Trường Đại học Công nghệ Thông tin – ĐHQG.HCM
Khoa Khoa học Máy tính

CS313.N21 – Khai thác dữ liệu và ứng dụng

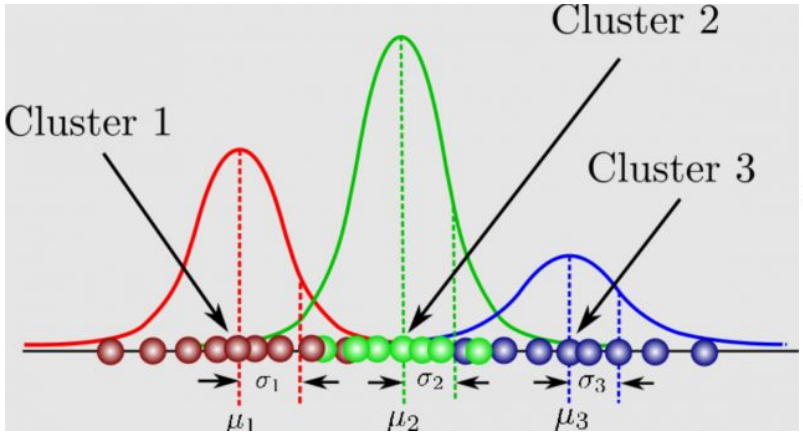
Thuật toán gom nhóm và đánh giá chất lượng gom nhóm

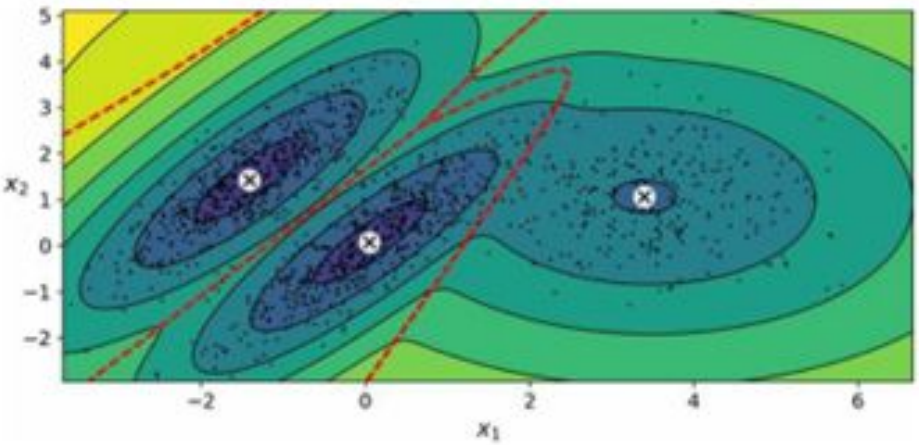
Nhóm 4
Nguyễn Quốc Huy Hoàng – 20520051
Lê Đoàn Phúc Minh – 20520243
Nguyễn Duy Đạt – 20520435
Huỳnh Hoàng Vũ – 20520867

Gaussian Mixture Model (GMM)

Gaussian Mixture Model

- GMM là mô hình phân cụm thuộc dạng unsupervised learning
- GMM giả định phân phối xác suất của mỗi cụm là phân phối Gaussian đa chiều
- Mỗi elip bên dưới tượng trưng như một cluster





Maximum Likelihood

- Ước lượng hợp lí tối đa:

+ θ* là nghiệm tối ưu sao cho giả thiết mô hình GMM khớp nhất với bộ dữ liệu

+ Giải bằng phương pháp EM (Expectation Maximization) để cập nhật dần dần nghiệm θ. MLE bất khả thi trong trường hợp nhiều cụm

