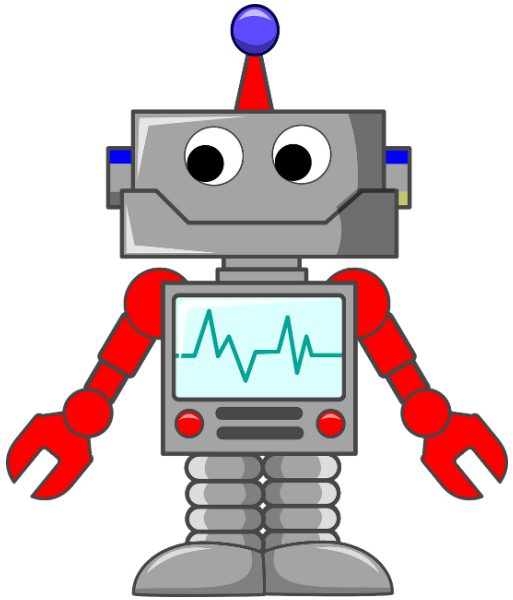
ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

CS116 – LẬP TRÌNH PYTHON CHO MÁY HỌC

BÀI 06

HỌC KHÔNG GIÁM SÁT - UNSUPERVISED LEARNING



TS. Nguyễn Vinh Tiệp

NỘI DUNG

1. GIỚI THIỆU HỌC KHÔNG GIÁM SÁT

2. CÁC MÔ HÌNH GOM NHÓM - CLUSTERING

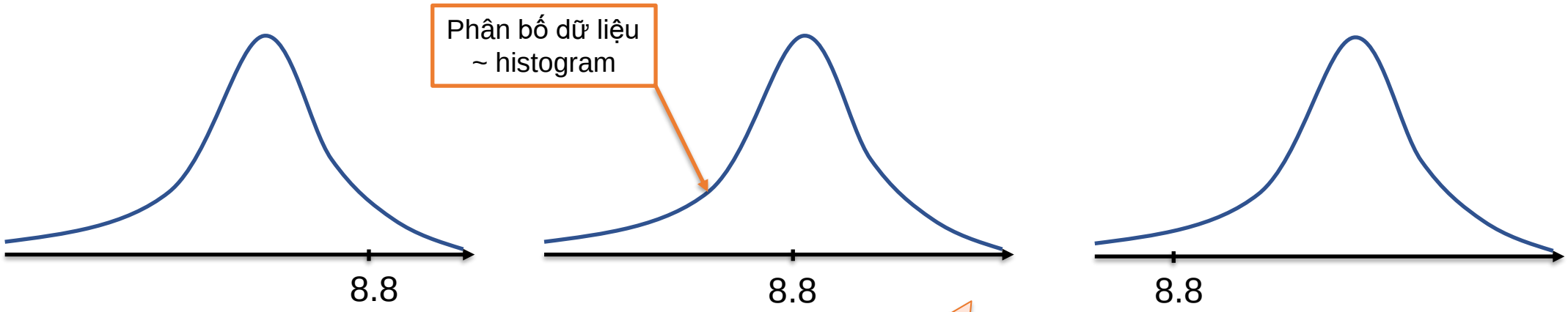
3. CÁC MÔ HÌNH GIẢM CHIỀU DỮ LIỆU

Giới thiệu

- **Học không giám sát (unsupervised learning)**: là một nhánh của ML, có nhiệm vụ học phân bố của dữ liệu, từ đó có thể biểu diễn dữ liệu hiệu quả hơn
- Dữ liệu cho thuật toán học không giám sát không cần gán nhãn (label)
  - Chỉ cần dữ liệu đầu vào x
  - Không cần nhãn đầu ra tương ứng
- Một số chủ đề chính:
  - Gom nhóm dữ liệu
  - Giảm chiều dữ liệu

Phân bố của dữ liệu

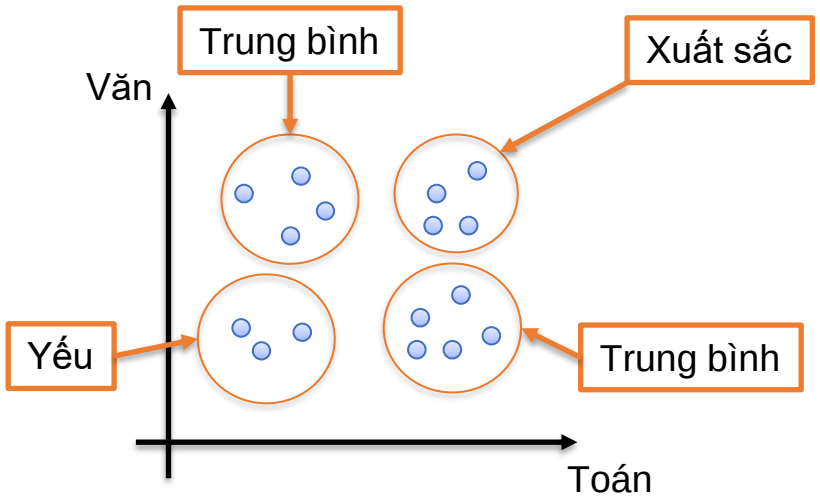
- Ví dụ: An có điểm TB là 8.8. Hỏi An xếp loại giỏi, khá hay trung bình?



