Báo cáo -Distributed Stochastic Neighbor Embedding (t-SNE)

Họ và tên: Trần Anh Duy

MSSV: 21120235

I. Động lực tìm hiểu t-SNE:

Trong các bài toán học máy thì dữ liệu có kích thước rất lớn. Máy tính có thể hiểu và thực thi các thuật toán trên dữ liệu này, tuy nhiên đối với con người để "nhìn" dữ liệu nhiều chiều thật sự là rất khó. Vì vậy bài toán giảm chiều dữ liệu ra đời giúp đưa ra cái nhìn mới cho con người về dữ liệu nhiều chiều. Ngoài để trực quan dữ liệu, các phương pháp giảm chiều dữ liệu còn giúp đưa dữ liệu về một không gian mới giúp khai phá các thuộc tính ẩn mà trong chiều dữ liệu ban đầu không thể hiện rõ, hoặc đơn giản là giảm kích thước dữ liệu để tăng tốc độ thực thi cho máy tính.

II. Giới thiệu t-SNE:

- **t-SNE (t-distributed Stochastic Neighbor Embedding)** là một kỹ thuật giảm chiều được sử dụng để visualize (hình ảnh hóa) dữ liệu nhiều chiều.
- Nó đặc biệt hữu ích cho việc hiển thị các tập dữ liệu có nhiều hơn ba chiều trong không gian hai hoặc ba chiều, giúp chúng ta dễ dàng quan sát các mẫu và cụm trong dữ liệu.
- Điểm khác biệt của t-SNE so với các phương pháp giảm chiều khác như PCA (Phân tích thành phần chính) là nó chú ý đến **giữ lại cấu trúc cục bộ** của dữ liệu.
- Quá trình tính toán của t-SNE có yếu tố ngẫu nhiên, dẫn đến việc có thể tạo ra các kết quả khác nhau mỗi khi chạy thuật toán trên cùng một dữ liệu. Tuy nhiên, điều này cũng giúp t-SNE thoát khỏi các điểm cực tiểu địa phương, dẫn đến kết quả tốt hơn.

III. Thuật toán t-SNE:

- t-SNE hoạt động bằng cách xây dựng một bảng đồ đa chiều của dữ liệu sau đó giảm số chiều của bản đồ này xuống còn 2 chiều để hiển thị dữ liệu.
- Trong quá trình xây dựng bản đồ đa chiều, t-SNE tối ưu hoá hàm chi phí để đảm bảo rằng các điểm dữ liệu tương đồng sẽ được đặt gần nhau trong bản đồ đa chiều. Có nhiều cách để tính toán độ tương đồng, ví dụ như sử dụng khoảng cách **Euclidean** hoặc **Kullback-Leibler.**
- Sau khi xây dựng bản đồ đa chiều, dựa trên các điểm tương đồng này, t-SNE tạo ra một biểu diễn hai hoặc ba chiều của dữ liệu, cố gắng giữ lại các điểm tương đồng cục bộ. Quá trình này sử dụng một kỹ thuật gọi là **phân bố t-distributed** để mô phỏng các mối quan hệ giữa các điểm dữ liệu trong không gian giảm chiều.

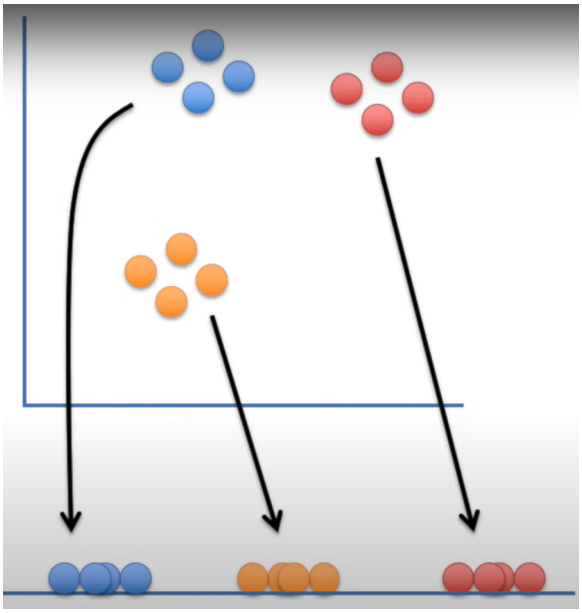
1. Đầu vào:

- Ma trận dữ liệu: Ma trận này chứa các điểm dữ liệu được biểu diễn trong không gian nhiều chiều. Mỗi hàng trong ma trận đại diện cho một điểm dữ liệu và mỗi cột đại diện cho một chiều. Kích thước của ma trận là n x m.
- **Số chiều**: Số chiều mong muốn cho biểu đồ nhúng. Giá trị phổ biến là 2 hoặc 3, nhưng t-SNE cũng có thể được sử dụng với nhiều chiều hơn.
- **Tham số** : Một số tham số khác có thể được điều chỉnh để ảnh hưởng đến kết quả của t-SNE, bao gồm:
    - **Perplexity**: Kiểm soát độ mịn của các cụm trong biểu đồ nhúng. Giá trị perplexity cao hơn sẽ dẫn đến các cụm mịn hơn, trong khi giá trị perplexity thấp hơn sẽ dẫn đến các cụm thô hơn.
    - **Learning rate**: Kiểm soát tốc độ thay đổi vị trí của các điểm dữ liệu trong quá trình tối ưu hóa. Giá trị learning rate cao hơn sẽ dẫn đến sự hội tụ nhanh hơn, nhưng có thể dẫn đến kết quả cục bộ.
    - **Số lần lặp**: Số lần thuật toán t-SNE sẽ được chạy. Giá trị số lần lặp cao hơn sẽ dẫn đến kết quả chính xác hơn, nhưng tốn thời gian tính toán hơn.

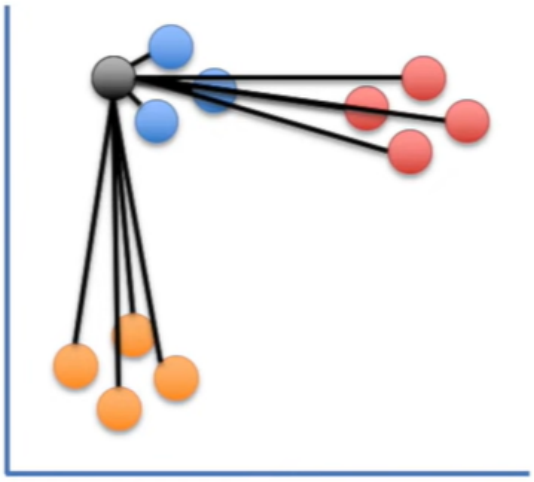
2. Đầu ra:

- **Biểu đồ nhúng**: Biểu đồ hiển thị vị trí của các điểm dữ liệu trong không gian nhúng. Các điểm dữ liệu gần nhau trong không gian ban đầu sẽ có xu hướng nằm gần nhau trong biểu đồ nhúng.
- **Thông tin bổ sung**: Một số thư viện t-SNE có thể cung cấp thông tin bổ sung về kết quả, chẳng hạn như:
    - **Khoảng cách giữa các điểm dữ liệu**: Khoảng cách giữa các điểm dữ liệu trong không gian nhúng.
    - **Mật độ điểm dữ liệu**: Mật độ điểm dữ liệu trong các khu vực khác nhau của biểu đồ nhúng.
    - **Phân bố điểm dữ liệu**: Phân bố điểm dữ liệu theo các chiều khác nhau của biểu đồ nhúng.