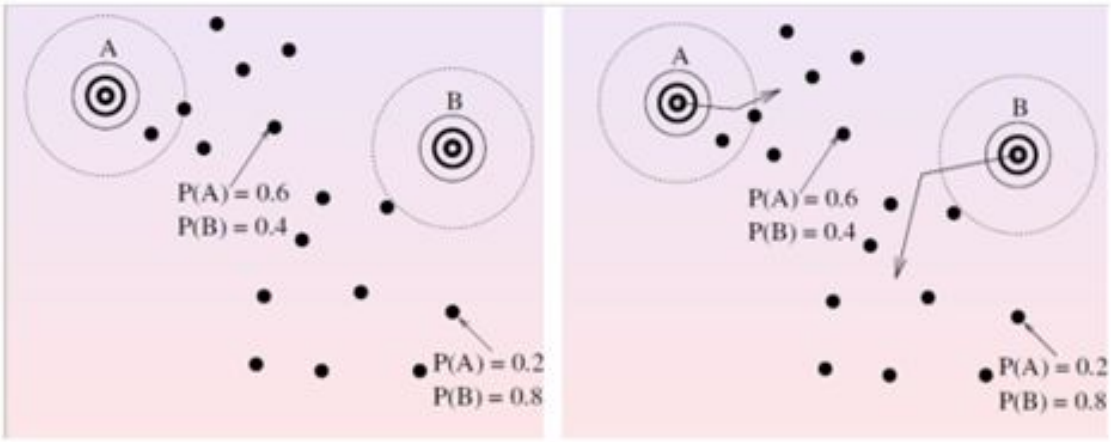
$$\theta^* = \arg\max_{\theta} p(\mathbf{X}|\theta) = \arg\max_{\theta} \prod_{i=1}^{N} p(\mathbf{x}_i|\theta)$$

Maximum Likelihood

*Trong EM, mỗi vòng lặp gồm 2 bước là E-step và M-step:

+ E-step: ước lượng phân phối biến ẩn z thể hiện phân phối xác suất của các cluster tương ứng với data và parameters.

+ M-step: Cực đại hóa phân phối xác suất đồng thời (join distribution probability) của data và biến ẩn z.



Bayesian Information Criteria

* Đánh giá chất lượng mô hình GMM dựa trên chỉ số BIC (Bayesian Information Criteria). Chỉ số này đo lường mức độ hợp lí của model với một bộ tham số được tính dựa trên giá trị tối đa của hàm hợp lí sau:

$$\mathrm{BIC} = k \ln(n) - 2 \ln(\hat{L})$$

+ k là số lượng tham số được ước lượng từ model

+ n là số lượng quan sát của bộ dữ liệu

+ L là giá trị ước lượng tối đa của hàm hợp lí

-> BIC càng nhỏ thì mức độ hợp lí của model đối với bộ dữ liệu càng cao

Ứng dụng

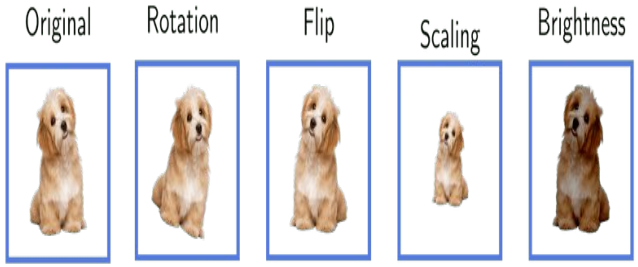
* Ứng dụng của GMM:

+ Gom nhóm khách hàng

+ Phát hiện bất thường, nhận diện lỗi sai sản phẩm

+ Dự đoán giá chứng khoán

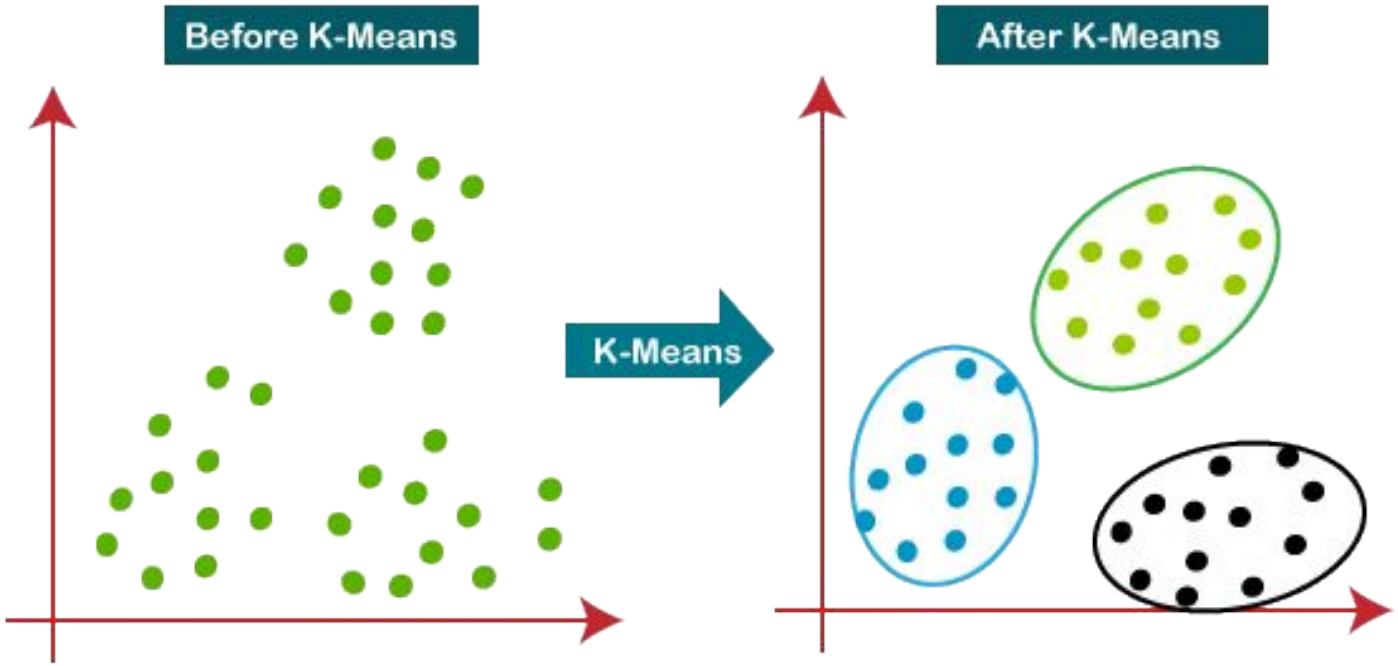
+ Dùng cho tác vụ data augmentation vì tính chất sinh dữ liệu