Không thể biết xếp loại nếu không biết phân bố dữ liệu

Biểu diễn dữ liệu

- Ví dụ: Cho điểm hai môn toán, văn của các bạn trong lớp. Phân loại học lực từng bạn?



Thay vì lưu dữ liệu "thô"
➔ chỉ cần lưu "đặc trưng" học lực

- Giảm chiều dữ liệu (VD: giảm 50% số thuộc tính)

- Thể hiện được đặc trưng theo nhóm của mẫu dữ liệu

Lưu ý: ở đây chỉ mượn các khái niệm trong cuộc sống ("yếu", "xuất sắc", "trung bình"). Thực tế, thuật toán UL sẽ mã hóa bằng các cluster ID nào đó!

NỘI DUNG

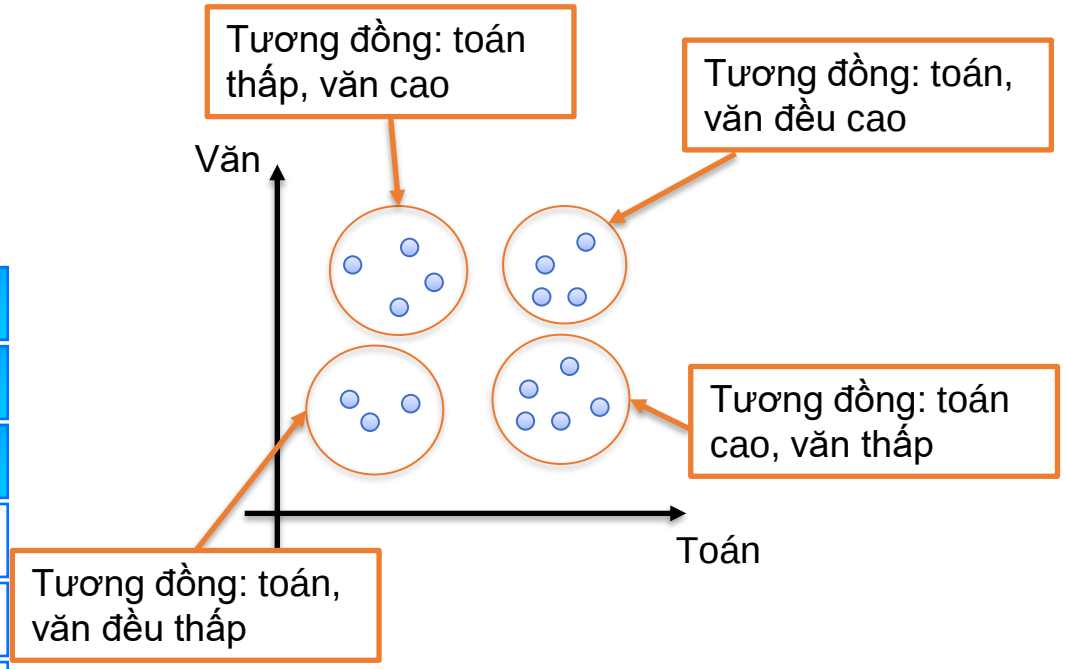
1. GIỚI THIỆU HỌC KHÔNG GIÁM SÁT

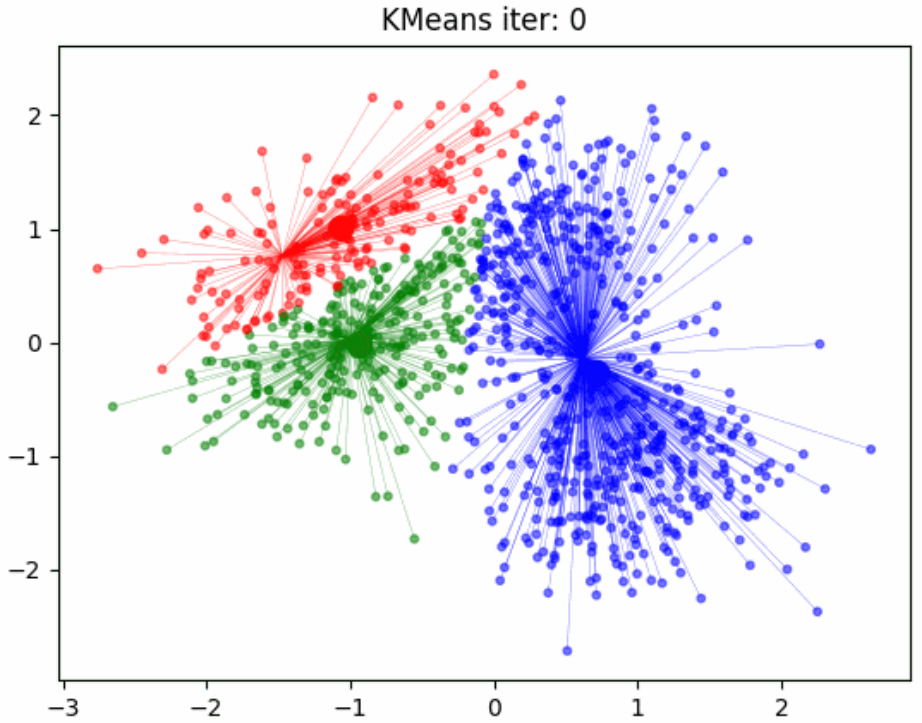
2. CÁC MÔ HÌNH GOM NHÓM - CLUSTERING

3. CÁC MÔ HÌNH GIẢM CHIỀU DỮ LIỆU

Bài toán 1: Gom nhóm dữ liệu

- **Gom nhóm (clustering):** là bài toán gom các đối tượng theo từng cụm sao cho các đối tượng trong cùng một cụm có sự tương đồng với nhau hơn so với những đối tượng thuộc các nhóm khác