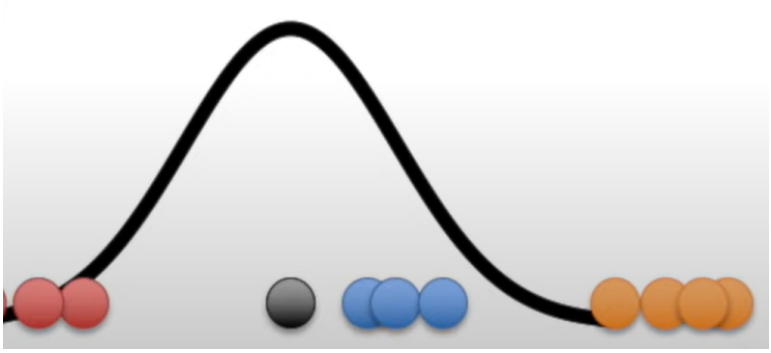
3. Mô tả hình học về t-SNE



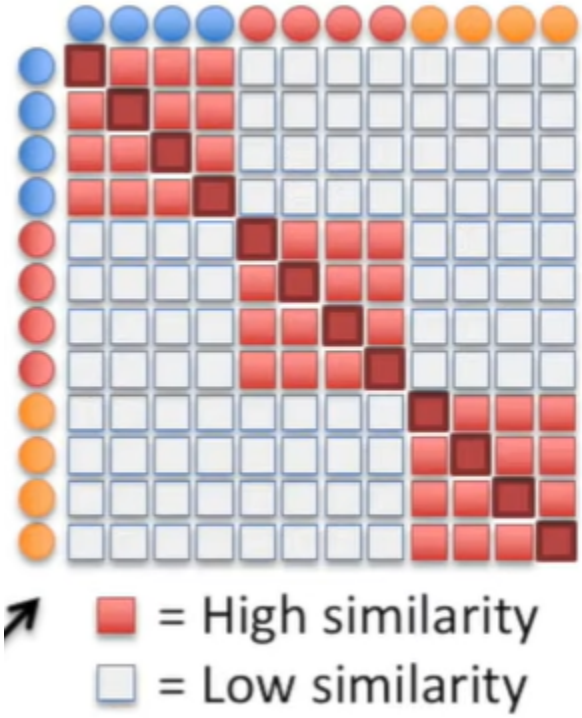
Khái niệm đằng sau t-SNE liên quan đến việc tính toán độ tương tự giữa từng điểm dữ liệu và tất cả các điểm khác trong tập dữ liệu. Tận dụng những điểm tương tự này, thuật toán sẽ giảm tính chiều của dữ liệu trong khi vẫn duy trì mối quan hệ cục bộ (tức là tìm các lân cận gần nhất) giữa các điểm.



Cách để tính toán sự tương tự giữa từng điểm: Thay vì chỉ dựa vào tính toán khoảng cách, t-SNE sử dụng phương pháp xác suất. Đối với mỗi điểm dữ liệu, phân bố Gaussian được vẽ xung quanh nó với giá trị trung bình bằng 0 và độ lệch chuẩn được xác định dựa trên mật độ của các điểm lân cận xung quanh nó.

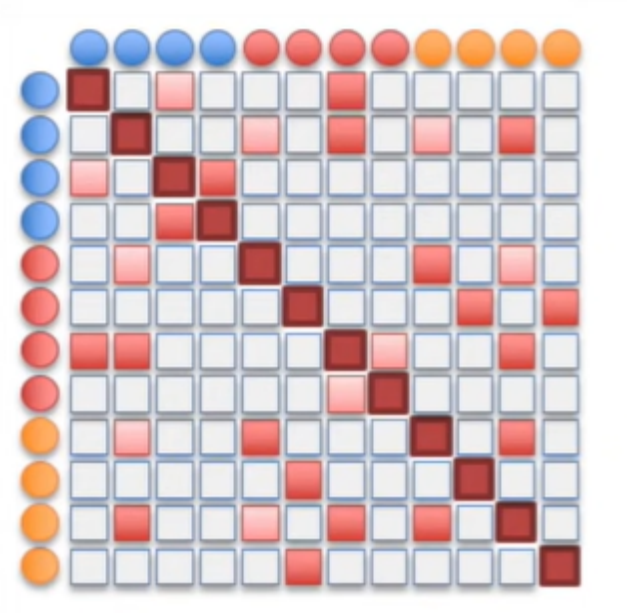


Tương tự, phương pháp này được áp dụng cho từng điểm trong tập dữ liệu, dẫn đến ma trận n * n trong đó điểm tương tự cho mỗi điểm dữ liệu được ghi lại so với mọi điểm dữ liệu khác. Giá trị của P cao hơn giữa hai điểm có nghĩa là chúng là hàng xóm của nhau trong khi giá trị thấp chỉ ra rằng chúng là những điểm không giống nhau.