K-Means

K-means

- Thuật toán K-means là một trong những thuật toán phân cụm dữ liệu phổ biến nhất.

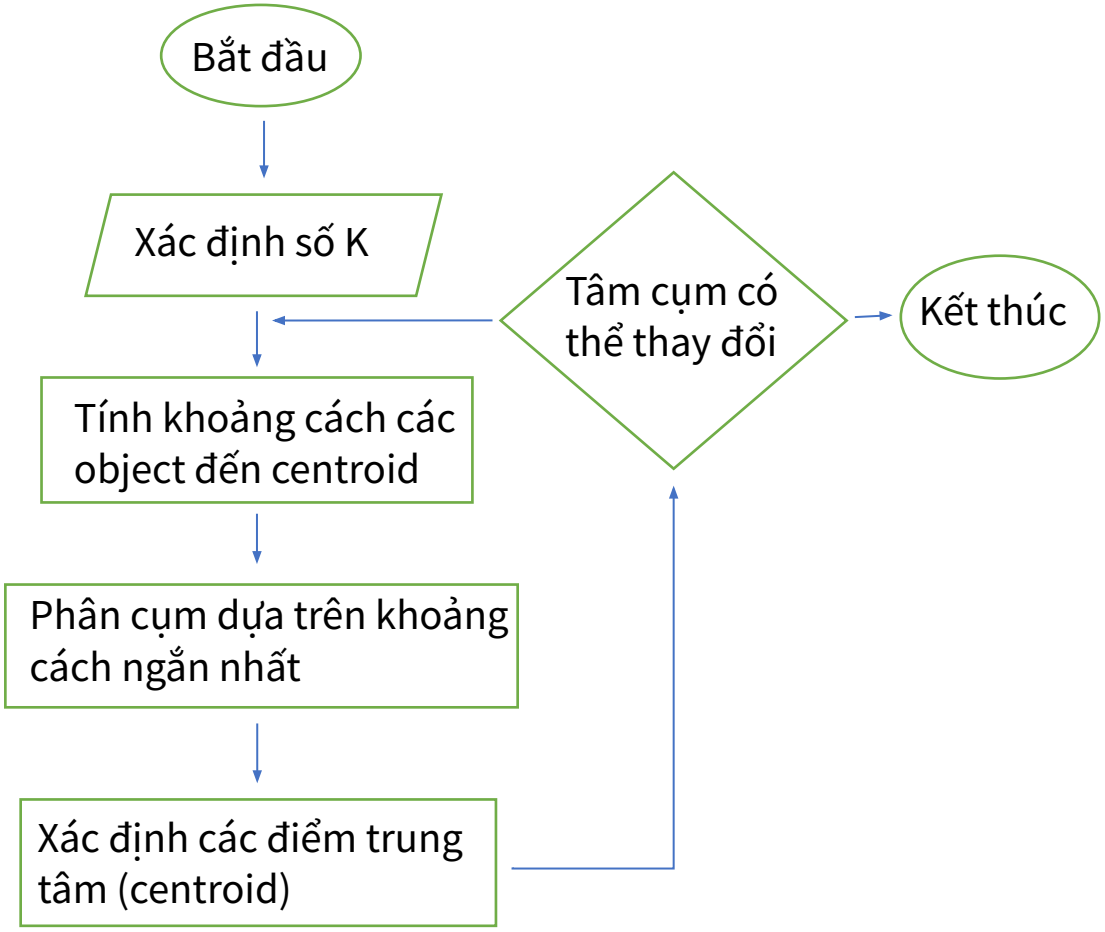


Source: javatpoint.com

K-means

•Thuật toán K-Means thực hiện qua các bước chính sau:

1. Chọn ngẫu nhiên K tâm (centroid) cho K cụm (cluster). Mỗi cụm được đại diện bằng các tâm của cụm.
2. Tính khoảng cách giữa các đối tượng (objects) đến K tâm (thường dùng khoảng cách Euclidean).
3. Nhóm các đối tượng vào cụm có tâm gần nó nhất.
4. Cập nhật lại tâm của các cụm mới được hình thành.
5. Lặp lại bước 2-4 cho đến khi không có sự thay đổi nào trong việc phân nhóm các đối tượng
6. Trả về kết quả



Ứng dụng của thuật toán K-means:

- Phân tích hành vi khách hàng:
- Xử lý ảnh và video
- Phân tích dữ liệu y tế
- Phân nhóm các quan sát trong khoa học
- Phân nhóm sản phẩm

K-means

Một số ưu điểm và nhược điểm của thuật toán K-means:

Ưu điểm:

1.Tốc độ tính toán nhanh.

2.Đơn giản và dễ hiểu.

3.Hiệu quả cho các cụm dạng cầu.

4.Dễ mở rộng và áp dụng.

Nhược điểm:

1.Phụ thuộc vào khởi tạo ban đầu.

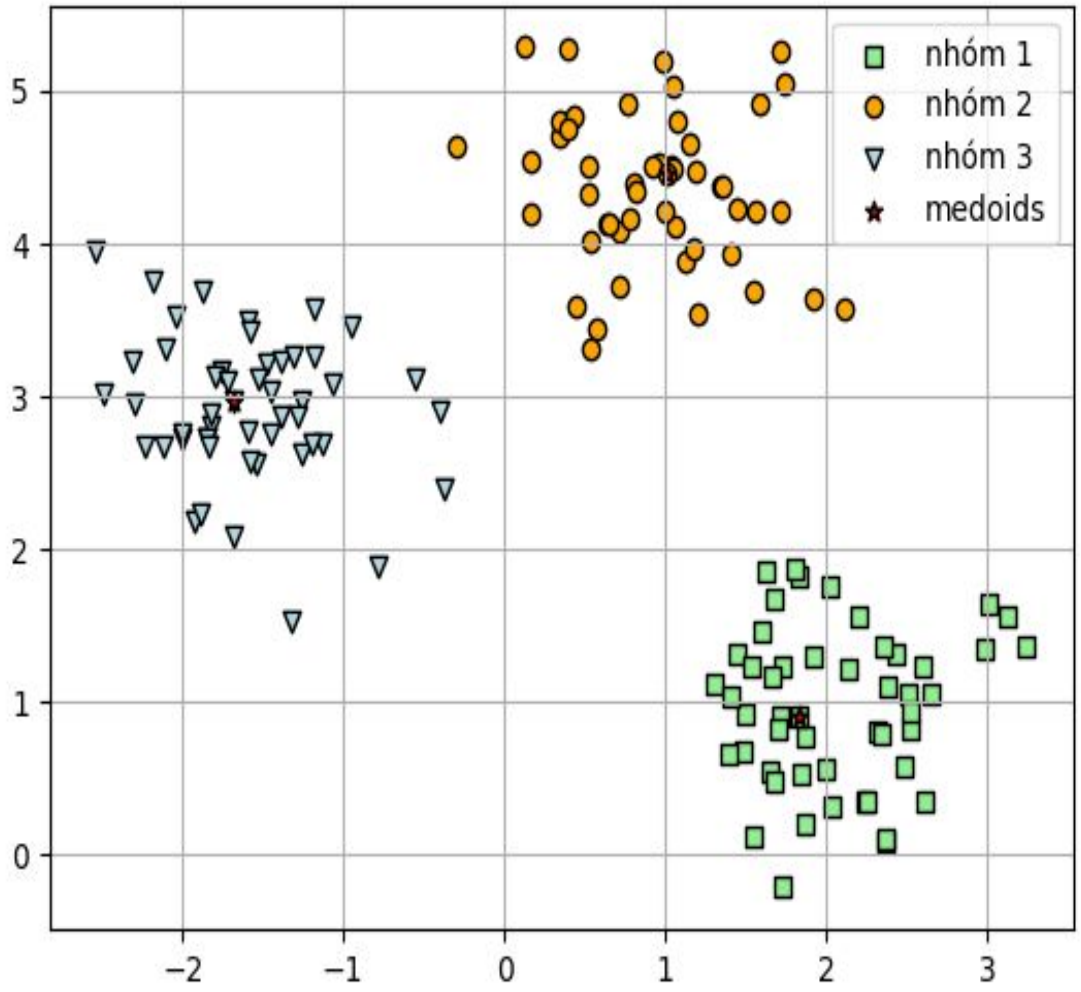
2.Nhạy cảm với nhiễu.