Ví dụ gom cụm với thuật toán K-Mean

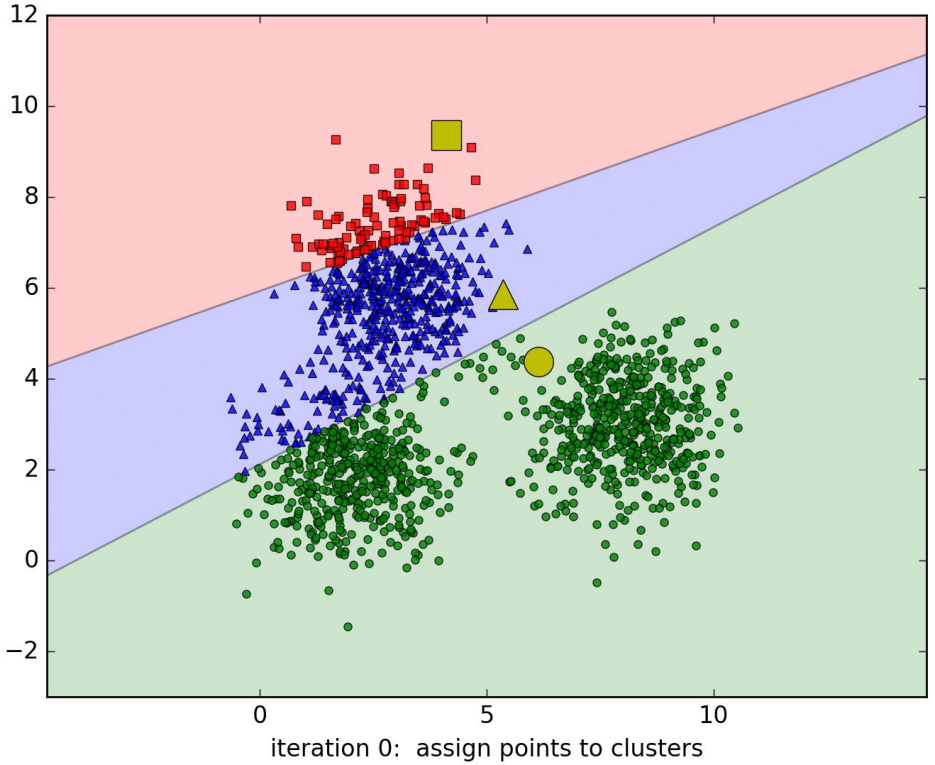
Bài toán 1: Gom nhóm dữ liệu

- Một số thuật toán gom nhóm dữ liệu:
  - K-Means
  - DBSCAN
  - Hierarchical clustering
  - Gaussian Mixture Models (GMM)
  - Spectral clustering

- Mỗi phương pháp có những ưu – khuyết điểm riêng. Sử dụng phương pháp nào tùy vào tính chất dữ liệu và mục tiêu cụ thể

Thuật toán gom nhóm - KMeans

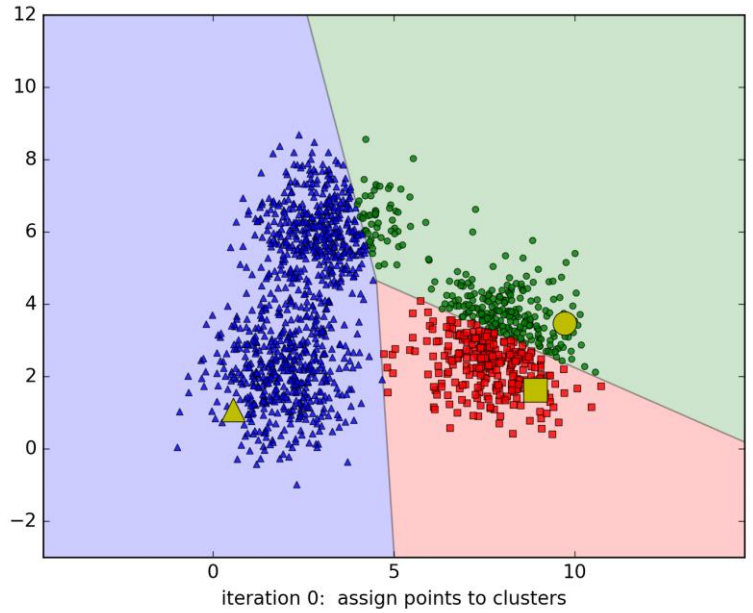
- **Ý tưởng**: khởi tạo ngẫu nhiên K tâm cụm, sau đó gán các điểm dữ liệu vào trọng tâm gần nhất. Quá trình này được lặp lại cho đến khi các trọng tâm không thay đổi nữa

