ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO BÀI TẬP CÁ NHÂN LAB 3
MÔN TRỰC QUAN HÓA DỮ LIỆU

Giảng viên: Bùi Tiến Lên
Lê Nhựt Nam
Lê Ngọc Thành
Nguyễn Thị Thu Hằng

Lớp: 21_21
MSSV: 21120201
Họ và tên: Bùi Đình Bảo

Học kỳ 2 – Năm học 2023–2024

Mục lục

TỰ ĐÁNH GIÁ MỨC ĐỘ HOÀN THÀNH:

Công việc	Mức độ hoàn thành
Studying t-SNE - Motivation - Problem statement - Mathematics behind t-SNE - t-SNE algorithms - Discussion	 100% 100% 100% 100% 100%
Implementation t-SNE	100%
Making comparison between t-SNE and PCA.	100%
Overall comprehension of the submitted source code.	100%
Bonus points	100%
Tổng	**110%**

CƠ SỞ LÝ THUYẾT CỦA TSNE:

1. Động lực nghiên cứu về TSNE:

t-SNE (t-distributed Stochastic Neighbor Embedding) giúp xóa bỏ những hạn chế trong việc trực quan hóa dữ liệu đa chiều. Não bộ con người của chúng ta khó khăn trong việc giải thích thông tin vượt quá ba chiều. Các nhà nghiên cứu cần một cách để phân tích các tập dữ liệu phức tạp với nhiều tính năng bằng cách nhúng chúng vào không gian chiều thấp hơn, thường là 2D hoặc 3D, cho mục đích trực quan hóa.

Những lý do nghiên cứu về t-SNE:

Trực quan hóa những điều không nhìn thấy: Dữ liệu thực tế thường có nhiều features. t-SNE cho phép chúng ta chiếu dữ liệu đa chiều này lên một không gian chiều thấp hơn mà chúng ta có thể trực quan hóa, cho phép chúng ta khám phá các cấu trúc và mối quan hệ ẩn trong dữ liệu.

Bảo tồn cấu trúc cục bộ: Không giống như Phân tích thành phần chính (PCA), tập trung vào việc nắm bắt phương sai toàn cục, t-SNE ưu tiên duy trì sự giống nhau giữa các điểm lân cận trong không gian đa chiều. Điều này cho phép chúng ta thấy cách các điểm dữ liệu nhóm lại với nhau dựa trên các thuộc tính vốn có của chúng.

Giải quyết các mối quan hệ phi tuyến tính: PCA giả định mối quan hệ tuyến tính giữa các tính năng. Tuy nhiên, t-SNE có thể xử lý hiệu quả hơn các mối quan hệ phi tuyến tính, cung cấp một cách biểu diễn chính xác hơn các cấu trúc dữ liệu phức tạp.

2. Đặt vấn đề:

Trong nhiều tình huống thực tế, dữ liệu có kích thước cao thường rất khó phân tích và diễn giải. Sự phức tạp của dữ liệu này thường làm che mờ các mẫu và mối quan hệ tiềm ẩn, khiến việc thu thập thông tin và đưa ra quyết định trở nên khó khăn. Các kỹ thuật trực quan hóa và giảm chiều truyền thống có thể không hiệu quả trong việc nắm bắt cấu trúc cục bộ và toàn cầu trong dữ liệu.

t-SNE là một phương pháp không tham số không giám sát, giảm chiều phi tuyến tính (unsupervised non-parametric method, non-linear dimensionality reduction) có thể giải quyết được những vấn đề bên trên.

Ý tưởng cốt lõi của t-SNE: là ánh xạ các điểm dữ liệu có độ chiều cao sang không gian có độ chiều thấp hơn, thường là hai hoặc ba chiều, theo cách bảo toàn các mối quan hệ cục bộ giữa các điểm. Nó đạt được điều này bằng cách đo lường sự tương đồng giữa các điểm dữ liệu trong không gian có độ chiều cao và biểu diễn sự tương đồng này dưới dạng xác suất. Sau đó, nó xây dựng một phân phối xác suất tương tự trong không gian có độ chiều thấp hơn và giảm thiểu sự khác biệt giữa hai phân phối bằng cách sử dụng một kỹ thuật gọi là gradient descent. Quá trình này cho phép t-SNE hiệu quả trong việc nắm bắt cấu trúc cục bộ của dữ liệu, làm cho nó đặc biệt hữu ích để trực quan hóa các tập dữ liệu phức tạp và khám phá các mẫu ý nghĩa.