3.Không phân biệt cỡ của các cụm.

4.Khó khăn trong việc chọn số cụm.

K-Medoids

K-Medoids



Ý tưởng chính của K-Medoids

- Mục tiêu là tìm k medoids sao cho tối thiểu hóa sự bất tương đồng (hay tổng chi phí) của việc phân cụm

- Chi phí này được tính bằng tổng khoảng cách của tất cả các điểm dữ liệu đến medoid gần nhất với chúng

- Liên tục lặp lại việc hoán đổi một medoid với một điểm không phải medoid nếu việc hoán đổi giảm chi phí

K-Medoids Pseudocode

- Một trong những thuật toán phổ biến nhất cho phân cụm K-Medoids là Partitioning Around Medoids (PAM), được đề xuất bởi Leonard Kaufman và Peter J. Rousseeuw vào năm 1987.

K-Medoids Pseudocode

```
Đầu vào: tập hợp dữ liệu D, số lượng cụm k
Đầu ra: k cụm với các medoid

M <- Chọn ngẫu nhiên k điểm dữ liệu từ D xem như các medoid
Gán các điểm dữ liệu trong D vào medoid gần chúng nhất
Tính toán chi phí ban đầu C

Lặp lại
  Với mỗi medoid m trong M
    Với mỗi non-medoid o trong D
      Hoán đổi m và o
      Gán các điểm dữ liệu trong D vào medoid gần chúng nhất
      Tính toán chi phí mới C'
      Nếu C' < C
        Chấp nhận thay đổi, C <- C'
      không thì
        Từ chối thay đổi, hoán đổi lại m và o
cho đến khi C không còn thay đổi

Trả về k cụm với các medoid
```

K-Medoids vs K-Means

K-Means	K-Medoids
Chọn điểm trung bình của mỗi cụm làm trung tâm -> Nhạy cảm với nhiễu và ngoại lệ -> Chạy nhanh, có khả năng mở rộng	Chọn một điểm thực tế trong cụm làm trung tâm -> Chịu được nhiễu và ngoại lệ -> Kém khả năng mở rộng
Thường dùng khoảng cách Euclid -> Phù hợp với cụm có hình dạng cầu	Có thể sử dụng bất kỳ phép đo khoảng cách nào -> Phù hợp với các loại dữ liệu khác nhau