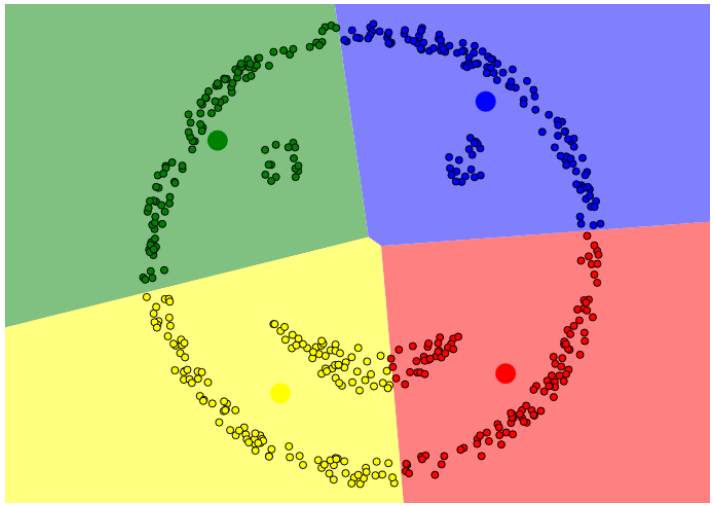


Nguồn: https://machinelearningcoban.com/2017/01/01/kmeans/

Thuật toán gom nhóm - KMeans

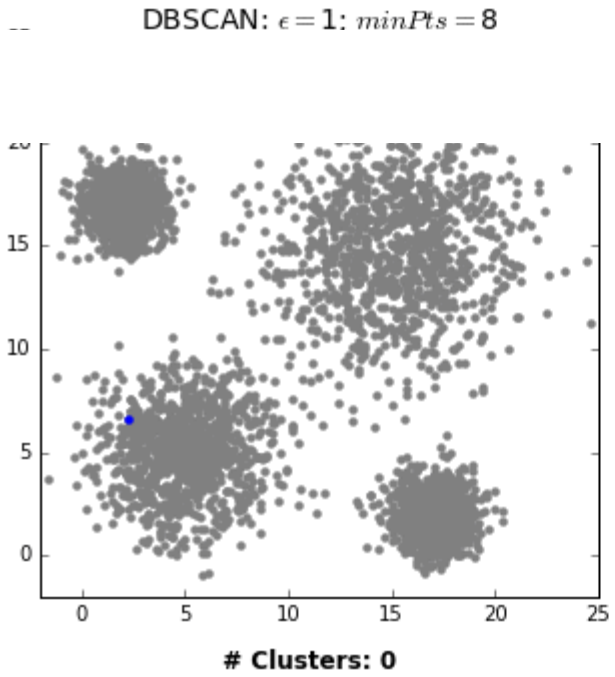
- **Ưu điểm**:
  - Đơn giản, dễ cài đặt
  - Hiệu quả với dữ liệu lớn
- **Khuyết điểm**:
  - Cần biết trước số lượng cụm K
  - Dễ bị rơi vào cực tiểu cục bộ
  - Phụ thuộc vào tâm cụm khởi tạo
  - Không hoạt động tốt với dữ liệu có phân bố phức tạp, không phải dạng hình cầu