Các yêu cầu chính để thực hiện t-SNE:

- Chuẩn Bị Dữ Liệu: Tiền xử lý và chuẩn hóa dữ liệu kích thước cao để đảm bảo tính tương thích với thuật toán t-SNE.
- Giảm Chiều: Áp dụng thuật toán t-SNE để chuyển đổi dữ liệu kích thước cao sang không gian có kích thước thấp hơn (thường là 2D hoặc 3D).
- Trực Quan Hóa: Tạo các biểu đồ trực quan của dữ liệu đã giảm chiều để xác định các mẫu, cụm và các điểm ngoại lai.
- Đánh Giá Hiệu Suất: Đánh giá chất lượng của biểu diễn t-SNE bằng cách xem xét việc bảo toàn cấu trúc cục bộ và toàn cục của dữ liệu.
- Ứng Dụng: Sử dụng các biểu đồ trực quan t-SNE để hướng dẫn phân tích sâu hơn, như lựa chọn đặc trưng, phân cụm, và phát triển các mô hình học máy, nhằm cải thiện hiệu suất và giảm tình trạng overfitting.

Những vấn đề thách thức trong t-SNE:

- Khả Năng Mở Rộng: t-SNE có thể yêu cầu nhiều tài nguyên tính toán cho các tập dữ liệu lớn, cần có các triển khai và tinh chỉnh tham số hiệu quả.
- Độ Nhạy Tham Số: Chất lượng đầu ra của t-SNE phụ thuộc vào các siêu tham số như perplexity và learning rate, cần được lựa chọn cẩn thận.
- Khả Năng Diễn Giải: Mặc dù t-SNE rất xuất sắc trong việc trực quan hóa, việc diễn giải các biểu đồ không gian thấp và liên kết chúng lại với không gian cao ban đầu có thể phức tạp.

3. Cơ sở toán học của TSNE:

Tính toán độ đo tương đồng bằng xác suất có điều kiện:

(Có dẫn chứng công thức ở phần thuật toán)

Đối với mỗi điểm dữ liệu trong không gian nhiều chiều, chúng ta tính toán độ tương đồng của nó với mọi điểm khác bằng cách sử dụng phân phối Gaussian. Độ tương đồng này dựa trên khoảng cách giữa các điểm.

Tương tự, trong không gian ít chiều, chúng ta tính toán độ tương đồng giữa các điểm bằng cách sử dụng phân phối Student-t.

Hàm mục tiêu (tối ưu hóa hàm chi phí):

(Có dẫn chứng công thức ở phần thuật toán)

Chúng ta muốn giảm thiểu sự khác biệt giữa độ tương đồng của các điểm trong không gian nhiều chiều và các điểm tương ứng của chúng trong không gian ít chiều. Chúng ta đo lường sự khác biệt này bằng cách sử dụng độ phân kỳ Kullback-Leibler (KL).

Độ phân kỳ KL đo lường mức độ một phân phối xác suất lệch khỏi phân phối khác. Trong trường hợp của chúng ta, nó định lượng mức độ khác biệt giữa các độ tương đồng cặp đôi trong không gian nhiều chiều và không gian ít chiều.

Thuật toán hạ dốc (Gradient Descent):

(Có dẫn chứng công thức ở phần thuật toán)

Để giảm thiểu độ phân kỳ KL, chúng ta sử dụng gradient descent. Kỹ thuật tối ưu hóa lặp này điều chỉnh vị trí của các điểm trong không gian ít chiều.