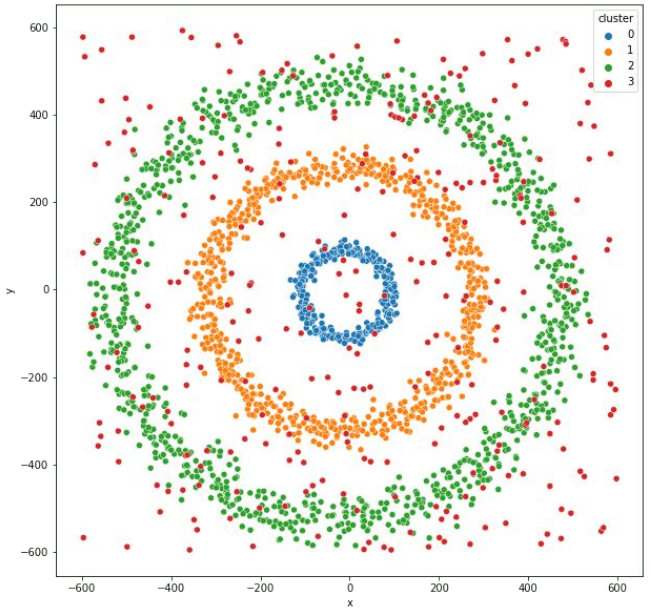
Ứng dụng của K-Medoids

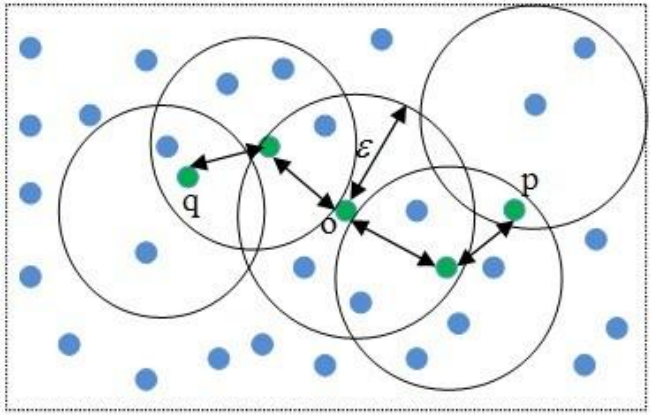
- Phân đoạn ảnh: chia một ảnh thành các vùng có giá trị điểm ảnh hoặc đặc trưng tương tự nhau.
- Phân cụm tài liệu: nhóm các tài liệu hoặc từ có chủ đề hoặc ý nghĩa tương tự nhau.
- Sinh học thông tin (Bioinformatics): phân cụm các gen hoặc protein có chức năng hoặc biểu hiện tương tự nhau.
- Tiếp thị: phân khúc khách hàng dựa trên sở thích hoặc hành vi của họ

DBSCAN

DBSCAN

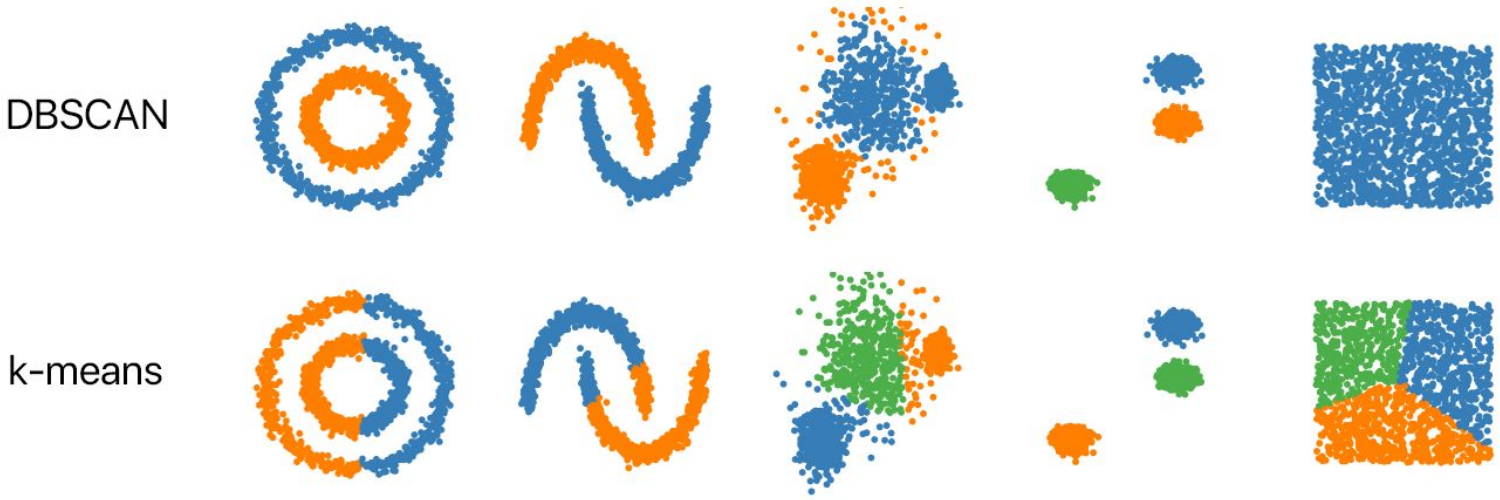
DBSCAN (Density-Based Spatial Clustering of Applications with Noise) là thuật toán gom nhóm hoạt động dựa trên giả thuyết các nhóm, cụm là các vùng dày đặc trong không gian được phân tách bằng các vùng có mật độ thấp hơn.





