||z|| \sup_{||\alpha=1||} \left\{ \sum_i^m \alpha_i z_i \right\}$. Áp dụng bổ đề *Gaussian Concentration* ta có

$$\mathbb{P}(|||z|| - \mathbb{E}||z||| \geq t) \leq 2exp(-t/4) \quad \text{với } t \geq 0$$

Khi đó phần bù của bất đẳng thức là

$$\mathbb{P}(|||z|| - \mathbb{E}||z||| \leq t) \leq 1 - 2exp(-t/4)$$

Ta có $\mathbb{E}||z||$ là hằng số nên đặt $t = \epsilon\mathbb{E}||z||$

$$\mathbb{P}\left(1 - \epsilon \leq \frac{||z||}{\mathbb{E}||z||} \leq 1 + \epsilon\right) \leq 1 - 2\exp\left(-\frac{\epsilon^2(\mathbb{E}||z||)^2}{4}\right)$$

Nhân 2 vế phương trình bên trong xác xuất với $||x^p - x^q||$ ta được

$$\mathbb{P}\left((1-\epsilon)||x^p - x^q|| \leq \frac{||z||(||x^p - x^q||)}{\mathbb{E}||z||} \leq (1+\epsilon)||x^p - x^q||\right) \geq 1 - 2exp\left(-\frac{\epsilon 2(\mathbb{E}||z||)^2}{4}\right)$$

Mà ta có $||y^p - y^q|| = (||x^p - x^q||)||z||$ nên suy ra

$$\mathbb{P}\left((1-\epsilon)||x^p - x^q|| \leq \frac{||y^p - y^q||}{\mathbb{E}||z||} \leq (1+\epsilon)||x^p - x^q||\right) \geq 1 - 2exp\left(-\frac{\epsilon^2(\mathbb{E}||z||)^2}{4}\right)$$

Tuy nhiên đây k phải xác xuất chính mà ta muốn tính toán, chúng ta muốn so sánh $||x^p - x^q||$ với $||y^p - y^q||$ không phả là $\frac{||y^p - y^q||}{E||z||}$. Để loại bỏ phần tử mẫu số $\mathbb{E}||z||$ chúng ta thay đổi ánh xạ tuyến tính G bằng cách chia cho hằng số $E||z||$, khi đó $\hat{G} = \frac{1}{\mathbb{E}||z||}G$. Nếu ta lấy $\hat{y} = \hat{G}x$ thì

$$||\hat{y}^p - \hat{y}^q|| = \frac{||z||(||x^p - x^q||)}{\mathbb{E}||z||}$$

Khi đó bất đẳng thức trở thành

$$\mathbb{P}((1-\epsilon)||x^p - x^q|| \leq ||\hat{y}^p - \hat{y}^q|| \leq (1+\epsilon)||x^p - x^q||) \geq 1 - 2exp\left(-\frac{\epsilon^2(\mathbb{E}||z||)^2}{4}\right)$$

Tới đây ta giả sử ánh xạ tuyến tính G ban đầu chính là $\hat{G}$ vậy thì ta có

$$\mathbb{P}((1-\epsilon)||x^p - x^q|| \leq ||y^p - y^q|| \leq (1+\epsilon)||x^p - x^q||) \geq 1 - 2exp\left(-\frac{\epsilon^2(\mathbb{E}||z||)^2}{4}\right) \quad (3)$$

Ta đã có xác suất để 2 điểm bất kỳ thuộc V thỏa mãn (1). Bây giờ ta cần tìm xác suất để tất cả các cặp điểm bất kỳ trong tập V thỏa mãn (1).Gọi A_{pq} là xác suất của biến cố mà 2 điểm bất kỳ trong tập V thỏa mãn

$$(1-\epsilon)||x^p - x^q|| \leq ||y^p - y^q|| \leq (1+\epsilon)||x^p - x^q||$$

Để tính mọi cặp điểm thỏa mảng bất đẳng thức (1), ta tính xác xuất của mọi giao điểm của A_{pq} tức là $\mathbb{P}(\cap A_{pq})$. Ta có

$$\mathbb{P}(A_{pq}) \geq 1 - 2exp\left(-\frac{\epsilon^2(\mathbb{E}||z||)^2}{4}\right)$$