Bây giờ chúng ta giảm chiều cao hơn thành chiều thấp hơn và các điểm được phân phối ngẫu nhiên trên trục x:

Ở giai đoạn này, ta sẽ tính toán lại điểm tương tự cho mỗi điểm so với các điểm khác, dẫn đến một ma trận n * n n khác.



Bây giờ chúng ta có hai ma trận: một đại diện cho điểm tương đồng ở chiều cao hơn và một đại diện cho điểm tương tự ở chiều thấp hơn.

Để duy trì mối quan hệ cục bộ giữa các điểm, đảm bảo rằng các điểm lân cận vẫn gần nhau ở chiều dưới, ta đặt mục tiêu căn chỉnh ma trận chiều thấp hơn với ma trận chiều cao hơn. Sự liên kết này liên quan đến việc điều chỉnh vị trí của các điểm lặp đi lặp lại cho đến khi ma trận tương tự ở chiều thấp giống với ma trận ở chiều cao hơn càng gần càng tốt.

4. Công thức toán học của t-SNE:

- Đối với mỗi điểm dữ liệu trong không gian chiều cao, chúng ta tính toán sự tương đồng của nó với mọi điểm khác bằng cách sử dụng phân phối Gaussian. Sự giống nhau này dựa trên khoảng cách giữa các điểm.
- Ví dụ ta có một tập dữ liệu X = {X1, X2, ..., XN} với độ tương đồng từng cặp Pij, trong đó Pij biểu thị xác suất có điều kiện chọn Xj làm láng giềng của Xi trong phân phối Gauss tập trung tại X, được định nghĩa như sau:

$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)} \tag{1}$$

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

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

$$q_{j|i} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k \neq i} \left(1 + \|y_i - y_j\|^2\right)^{-1}} \tag{3}$$

$$q_{ij} = \frac{q_{j|i} + q_{i|j}}{2N} \tag{4}$$

IV. Ứng dụng của t-SNE:

- Phân cụm và phân loại: phân cụm các điểm dữ liệu tương tự lại với nhau trong không gian có chiều thấp hơn. Nó cũng có thể được sử dụng để phân loại và tìm kiếm các mẫu trong dữ liệu.
- Phát hiện bất thường: để xác định các ngoại lệ và sự bất thường trong dữ liệu.
- Xử lý ngôn ngữ tự nhiên: để trực quan hóa các từ nhúng được tạo từ một kho văn bản lớn giúp xác định điểm tương đồng và mối quan hệ giữa các từ dễ dàng hơn.
- Bảo mật máy tính: để trực quan hóa các mẫu lưu lượng mạng và phát hiện sự bất thường.
- Nghiên cứu ung thư: để hình dung hồ sơ phân tử của các mẫu khối u và xác định các loại ung thư.
- Giải thích miền địa chất: để hình dung các thuộc tính địa chấn và xác định các dị thường địa chất.
  Xử lý tín hiệu y sinh: để hình dung điện não đồ (EEG) và phát hiện các mô hình hoạt động của não.

V. Vấn đề của t-SNE:

Mặc dù là một công cụ mạnh mẽ để visualize dữ liệu nhiều chiều, t-SNE cũng có một số vấn đề cần lưu ý:

- Kết quả của t-SNE có thể phụ thuộc đáng kể vào các tham số được sử dụng, bao gồm số chiều, perplexity và learning rate. Việc lựa chọn tham số không

phù hợp có thể dẫn đến biểu đồ nhúng không chính xác hoặc khó diễn giải.

- Biểu đồ nhúng t-SNE có thể khó diễn giải hơn so với các phương pháp giảm chiều khác như PCA. Điều này là do t-SNE là một thuật toán phi tuyến tính và không bảo tồn các mối quan hệ tuyến tính trong dữ liệu ban đầu.
- Quá trình tính toán của t-SNE có thể tốn thời gian, đặc biệt cho các tập dữ liệu lớn. Điều này là do t-SNE là một thuật toán lặp và cần nhiều lần tối ưu hóa vị trí của các điểm dữ liệu.
- t-SNE có thể tạo ra các nhiễu và biến dạng trong biểu đồ nhúng. Điều này có thể che khuất các cấu trúc thực sự trong dữ liệu và gây khó khăn cho việc giải thích kết quả.