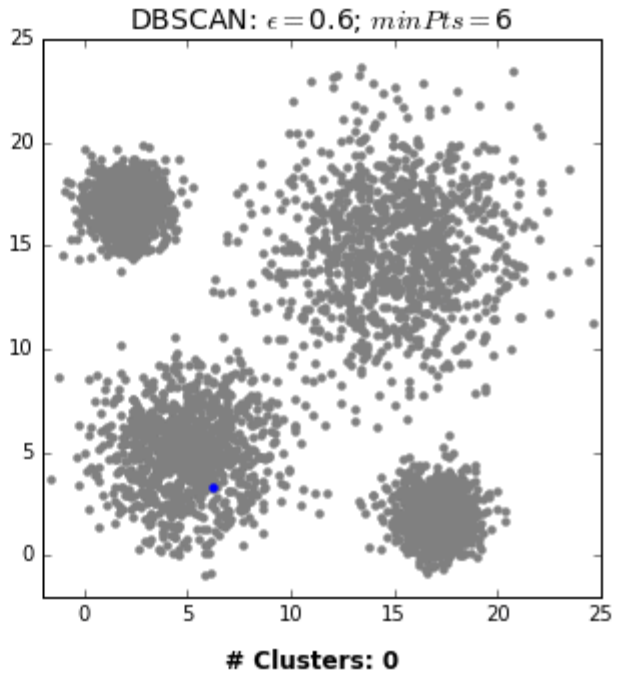




Thuật toán gom nhóm - DBSCAN

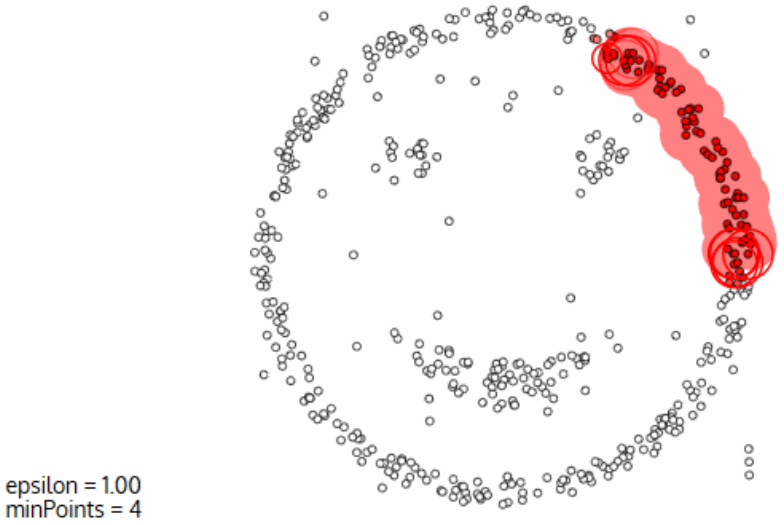
- **Ý tưởng**: phân cụm dựa trên mật độ, gom các điểm gần nhau (có khoảng cách nhỏ hơn ε) và có mật độ cao (số điểm tối thiểu trong cụm là $minPts$)
- Các điểm nằm trong các cụm mật độ thấp được gán là nhiễu (noise)



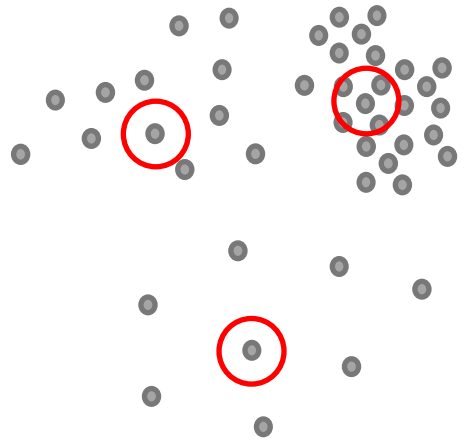


Thuật toán gom nhóm - DBSCAN

- **Ưu điểm**:
  - Không cần biết trước số cụm
  - Hiệu quả với dữ liệu mật độ cao
- **Khuyết điểm**:
  - Không hiệu quả khi dữ liệu có mật độ biến động
  - Phải chọn tham số ε và $minPts$



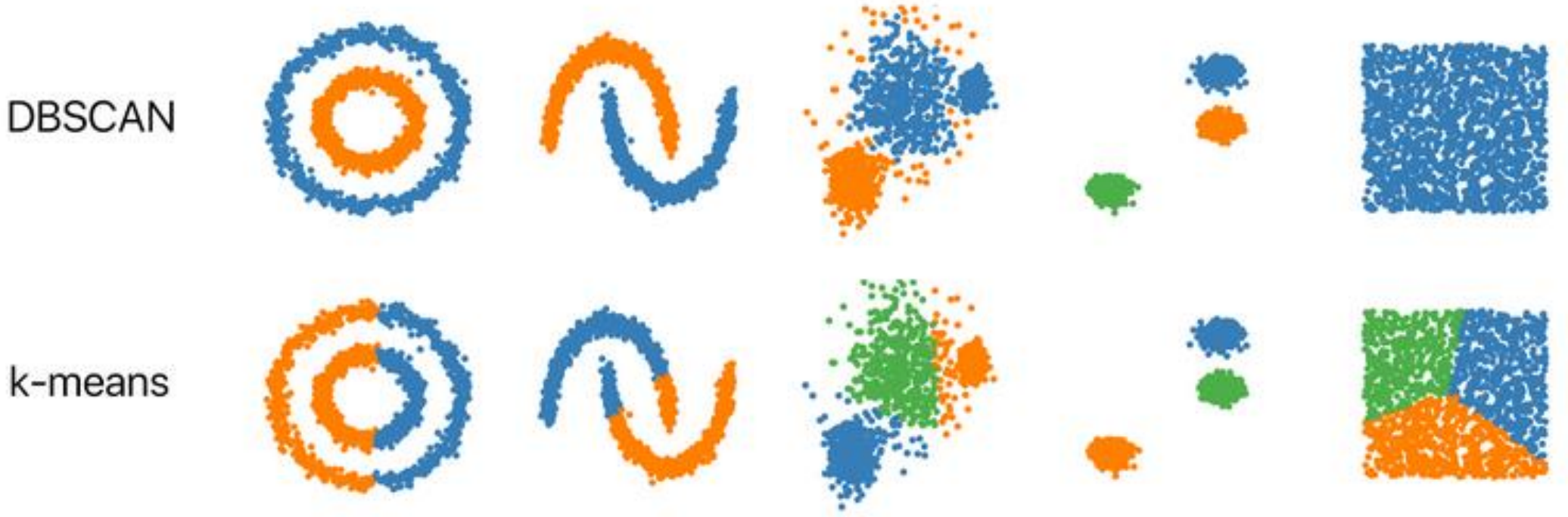
Nguồn: https://www.digitalvidya.com/blog/the-top-5-clustering-algorithms-data-scientists-should-know/



Tham số ε cố định không hiệu quả khi số điểm giảm

So sánh K-Means và DBSCAN

- So sánh K-Means với DBSCAN khi gom nhóm trên các “toy example”:



Nguồn: scikit-learn.org

NỘI DUNG

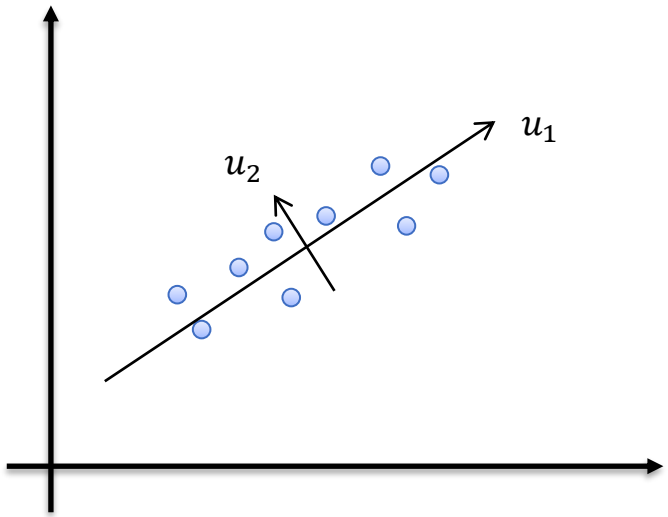
1. GIỚI THIỆU HỌC KHÔNG GIÁM SÁT

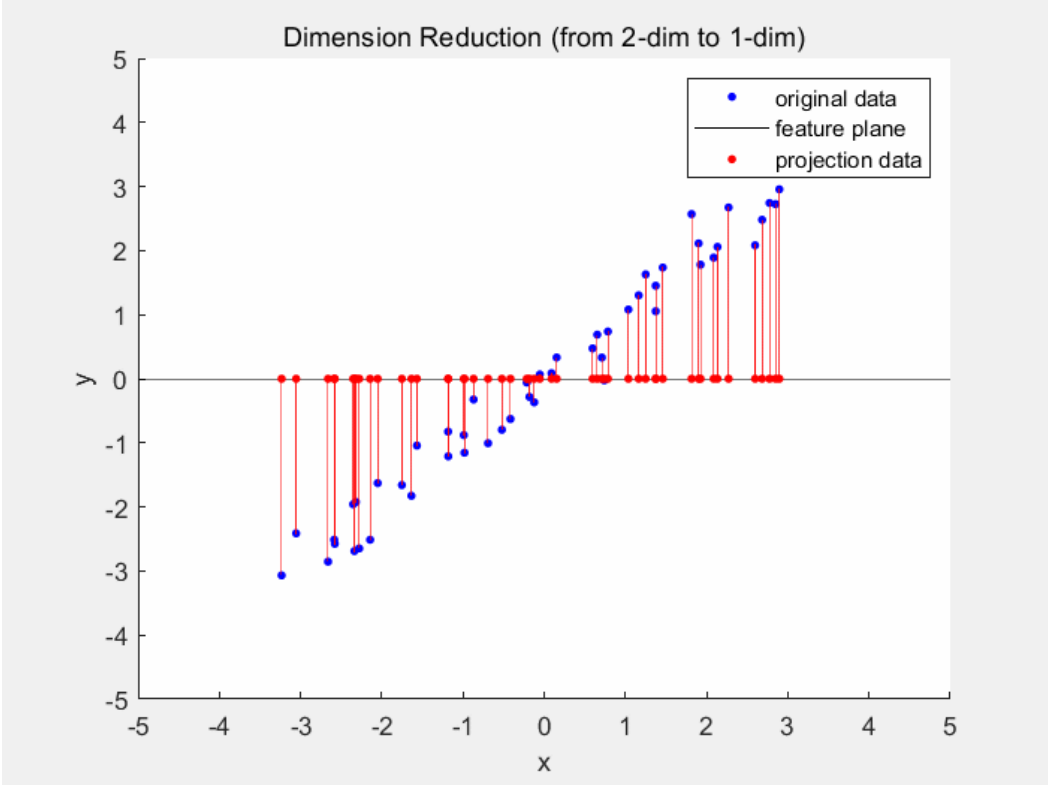
2. CÁC MÔ HÌNH GOM NHÓM - CLUSTERING

3. CÁC MÔ HÌNH GIẢM CHIỀU DỮ LIỆU

Bài toán 2: Giảm chiều dữ liệu

- **Giảm chiều dữ liệu**: là quá trình chuyển đổi dữ liệu từ không gian đa chiều sang không gian ít chiều sao cho biểu diễn không gian ít chiều vẫn giữ được một số tính chất có ý nghĩa của dữ liệu gốc





Trực quan hóa thuật toán PCA

Nguồn: https://www.programmersought.com/article/63044502718/

Bài toán 2: Giảm chiều dữ liệu

- Một số thuật toán giảm chiều dữ liệu:
  - Phân tích thành phần chính – PCA
  - Nhúng t-SNE

- Mỗi phương pháp có những đặc điểm và ứng dụng riêng, tùy thuộc vào bộ dữ liệu và mục tiêu cụ thể

Thuật toán Principal Component Analysis - PCA

- **Ý tưởng**: Cho bảng dữ liệu điểm của một lớp học như sau, bạn có nhận xét gì?

Giảm được $\approx \frac{1}{4} = 25\%$ dữ liệu

Họ tên	Giới tính	Toán	Văn
Họ tên 1	Nam	6.5	7
Họ tên 2	Nữ	7.5	7
Họ tên 3	Nam	9.0	7
Họ tên 4	Nữ	6.0	7
Họ tên 5	Nữ	9.5	7

Độ lệch bằng 0 nên có thể xóa cột này, **chỉ cần lưu điểm chung là 7**

Thuật toán Principal Component Analysis - PCA

- **Ý tưởng**: Cho bảng dữ liệu điểm của một lớp học như sau, bạn có nhận xét gì? Ví dụ khác.