Tại mỗi bước lặp, chúng ta tính toán gradient của hàm chi phí đối với vị trí của các điểm trong không gian ít chiều.

Gradient này chỉ ra hướng mà chúng ta nên di chuyển mỗi điểm để giảm sự khác biệt giữa độ tương đồng nhiều chiều và ít chiều.

Bằng cách cập nhật vị trí của các điểm dựa trên gradient này, chúng ta dần dần hội tụ đến một điểm mà ở đó độ tương đồng ít chiều khớp chặt chẽ với độ tương đồng trong không gian nhiều chiều.

Lưu ý thêm:

SNE thực hiện tìm kiếm nhị phân cho giá trị của σ_i tạo ra một P_i với một perplexity cố định được chỉ định bởi người dùng. Perplexity được định nghĩa là:

$$\text{Perp}(P_i) = 2^{H(P_i)}$$

Trong đó $H(P_i)$ là entropy Shannon của P_i đo bằng bit:

$$H(P) = -\sum_j p_j \log_2 p_j$$

4. Thuật toán TSNE:

Mã giả của thuật toán t-SNE:

Algorithm: Simple version of t-Distributed Stochastic Neighbor Embedding

Data: Dataset: $\mathcal{X} = \{x_1, x_2, \dots, x_n\}$

Cost function parameters: Perplexity $(Perp)$

Optimization parameters: Number of iterations (T), Learning rate (η), Momentum $(\alpha(t))$

Result: Low-dimensional data representation $\mathcal{Y}^{(T)} = \{y_1, y_2, \dots, y_n\}$

begin

compute pairwise affinities $p_{j|i}$ with perplexity $(Perp)$

set $p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$

sample initial solution $\mathcal{Y}^{(0)} = \{y_1, y_2, \dots, y_n\}$ from $\mathcal{N}(0, 10^{-4}I)$

for $t = 1$ **to** T **do**

compute low-dimensional affinities q_{ij}

compute gradient $\frac{\partial C}{\partial y}$

set $\mathcal{Y}^{(t)} = \mathcal{Y}^{(t-1)} + \eta \frac{\partial C}{\partial y} + \alpha(t)\left(\mathcal{Y}^{(t-1)} - \mathcal{Y}^{(t-2)}\right)$

end

end

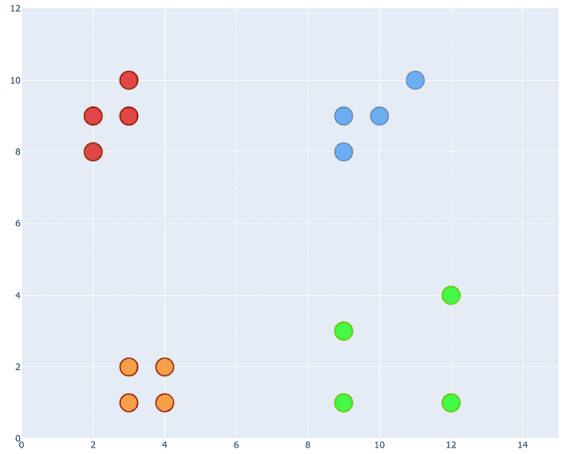
Diễn giải thuật toán t-SNE:

Bước 1: Nhập dữ liệu và thiết lập các tham số

- Dữ liệu: $\mathcal{X} = \{x_1, x_2, \dots, x_n\}$

- Các tham số của hàm mục tiêu: perplexity ($Perp$)

- Các tham số tối ưu hóa: số lần lặp (T), tốc độ học (η), momentum ($\alpha(t)$)