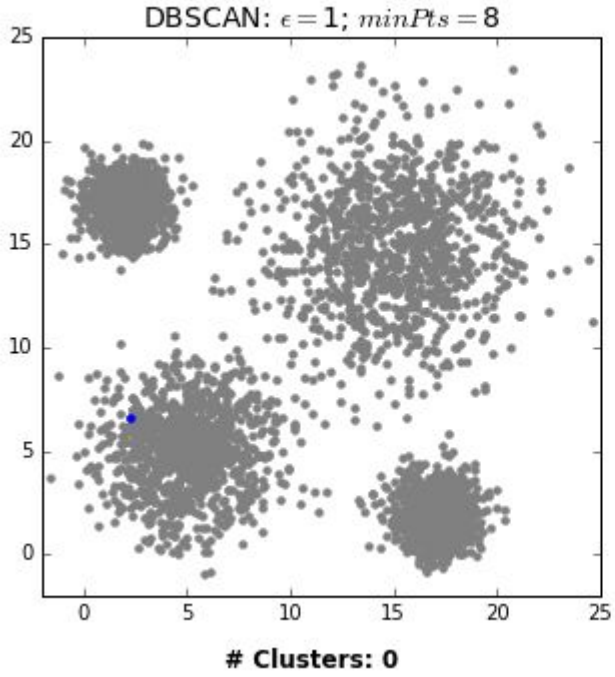
Hạn chế của K-Means

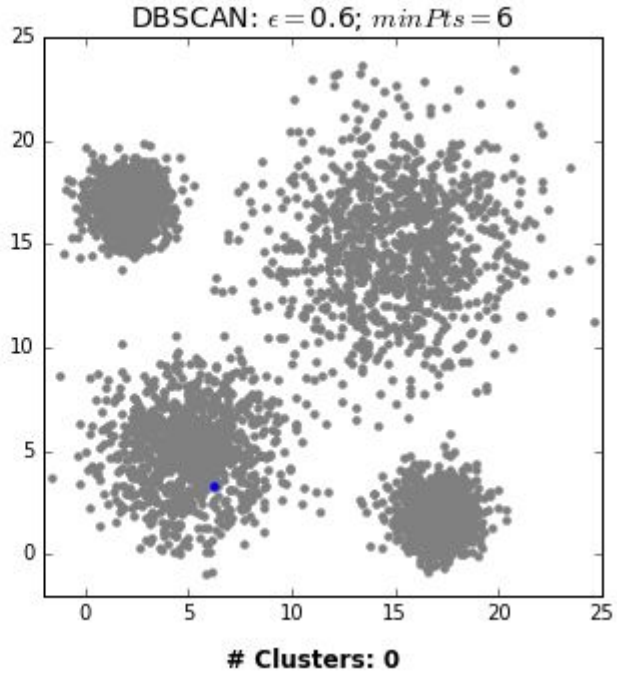
- Phụ thuộc rất nhiều vào vị trí Centroids
- Yêu cầu phải cho biết số lượng Centroids (Đối với K-Means truyền thống)
- Không phù hợp với cụm dữ liệu có hình thù không rõ ràng
- Nhạy cảm với dữ liệu nhiễu



Ý tưởng thuật toán

- Lựa chọn một điểm dữ liệu bất kì. Sau đó tiến hành xác định các điểm lõi và điểm biên thông qua vùng lân cận bằng cách lan truyền theo liên kết chuỗi các điểm thuộc cùng một cụm.
- Cụm hoàn toàn được xác định khi không thể mở rộng được thêm. Khi đó lặp lại đệ qui toàn bộ quá trình với điểm khởi tạo trong số các điểm dữ liệu còn lại để xác định một cụm mới.