Lấy phần bù của nó ta được

$$\mathbb{P}({A_{pq}}^c) \leq 2exp\left(-\frac{\epsilon^2(\mathbb{E}||z||)^2}{4}\right)$$

Nhận thấy có M điểm trong V nên tổng số cặp điểm là $\binom{M}{2} < \frac{M^2}{2}$, do đó dựa theo bất đẳng thức Boole, ta có

$$\mathbb{P}(\cap {A_{pq}}^c) \leq 2\binom{M}{2}exp\left(-\frac{\epsilon^2(\mathbb{E}||z||)^2}{4}\right) \leq 2M^2exp\left(-\frac{\epsilon^2(\mathbb{E}||z||)^2}{4}\right)$$

Theo định luật DeMorgan ta có $\cup {A_{pq}}^c = (\cap A_pq)^c$, suy ra

$$\mathbb{P}((\cap A_{pq})^c) \leq 2M^2exp\left(-\frac{\epsilon^2(\mathbb{E}||z||)^2}{4}\right)$$

Lấy phần bù ta được

$$\mathbb{P}(\cap A_{pq}) \geq 1 - 2M^2exp\left(-\frac{\epsilon^2(\mathbb{E}||z||)^2}{4}\right)$$

Từ đây ta suy ra

$$\mathbb{P}(\forall x^p, x^q \in V : (1-\epsilon)||x^p - x^q|| \leq ||y^p - y^q|| \leq (1+\epsilon)||x^p - x^q||) \geq 1 - 2M^2exp\left(-\frac{\epsilon^2(\mathbb{E}||z||)^2}{4}\right)$$

Cuối cùng để thỏa mãn (1) chúng ta cần điều kiện là $1 - 2M^2exp\left(-\frac{\epsilon^2(\mathbb{E}||z||)^2}{4}\right) > 0$ điều này sẽ dẫn đến

$$1 - 2M^2exp\left(-\frac{\epsilon^2(\mathbb{E}||z||)^2}{4}\right) > 0 \Leftrightarrow \ln\left(\frac{1}{M^2}\right) > -\frac{\epsilon^2}{4}(\mathbb{E}||z||)^2 \Leftrightarrow (\mathbb{E}||z||)^2 > \frac{8}{\epsilon^2}\ln M$$

Điều kiện cần thỏa mảng là $(\mathbb{E}||z||)^2 > \frac{8}{\epsilon^2}\ln M$ mà $\mathbb{E}||z|| \geq k\sqrt{n}$ khi k là hằng số và n là số chiều của vector z do đó nếu $n \geq \frac{8c}{\epsilon^2}\ln M$ với c là hằng số thì điều kiện $1 - 2M^2exp\left(-\frac{\epsilon^2(\mathbb{E}||z||)^2}{4}\right) > 0$ là đúng. Vậy nên ta có điều phải chứng minh □

5 Ứng dụng của bổ đề Johnson - Lindenstrauss trong khoa học dữ liệu

Bổ đề Johnson - Lindenstrauss được sử dụng nhiều trong khoa học dữ liệu, bổ dề nói rằng các *phép chiếu ngẫu nhiên (Random Projection* tạo thành một cơ sở để xây dựng các phép nhúng *almost isometry* - Phép nhúng *almost isometry* là một phương pháp được sử dụng để chiếu tập dữ liệu trong không gian Euclide lên một không gian mới sao cho khoảng cách giữa các điểm trong không gian mới được bảo toàn một cách gần đúng, đồng thời số chiều của không gian mới không phụ thuộc vào số chiều của không gian gốc. *Phép chiếu ngẫu nhiên* được sử dụng rộng rãi trong các bài toán sau đây

5.1 Giảm chiều

Giảm chiều (Dimensionality Reduction): là quá trình giảm số lượng các biến đầu vào trong tập dữ liệu mà vẫn giữ lại thông tin quan trọng. Khi làm việc với các tập dữ liệu đa chiều, việc giảm chiều dữ liệu giúp giảm độ phức tạp tính toán của thuật toán và cải thiện hiệu suất của chúng.