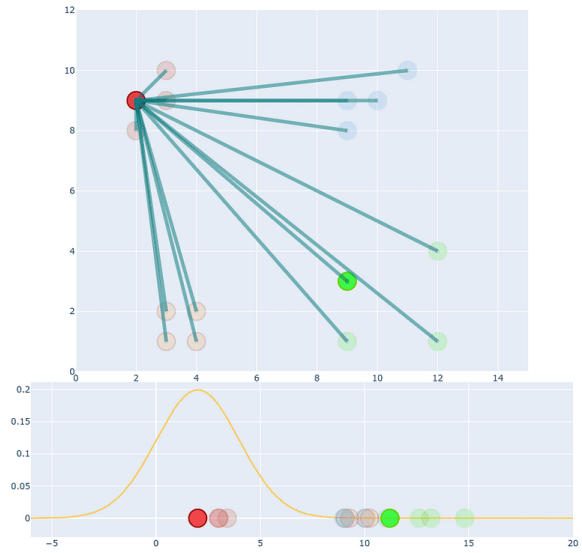


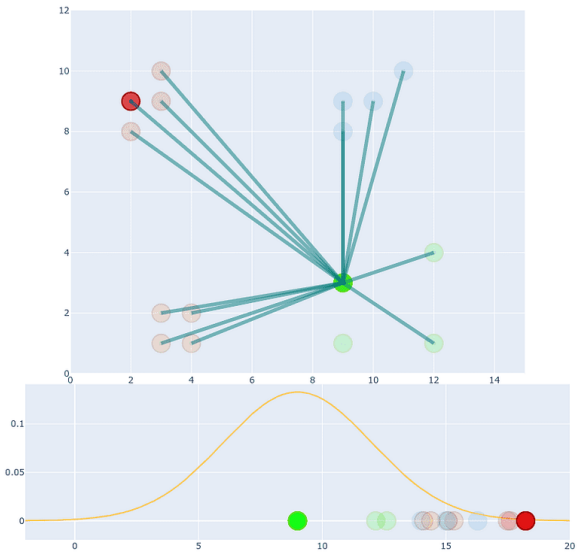
Bước 2: Tính toán các xác suất có điều kiện (trên tập dữ liệu nhiều chiều hay chưa được giảm chiều) $p_{j|i}$ với perplexity ($Perp$)

$$p_{j|i} = \frac{\exp\left(-\left|\left|x_i - x_j\right|\right|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\left|\left|x_i - x_k\right|\right|^2 / 2\sigma_i^2\right)}$$

⇨ Xác suất đối xứng p_{ij}

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$$





Bước 3: Khởi tạo ngẫu nhiên tập hợp ban đầu $\mathcal{Y}^{(0)}$, chính là tập dữ liệu sau khi giảm chiều $\mathcal{Y}^{(0)} = \{y_1, y_2, \dots, y_n\}$ được lấy mẫu từ phân phối chuẩn với phương sai nhỏ $\mathcal{N}(0, 10^{-4}I)$

Bước 4: Chạy vòng lặp $t = 1$ đến $t = T$, với T là số vòng lặp trong t-SNE

- Tính toán các xác suất có điều kiện (trên tập dữ liệu ít chiều hay sau khi giảm chiều) q_{ij}

$$q_{ij} = \frac{\left(1 + \left||y_i - y_j|\right|^2\right)^{-1}}{\sum_{k \neq l}\left(1 + \left||y_k - y_l|\right|^2\right)^{-1}}$$

- Tính toán gradient của hàm chi phí C

$$C = KL(P \parallel Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

⇨ Gradient của hàm chi phí

$$\frac{\partial C}{\partial y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)\left(1 + |y_i - y_j|^2\right)^{-1}$$

- Cập nhật các điểm $\mathcal{Y}^{(\mathcal{T})}$ trên tập dữ liệu ít chiều bằng Gradient Descent

$$\mathcal{Y}^{(t)} = \mathcal{Y}^{(t-1)} + \eta \frac{\partial C}{\partial \mathcal{Y}} + \alpha(t)\left(\mathcal{Y}^{(t-1)} - \mathcal{Y}^{(t-2)}\right)$$

Bước 5: Xuất kết quả

Biểu diễn dữ liệu trong không gian ít chiều $\mathcal{Y}^{(\mathcal{T})} = \{y_1, y_2, \dots, y_n\}$

Các siêu tham số chính trong thuật toán t-SNE:

Độ phân kỳ (Perplexity):

- Độ phân kỳ có lẽ là siêu tham số quan trọng nhất trong t-SNE. Nó có thể được coi như là một thước đo của số lượng lân cận hiệu quả cho mỗi điểm.