Họ tên	Giới tính	Toán	Văn
Họ tên 1	Nam	6.5	7.0
Họ tên 2	Nữ	7.5	6.5
Họ tên 3	Nam	9.0	7.0
Họ tên 4	Nữ	6.0	7.0
Họ tên 5	Nữ	9.5	7.5

Độ lệch bằng ~0.32 nên vẫn có thể xóa cột này, **chỉ lưu điểm TB là 7**

Thuật toán Principal Component Analysis - PCA

- **Ý tưởng**: Cho bảng dữ liệu điểm của một lớp học như sau, bạn có nhận xét gì? Ví dụ khác.

Vẫn giảm được $\approx \frac{1}{4} = 25\%$ dữ liệu

Nhưng không khôi phục được dữ liệu gốc

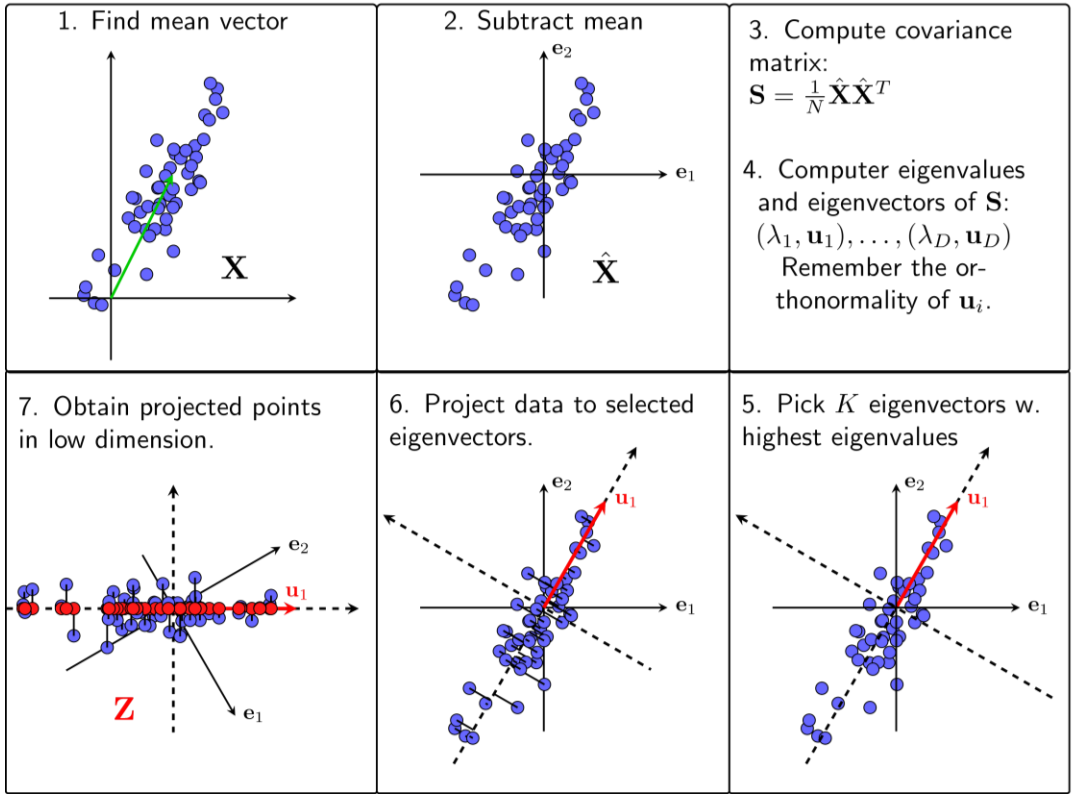
Họ tên	Giới tính	Toán	Văn
Họ tên 1	Nam	6.5	7.0. **7**
Họ tên 2	Nữ	7.5	6.5
Họ tên 3	Nam	9.0	
Họ tên 4	Nữ	6.0	
Họ tên 5	Nữ	9.5	