Để làm được điều này, bổ đề Johson-Lindenstrauss sử dụng một phép chiếu ngẫu nhiên từ không gian ban đầu sang không gian mới có số chiều thấp hơn. Phép chiếu ngẫu nhiên sẽ bảo toàn khoảng cách Euclid giữa các điểm dữ liệu với độ sai số xác định được. Cụ thể, nếu chọn một số dương ϵ và một số nguyên dương k thỏa mãn điều kiện: $k \geq \Omega(\epsilon, n)$ thì với mỗi cặp điểm dữ liệu trong không gian ban đầu, khoảng cách Euclid giữa chúng sẽ được bảo toàn với độ sai số không quá $1 \pm \epsilon$ trong không gian mới.

Tuy nhiên, việc giảm chiều dữ liệu bằng phép chiếu ngẫu nhiên cũng có một số hạn chế nhất định chẳng hạn như làm mất mát một số thông tin quan trọng. Hơn nữa, bổ đề Johnson-lindenstrauss chỉ hoạt động hiệu quả đối với tập dữ liệu lớn. Khi tập dữ liệu không đủ lớn, phép chiếu ngẫu nhiên có thể làm tăng số chiều của tập dữ liệu nếu ta chọn các tham số không khéo, hoặc có giảm chiều dữ liệu nhưng không đáng kể

Ngoài ra, việc kết hợp JL với các kỹ thuật khác như *PCA (Principal Component Analysis)*, *LLE (Locally Linear Embedding)* hay *MDS (Multidimensional Scaling)* cũng là một phương pháp hiệu quả để giảm chiều dữ liệu đối với các tập dữ liệu lớn. Để giải thích rõ hơn chúng ta sẽ tìm hiểu thêm trong phần code.

5.2 Locality Sensitive Hashing

Locality Sensitive Hashing (LSH) là một kỹ thuật được sử dụng để *tìm kiếm xấp xỉ lân cận gần nhất (approximate nearest neighbor search)* trong không gian có số chiều lớn một cách nhanh chóng và hiệu quả.

Một trong nhưng bài toán quan trọng trong Khoa học máy tính là *tìm kiếm lân cận gần nhất (nearest neighbor search)*. Bài toán này được phát biểu như sau: Cho một tập các điểm $\mathcal{P}$ gồm n điểm trong không gian d chiều và một số thực r. Thiết kế cấu trúc dữ liệu trả lời truy vấn dạng: cho trước một điểm $p \in \mathcal{P}$, tìm một (vài) điểm $q \in \mathcal{P}$ mà $d(p, q) \leq r$, trong đó $d(p, q)$ là khoảng cách từ điểm p tới điểm q.

Ý tưởng của LSH là sử dụng một hàm băm $h(.)$ để băm các điểm dữ liệu trong $\mathcal{P}$ vào một bảng $\mathcal{T}[.]$ sao cho nếu hai điểm p, q gần nhau thì khả năng cao p, q sẽ được băm vào cùng một ô. Ngược lại, p, q sẽ bị băm vào hai ô khác nhau. Khi trả lời truy vấn, ta chỉ việc tính mã băm $h(p)$ và đưa ra tất cả các điểm lưu trong ô $\mathcal{T}[h(p)]$. Hàm băm như vậy đươc gọi là kiểu *locality sensitive*.

Khi kết hợp với bổ đề Johnson-Lindenstrauss với LSH sẽ làm giảm độ phức tạp tính toán. cụ thể nó sẽ đưa độ phức tạp tính toán của LSH về $\mathcal{O}(\log(n))$. Điều này giúp cải thiện đáng kể thời gian truy vấn của LSH đối với các tập dữ liệu lớn.