- Giá trị của độ phân kỳ ảnh hưởng đến sự cân bằng giữa các khía cạnh cục bộ và toàn cục của dữ liệu. Độ phân kỳ nhỏ nhấn mạnh cấu trúc cục bộ, trong khi độ phân kỳ lớn hơn mang lại nhiều cấu trúc toàn cục hơn.

- Các giá trị điển hình cho độ phân kỳ nằm trong khoảng từ 5 đến 50, nhưng điều này có thể thay đổi tùy thuộc vào tập dữ liệu. Thường được khuyến nghị thử nghiệm với các giá trị khác nhau để xem chúng ảnh hưởng như thế nào đến kết quả.

Tốc độ học (Learning Rate):

- Tốc độ học xác định kích thước bước tại mỗi lần lặp khi di chuyển về phía một cực tiểu của hàm chi phí.

- Tốc độ học quá cao có thể khiến thuật toán dao động và bỏ lỡ cực tiểu toàn cục, trong khi tốc độ học quá thấp có thể dẫn đến quá trình huấn luyện dài và có thể bị kẹt trong một cực tiểu cục bộ.

- Các giá trị phổ biến cho tốc độ học nằm trong khoảng từ 10 đến 1000. Một lần nữa, thử nghiệm với các giá trị khác nhau là chìa khóa để tìm ra cài đặt tốt nhất cho một tập dữ liệu cụ thể.

Số lần lặp (Number of Iterations):

- Siêu tham số này kiểm soát số lần lặp mà thuật toán chạy trước khi kết thúc.

- Nếu số lần lặp quá ít, thuật toán có thể không hội tụ hoàn toàn. Nếu quá nhiều, bạn có thể lãng phí tài nguyên tính toán mà không đạt được nhiều về chất lượng của việc nhúng.

- Số lần lặp mặc định thường được đặt là 1000, nhưng con số này có thể cần phải tăng lên đối với các tập dữ liệu lớn hơn.

5. Thảo luận về thuật toán TSNE:

Các điểm lý thuyết liên quan đến t-SNE:

Diễn giải Cụm (Interpreting Clusters): t-SNE có thể tiết lộ các cụm và cấu trúc địa phương rất hiệu quả. Tuy nhiên, khoảng cách giữa các cụm hoặc vị trí tương đối của các cụm trong biểu đồ có thể không có ý nghĩa giải thích. Tránh diễn giải quá mức các mối quan hệ toàn cầu.

Trục không có ý nghĩa (Interpreting Clusters): Trục trong t-SNE không có ý nghĩa giải thích, đó là lý do tại sao t-SNE chỉ được sử dụng để trực quan hóa và không để dự đoán.

Sự quan trọng của Perplexity (Perplexity Matters): Perplexity là một siêu tham số quan trọng trong t-SNE. Nó tương ứng với số lượng lân cận hiệu quả. Không có giá trị duy nhất phù hợp với mọi trường hợp; các giá trị khác nhau có thể tiết lộ các cấu trúc khác nhau, vì vậy hãy thử nghiệm với một loạt các giá trị. Các giá trị phổ biến là từ 5 đến 50.

Tính tái lập (Reproducibility): t-SNE bắt đầu với một khởi tạo ngẫu nhiên, dẫn đến kết quả khác nhau mỗi lần chạy. Nếu tính tái lập là quan trọng, hãy đặt một seed ngẫu nhiên. Ngoài ra, nhiều lần chạy với các khởi tạo khác nhau có thể cho thấy một bức tranh toàn diện hơn về cấu trúc dữ liệu của bạn.

Cân bằng dữ liệu (Scaling the Data): Các bước tiền xử lý như cân bằng hoặc chuẩn hóa dữ liệu của bạn, đặc biệt nếu các tính năng có các thang đo khác nhau, có thể có ảnh hưởng đáng kể đến kết quả của t-SNE.