Chiều dữ liệu nào ít biến động →có thể loại bỏ

Độ lệch bằng ~0.32 nên vẫn có thể xóa cột này, **chỉ lưu điểm TB là 7**

Thuật toán Principal Component Analysis - PCA

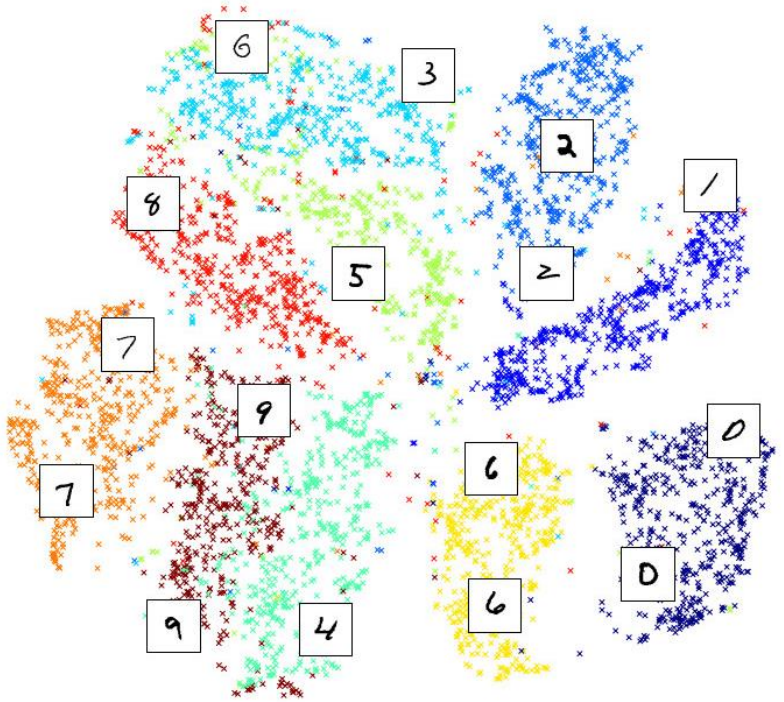
- PCA tìm một không gian con tuyến tính mới mà dữ liệu được biểu diễn một cách hiệu quả nhất
- Trong không gian mới này, các chiều được chọn sao cho tối đa hóa phương sai của dữ liệu



Nguồn: https://machinelearningcoban.com/2017/06/15/pca/

tSNE (t-Distributed Stochastic Neighbor Embedding)

- **t-SNE**: là kỹ thuật giảm chiều phi tuyến tính, để trực quan hóa dữ liệu đa chiều trong không gian có số chiều thấp hơn (thường là 2D hoặc 3D)
- **Ý tưởng**: tạo ra một phân phối xác suất tương tự trong không gian có số chiều thấp hơn

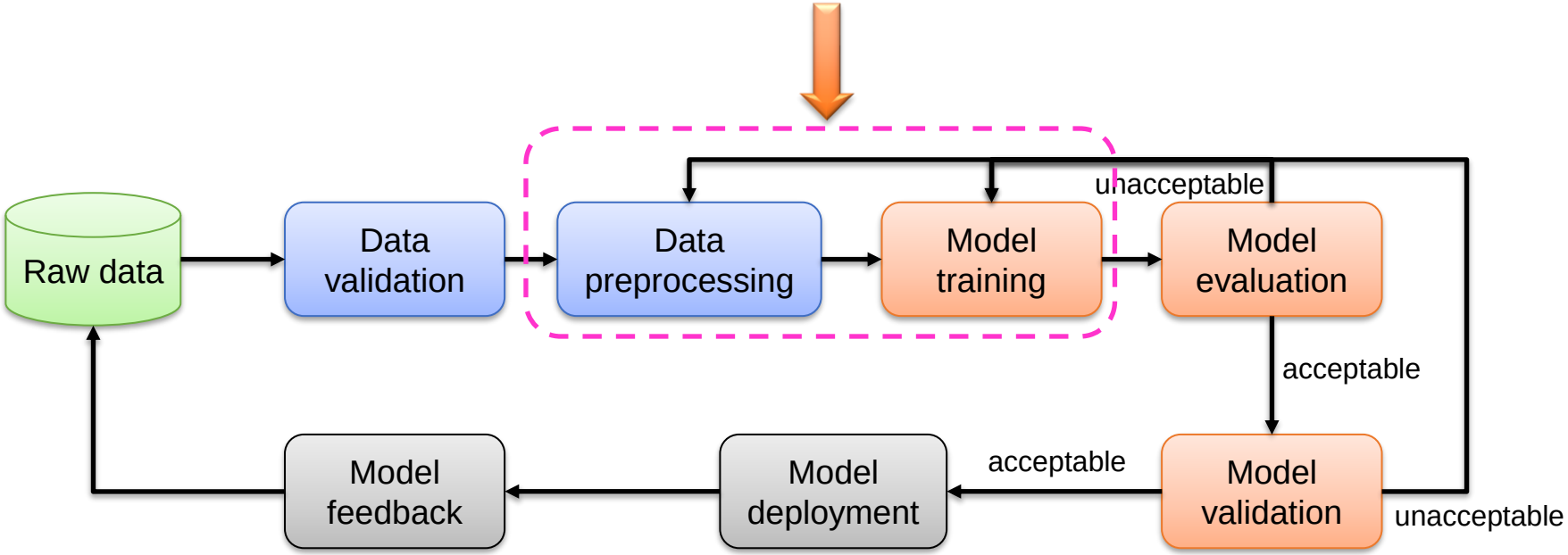


Trực quan hóa tập MNIST sử dụng tSNE

So sánh PCA và tSNE

	PCA	tSNE
Ưu điểm	- Nhanh, hiệu quả với dữ liệu lớn do thực hiện biến đổi tuyến tính - Giữ lại các thành phần chính với phương sai lớn	- Có khả năng bắt cấu trúc phi tuyến - Thường được sử dụng để trực quan hóa dữ liệu
Khuyết điểm	- Khả năng bắt cấu trúc phi tuyến kém - Dễ bị ảnh hưởng bởi nhiễu (outlier)	- Không ổn định do yếu tố ngẫu nhiên - Độ phức tạp cao - Khó giải thích hơn so với PCA

Tổng kết – Vị trí của bài học



BÀI QUIZ VÀ HỎI ĐÁP