5.3 Compressed Sensing

Compressed Sensing (hay còn gọi là compressive sampling, hoặc sparse sampling là một kỹ thuật xử lý tín hiệu để thu và tái tạo tín hiệu một cách hiệu quả bằng cách tìm nghiệm của các hệ phương trình tuyến tính.

Một trong những kỹ thuật quan trọng trong Compressed Sensing là *Restricted Isometry Property (RIP)*, được sử dụng để đánh giá tính khả dụng của các ma trận ánh xạ trong quá trình nén dữ liệu, chính là một ứng dụng của bổ đề Johnson-Lindenstrauss. RIP đảm bảo rằng ma trận ánh xạ được sử dụng trong quá trình nén dữ liệu bảo toàn được khoảng cách giữa các điểm dữ liệu. Cụ thể, một ma trận A được gọi là thỏa mãn RIP với hằng số ϵ nếu nó giữ lại khoảng cách giữa các điểm dữ liệu trong không gian mới với sai số tương đối không quá ϵ.

Việc đánh giá tính khả dụng của ma trận ánh xạ thông qua RIP rất quan trọng trong quá trình nén dữ liệu, vì nó đảm bảo thông tin của tập dữ liệu trong không gian mới được giữ lại với độ chính xác cao, giúp đảm bảo tính chính xác của quá trình phân tích dữ liệu.

Tuy nhiên, việc tìm kiếm ma trận ánh xạ thỏa mãn RIP có thể là một vấn đề khó khăn, vì nó liên quan đến việc tối ưu hóa một hàm số phi tuyến. Một số phương pháp được đề xuất để tìm kiếm ma trận ánh xạ thỏa mãn RIP, bao gồm phương pháp *Gradient Descent*, phương pháp sử dụng ma trận *Toeplitz* và phương pháp tối ưu hóa hàm lồi.

5.4 Một số ứng dụng khác

Bên cạnh các ứng dụng đã được giới thiệu, bổ đề Johnson - Lindenstrauss còn được sử dụng trong các bài toán như *Phân cụm dữ liệu (Data Clustering)*, *Phân loại dữ liệu (Data Classification)* hay *Phân tích hồi quy (Regression Analysis)*.

6 Minh họa ứng dụng của bổ đề Johnson-Lindenstrauss

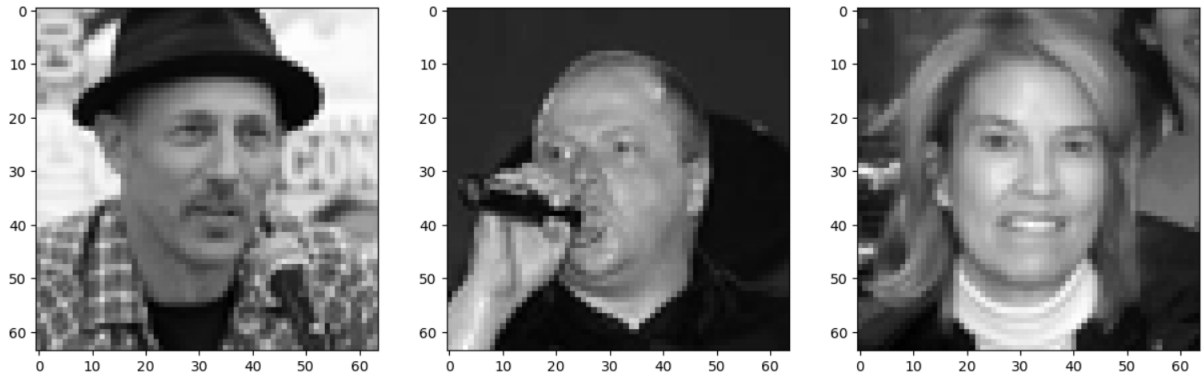
6.1 Tóm tắt kết quả code

Trong phần này nhóm sẽ trình bày phần kết quả code ứng dụng của bổ đề Johnson-Lindenstrauss trong bài toán giảm chiều dữ liệu. Mục tiêu của nhóm là so sánh mức độ hiệu quả của việc sử dụng thuật toán PCA riêng lẻ với việc kết hợp thuật toán PCA với Random Projection. Hai phương pháp Random Projection mà nhóm sử dụng là Gaussian Random Projection và Sparse Random Projection. Ngôn ngữ lập trình mà nhóm sử dụng là Python.