Lời nguyền của chiều không gian (Curse of Dimensionality): t-SNE có thể giảm thiểu nhưng không thể hoàn toàn khắc phục lời nguyền của chiều không gian. Dữ liệu có chiều rất cao có thể yêu cầu các bước khác, như giảm chiều không gian ban đầu bằng PCA, trước khi áp dụng t-SNE.

Tốc độ học và số lần lặp lại (Learning Rate and Number of Iterations): Ngoài perplexity, các tham số khác như tốc độ học và số lần lặp lại cũng ảnh hưởng đến kết quả. Tốc độ học

quá cao hoặc quá thấp có thể dẫn đến nhúng kém, và số lần lặp lại không đủ có thể nghĩa là thuật toán không hoàn toàn hội tụ.

t-SNE không phải là một giải pháp toàn diện (It's Not a Silver Bullet): Mặc dù t-SNE là một công cụ mạnh mẽ, nó không phù hợp với mọi loại tập dữ liệu hoặc phân tích. Đôi khi các kỹ thuật giảm chiều khác như PCA, UMAP, hoặc MDS có thể phù hợp hơn.

Ưu điểm của t-SNE:

- Nếu được thực hiện đúng cách, có thể cung cấp các hình ảnh trực quan rất trực quan vì nó bảo toàn cấu trúc cục bộ của dữ liệu trong các chiều ít hơn.

Nhược điểm của t-SNE:

- Tốn kém về mặt tính toán.

- Không tốt trong việc bảo toàn cấu trúc toàn cục.

- Nhạy cảm với các siêu tham số.

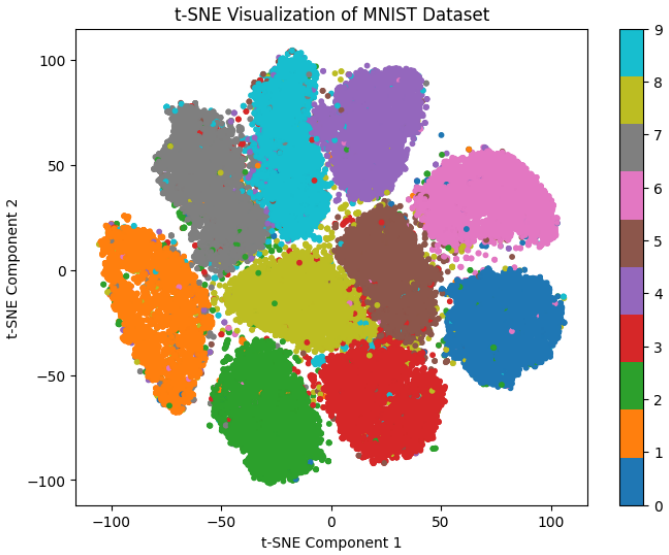
- Có thể bị kẹt ở cực tiểu địa phương.

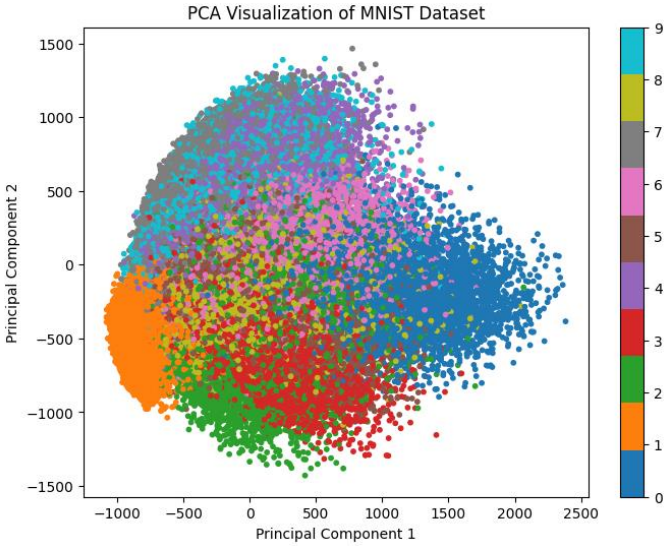
- Việc diễn giải là một thách thức.