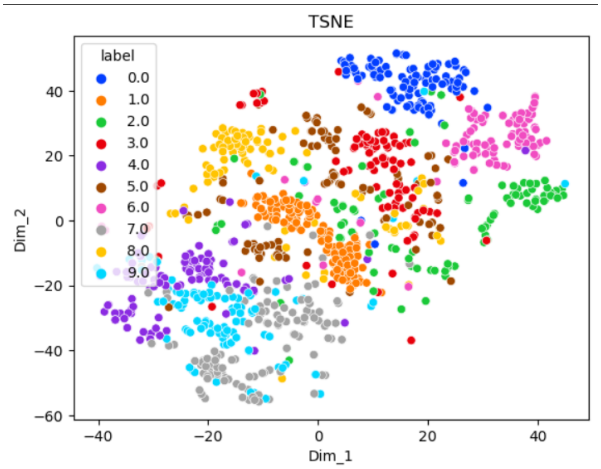
VI. So sánh t-SNE với PCA

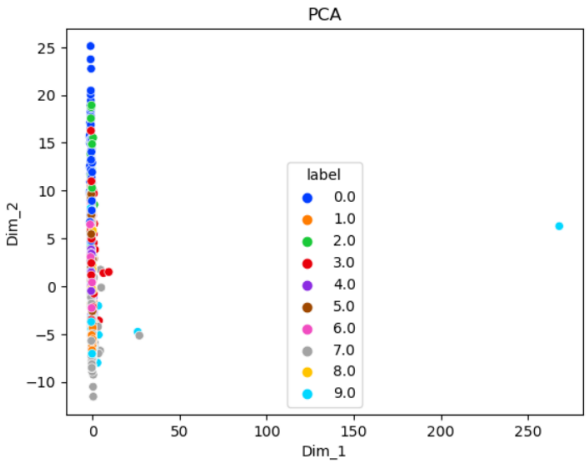
Cả t-SNE và PCA đều là các kỹ thuật giảm kích thước có các cơ chế khác nhau và hoạt động tốt nhất với các loại dữ liệu khác nhau.

- PCA (Phân tích thành phần chính) là một kỹ thuật tuyến tính hoạt động tốt nhất với dữ liệu có cấu trúc tuyến tính. Nó tìm cách xác định các thành phần chính cơ bản trong dữ liệu bằng cách chiếu lên các kích thước thấp hơn, giảm thiểu phương sai và duy trì khoảng cách lớn theo cặp.
- t-SNE là một kỹ thuật phi tuyến tập trung vào việc duy trì sự tương đồng theo cặp giữa các điểm dữ liệu trong không gian có chiều thấp hơn. t-SNE quan tâm đến việc duy trì khoảng cách nhỏ theo cặp trong khi PCA tập trung vào việc duy trì khoảng cách lớn theo cặp để tối đa hóa phương sai.

Tóm lại, PCA bảo toàn phương sai trong dữ liệu, trong khi t-SNE bảo toàn mối quan hệ giữa các điểm dữ liệu trong không gian có chiều thấp hơn, khiến nó trở thành một thuật toán khá tốt để hiển thị dữ liệu chiều cao phức tạp.

Với 1 bộ dataset MNIST, 2 kỹ thuật đưa ra những kết quả khác nhau. Dễ dàng ta thấy được t-SNE có hiệu quả hơn so với PCA ở bộ dữ liệu này.





VII. Đánh giá báo cáo:

Tiêu chí đánh giá	Mức độ hoàn thành
Tìm hiểu t-SNE	100%
Cài đặt t-SNE	100%
So sánh PCA và t-SNE	100%

VIII. Tài liệu tham khảo:

t-SNE from Scratch (ft. NumPy). Acquire a deep understanding of the... | by Jacob Pieniazek | Towards Data Science

https://www.youtube.com/watch?v=_CzYVI8axao

https://www.geeksforgeeks.org/ml-t-distributed-stochastic-neighbor-embedding-t-sne-algorithm/Python Tutorial: t-SNE visualization of high-dimensional data (youtube.com)