Các thư viện được nhóm sử dụng

```python
import numpy as np
from sklearn.preprocessing import MinMaxScaler
from sklearn.decomposition import PCA
import matplotlib.pyplot as plt
from sklearn.random_projection import SparseRandomProjection, GaussianRandomProjection
from sklearn.metrics import accuracy_score
from sklearn.linear_model import LogisticRegression
import time
```

Tập dữ liệu mà nhóm sử dụng là 11938 hình ảnh khuôn mặt 64×64. Nhãn dán của tập dữ liệu là nam / nữ với 0 là nam và 1 là nữ.



Đầu tiên ta sử dụng MinMaxscaler để tiền xử lý tập dữ liệu trên, sau đó ta chọn α và ϵ phù hợp để số chiều giảm đủ nhỏ với xác suất xảy ra đủ lớn. Ở đây nhóm chọn $\alpha = 0.2$ và $\epsilon = 0.6$, khi này thì số nguyên K nhỏ nhất thoả mãn bổ đề Jonhson-Lindentrauss là 383 với xác suất xảy ra không ít hơn $84,7\%$.

Tiếp theo, ta sử dụng LogisticsRegression để phân lớp cho tập dữ liệu gốc thì ta thấy thời gian chạy của mô hình này là 3.94328s với độ chính xác là xấp xỉ 0.83112.

Bây giờ ta sẽ so sánh kết quả của mô hình LogisticsRegression trước và sau khi giảm chiều dữ liệu. Đàu tiên, ta viết thuật toán JL-PCA như sau

```python
def jl_pca(X, epsilon, alpha, n_components):
    # Tính toán số chiều mới theo JL lemma
    n, d = X.shape # số chiều ban đầu
    k = int(np.ceil((4 + 2*alpha) * np.log(n) / (epsilon**2/2 - epsilon**3/3)))
    print(k)
    # Tạo ma trận ngẫu nhiên cho JL lemma
    R = np.random.normal(size=(d, k))/np.sqrt(k)
    print("Voi xac suat khong nho hon:", 1 - n**(-alpha))
    # Áp dụng JL lemma lên dữ liệu X
    X_jl = X@R

    # Thực hiện PCA trên dữ liệu mới giảm chiều
    pca = PCA(n_components=n_components)
    pca.fit(X_jl)
    print(f"Explained variance ratio: {sum(pca.explained_variance_ratio_)}")

    X_pca = pca.transform(X_jl)

    return X_pca
```

Bây giờ ta sử dụng LogisticsRegression cho tập dữ liệu sau khi giảm chiều bằng thuật toán JL-PCA thì ta thấy thời gian chạy của mô hình này là 0.32229s với độ chính xác xấp xỉ 0.79954.

So sánh kết quả của hàm mà nhóm tự viết với hàm Gaussian Random Projection của thư viện Sklearn thì ta thấy độ chính xác gần như tương tự nhau, nhưng thời gian chạy của thư viện sklearn lại cao hơn so với hàm của nhóm tự viết, cụ thể là 0.563311s. Lý do là bởi hàm của nhóm tự viết có tham số ϵ còn hàm của thư viện sklearn không có sẵn tham số này, nên nó phải tự tính toán ϵ, do đó thời gian chạy lâu hơn.

Tiếp theo, ta sử dụng mô hình KNeighborsClassifier thay cho mô hình LogisticRegression trên. Mục đính là để kiểm chứng thuật toán JL-PCA. Đầu tiên ta vẫn sử dụng KNeighborsClassifier với tập dữ liệu gốc thì ta thấy thời gian chạy là 7.51282s với độ chính xác xấp xỉ 0.84184. Sau đó ta sử dụng mô hình này với tập dữ liệu đã được giảm chiều bằng thuật toán JL-PCA thì ta thấy thời gian chạy là 0.51849s với độ chính xác xấp xỉ 0.83992. Nhìn chung độ chính xác của mô hình trước và sau khi giảm chiều dữ liệu là tương đương nhau, tuy nhiên thời gian chạy lại tăng nhanh gần mười lần.