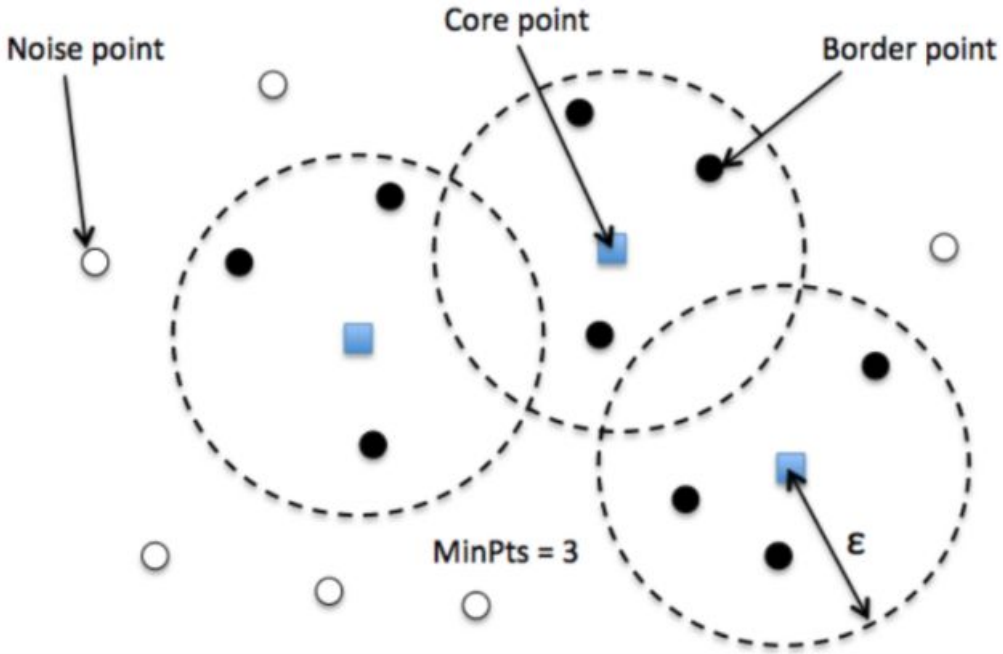


Cơ chế tìm điểm lân cận

Có 3 tham số chính quyết định quy trình xử lý của DBSCAN:

- ***Eps*** cho biết bán kính vùng quét kể từ điểm được chọn
- ***minPoint*** là số điểm tối thiểu để hợp lại thành một nhóm
- Độ đo đánh giá điểm dữ liệu (mặc định là độ đo Euclidean)

Nếu như số điểm đạt chỉ tiêu theo *minPoint* trong vùng bán kính *eps* thì sẽ gom các điểm đó vào nhóm.



Ưu điểm và nhược điểm

Ưu điểm:

- Có thể tự loại bỏ được các dữ liệu nhiễu nhờ có tính linh hoạt
- Hoạt động tốt đối với những dữ liệu có hình dạng phân phối đặc thù
- Tốc độ tính toán nhanh

Nhược điểm:

- Thường không hiệu quả đối với những dữ liệu có phân phối đều khắp nơi
- Các trọng số có ảnh hưởng rất lớn đến kết quả phân cụm

Ứng dụng của thuật toán

- Phân cụm, gom nhóm văn bản (Document Clustering)
- Lõi hệ thống khuyến nghị (Recommendation Engine)
- Phân vùng ảnh (Image Segmentation)
- Phân vùng thị trường (Market Segmentation)
- Gom nhóm kết quả tìm kiếm (Search Result Grouping)
- Phát hiện bất thường (Anomaly Detection)

Chỉ số phân cụm

Chỉ số phân cụm

- Các chỉ số phân cụm là các số cho chúng ta biết các cụm có tốt không
- Chúng giúp đánh giá và so sánh các kết quả phân cụm khác nhau trên một tập dữ liệu cho trước
- Giúp chúng ta chọn các thuật toán, tham số hoặc số lượng cụm tốt nhất cho một tập dữ liệu

Chỉ số phân cụm

- Có ba loại chỉ số phân cụm:
  - Các chỉ số nội bộ (internal metrics)

    chỉ cần dữ liệu và các cụm
  - Các chỉ số bên ngoài (external metrics)

    cần nhãn hoặc thông tin khác
  - Các chỉ số tương đối (relative metrics)

    so sánh các kết quả phân cụm khác nhau

Chỉ số phân cụm

- Các chỉ số nội bộ: chỉ cần dữ liệu và các cụm
  - Chúng đo lường độ hợp lệ, gọn gàng và tách biệt của các cụm
  - Ví dụ là silhouette coefficient, Davies-Bouldin index, Calinski-Harabasz index

Chỉ số phân cụm

- Các chỉ số bên ngoài: cần nhãn hoặc thông tin khác
  - Chúng đo lường mức độ khớp của các cụm với cấu trúc thực sự của dữ liệu
  - Ví dụ là chỉ số adjusted Rand index, normalized mutual information, F1-score

Chỉ số phân cụm

- Các chỉ số tương đối: so sánh các kết quả phân cụm khác nhau
  - Chúng đo lường sự giống hoặc khác nhau của hai kết quả phân cụm
  - Ví dụ là variation of information, cluster matching index, cophenetic correlation coefficient

Chỉ số phân cụm

- Các chỉ số phân cụm là hữu ích nhưng không hoàn hảo
- Chúng có các giả định và hạn chế khác nhau có thể không phù hợp với mục tiêu hoặc trực giác của chúng ta
- Chúng có thể không bao quát được tất cả các khía cạnh về chất lượng hoặc ý nghĩa của phân cụm
- Vì vậy chúng ta cần biết được ưu và nhược điểm của chúng và sử dụng chúng một cách khôn ngoan và thận trọng

Demo

Demo

https://colab.research.google.com/drive/1P5fVhDcsrOdCGlQdM_dbeRj-sgXahbQA?usp=sharing

Thank you for your attention