SO SÁNH GIỮA TSNE VÀ PCA

t-SNE	PCA
Đây là một kỹ thuật giảm kích thước phi tuyến tính.	Nó là một kỹ thuật giảm kích thước tuyến tính.
Kết quả tạo ra không duy nhất.	Kết quả tạo ra là duy nhất.
Nó cố gắng bảo toàn cấu trúc (cụm) dữ liệu cục bộ (bao gồm cả toàn cục nhưng không phổ biến).	Nó cố gắng bảo tồn cấu trúc toàn cục của dữ liệu.
Đây là một trong những kỹ thuật giảm chiều tốt nhất.	Nó không hoạt động tốt so với t-SNE.
Nó liên quan đến các siêu thông số như độ phức tạp, tốc độ học và số bước.	Nó không liên quan đến siêu tham số.
Nó có thể xử lý các ngoại lệ.	Nó bị ảnh hưởng rất nhiều bởi các ngoại lệ.
Đây là một thuật toán không xác định hoặc ngẫu nhiên (Stochastic).	PCA là một thuật toán xác định (Deterministic).
Nó hoạt động bằng cách giảm thiểu khoảng cách giữa các điểm trong không gian Gaussian.	Nó hoạt động bằng cách xoay các vectơ để bảo toàn phương sai.
Chúng ta không thể bảo toàn phương sai thay vào đó chúng ta có thể bảo toàn khoảng cách bằng cách sử dụng siêu tham số.	Chúng ta có thể quyết định mức độ phương sai cần duy trì bằng cách sử dụng các giá trị riêng.
t-SNE có thể tốn kém về mặt tính toán, đặc biệt đối với các bộ dữ liệu nhiều chiều có số lượng điểm dữ liệu lớn.	PCA ít tốn kém về mặt tính toán hơn t-SNE, đặc biệt đối với các bộ dữ liệu lớn.
Nó được thiết kế đặc biệt để trực quan hóa và được biết là hoạt động tốt hơn về mặt này.	Nó có thể được sử dụng để trực quan hóa dữ liệu nhiều chiều trong không gian ít chiều.

Nó phù hợp hơn cho các bộ dữ liệu có thể phân tách phi tuyến tính (không bị giới hạn bởi loại bộ dữ liệu).	Nó phù hợp cho các bộ dữ liệu có thể phân tách tuyến tính.
Nó chủ yếu được sử dụng để trực quan hóa và EDA.	Nó có thể được sử dụng để trích xuất đặc trưng.
t-SNE ít nhạy cảm hơn với thứ tự của các điểm dữ liệu.	PCA có thể nhạy cảm với thứ tự của các điểm dữ liệu.
Cung cấp diễn giải chủ quan về dữ liệu thành các cụm.	Chỉ đơn thuần xoay các trục tọa độ để giảm chiều.





REFERENCES

1. Visualizing Data using t-SNE by Laurens van der Maaten: https://drive.google.com/file/d/1nkFmQSQq9gjqf8d4DbMd7mZjAyVkCQE-/view
2. Mastering t-SNE(t-distributed stochastic neighbor embedding) | by Sachinsoni | Medium: https://medium.com/@sachinsoni600517/mastering-t-sne-t-distributed-stochastic-neighbor-embedding-0e365ee898ea
3. t-SNE clearly explained. An intuitive explanation of t-SNE… | by Kemal Erdem (burnpiro) | Towards Data Science: https://towardsdatascience.com/t-sne-clearly-explained-d84c537f53a
4. Difference between PCA VS t-SNE – GeeksforGeeks: https://www.geeksforgeeks.org/difference-between-pca-vs-t-sne/
5. The Ultimate Comparison Between PCA and t-SNE Algorithm: https://blog.dailydoseofds.com/p/the-ultimate-comparison-between-pca
6. Sử dụng ChatGPT và Gemini để giải đáp thắc mắc.

----------HẾT----------