Cũng với mô hình KNeighborsClassifier ta thử so sánh tốc độ chạy và độ chính xác của hai phương pháp Gaussian Random Projection và Sparse Random Projection. Ở đây ta sẽ sử dụng hàm của thư viện sklearn. Cụ thể là hàm Gaussian Random Projection cho thời gian chạy là 0.86888s với độ chính xác xấp xỉ 0.83481, còn hàm Sparse Random Projection cho thời gian chạy là 2.05143s với độ chính xác xấp xỉ 0.83774. Qua đây ta thấy với độ chính xác gần như tương đương nhau, tuy nhiên phương pháp Gaussian Random Projection lại cho kết quả chạy nhanh hơn hai lần so với phương pháp Sparse Random Projection.

6.2 Nhận xét chung

Ta thấy rằng chênh lệch độ chính xác giữa sử dụng dữ liệu gốc và dữ liệu được giảm chiều trong mô hình KNeighborsClassifier hơn hẳn mô hình LogisticRegression, lí do là trong KNeighborsClassifier tính toán các láng giềng dựa theo khoảng cách mà thuật toán JL-PCA lại giảm chiều mà bảo toàn khoảng cách, vì vậy sử dụng giảm chiều bằng thuật toán JL-PCA cho mô hình KNeighborsClassifier thì bảo toàn thông tin của tập dữ liệu tốt hơn. Ngoài ra đối với bài toán phân loại có tập dữ liệu dạng hình ảnh như bài toán trên thì phương pháp Gaussian Random Projection sẽ cho kết quả tốt hơn so với phương pháp Sparse Random Projection.

7 Lời kết

Trong bài báo cáo này nhóm đã thực hiện chứng minh bổ đề Johnson-Lindenstrauss và chỉ ra một số ứng dụng của nó trong khoa học dữ liệu, Tuy bài báo cáo còn nhiều hạn chế nhưng các thành viên trong nhóm cũng đã hoàn thành tốt nhất có thể, qua lần nghiên cứu này nhóm đã hiểu rỏ hơn về bổ đề Johnson-Lindenstrauss và một số vấn đề liên quan tới nó. Nhóm cũng muốn cảm ơn những người đã giúp đỡ và hỗ trợ nhóm trong quá trình nghiên cứu và thực hiện báo cáo. Đặt biệt nhóm xin chân thành cảm ơn thầy Kha Tuấn Minh đã tạo điều kiện thuận lợi trong cho nhóm trong quá trình nghiên cứu và thực hiện báo cáo.

Dưới đây là bảng phân công công việc của các thành viên trong nhóm

Tên công việc	**Tên thành viên thực hiện**
Trình bày chứng minh bổ đề	Cách 1: Nguyễn Lộc Phúc, cách 2: Bùi Thị Thanh Phương
Trình bày ứng dụng bổ đề	Bùi Thị Thanh Phương, Nguyễn Lộc Phúc
Trình bày code	Lê Thành Đạt
Chuẩn bị Silde	Bùi Thị Thanh Phương
Thuyết trình	Nguyễn Lộc Phúc, Lê Thành Đạt

Tài liệu

[1] Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of johnson and lindenstrauss. Random Structures and Algorithms, 22(1):60–65, 2002.

[2] Benyamin Ghojogh, Ali Ghodsi, Fakhri Karray, and Mark Crowley. Johnson- lindenstrauss lemma, linear and nonlinear random projections, random fourier features, and random kitchen sinks: Tutorial and survey. arXiv preprint arXiv:2108.04172, 2021.

[3] https://www.math.toronto.edu/undergrad/projects-undergrad/Project03.pdf

[4] https://www.cs.princeton.edu/~smattw/Teaching/Fa19Lectures/lec9/lec9.pdf