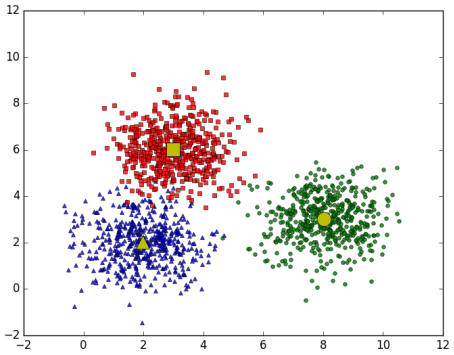
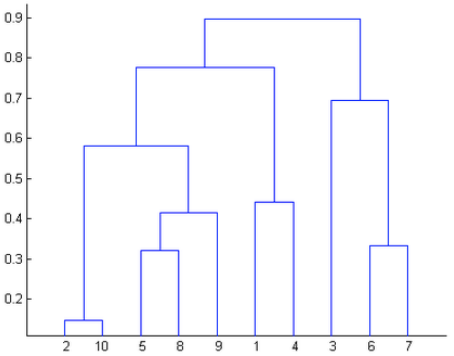
Session 4.3 DBSCAN

I. Introduction

- Đối với thuật toán *k-Means* : khởi tạo ngẫu nhiên các centroids $\rightarrow$ cập nhật cụm bằng cách cập nhật lại centroids.



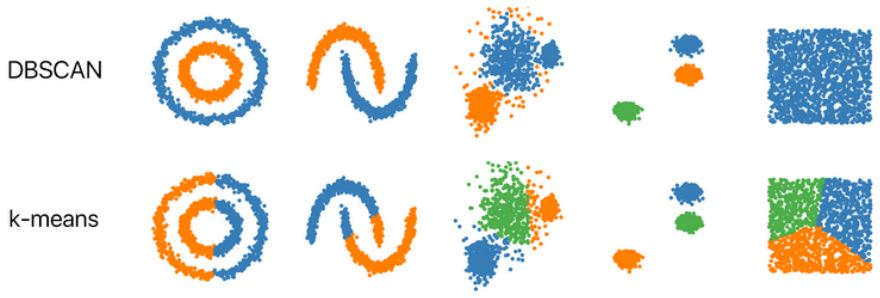
- Thuật toán phân cụm phân cấp (Hierarchical Clustering): - Thực hiện liên tiếp truy hồi quá trình gộp hoặc chia cụm. - Yoàn bộ quá trình này có thể biểu diễn thông qua một biểu đồ *dendogram* và dựa trên biểu đồ *dendogram* ta có thể xác định số lượng cụm phù hợp.



- **Nhược điểm**:
  - k-Means: phải xác định trước số cụm, tâm cụm sẽ bị ảnh hưởng bởi các điểm khởi tạo.
  - Hỉeachical Clustering: chi phí tính toán lớn $O(N^3)$ (với N là số lượng mẫu dữ liệu), không phù hợp với dữ liệu lớn.

1. So sánh K-means và DBSCAN

	K-means	DBSCAN
Nguyên lý	Phân cụm dựa trên tâm cụm (centroid-based), gán mỗi điểm đến tâm gần nhất.	Phân cụm dựa trên mật độ (density-based), hình thành cụm dựa trên các vùng có mật độ điểm cao.
Ưu điểm	Dễ hiểu, dễ triển khai.	- Khả năng phát hiện các cụm có hình dạng bất kỳ; - Loại bỏ điểm nhiễu hiệu quả; - Không cần xác định trước số lượng cụm.
Nhược điểm	- Nhạy cảm với điểm ngoại lai (outliers); - Yêu cầu xác định trước số lượng cụm; - Không phù hợp với dữ liệu có mật độ không đồng đều.	- Độ phức tạp tính toán có thể cao hơn k-Means; - Cần tinh chỉnh hai tham số eps và minPts.
Xử lý nhiễu	- Ảnh hưởng tâm cụm $\rightarrow$ chất lượng phân cụm.	- Phân loại riêng, không ảnh hưởng quá trình hình thành.
Hình dạng cụm	- Phù hợp với các cụm có hình dạng gần tròn.	- Có thể phát hiện các cụm có hình dạng bất kỳ, bao gồm cả các cụm có hình dạng lõm hoặc chồng chéo.
Số lượng cụm	Cần xác định trước.	Tự động, dựa trên mật độ dữ liệu.
Tham số	Số lượng cụm k.	Bán kính eps và số điểm tối thiểu trong bán kính $\min Pts$.
Ứng dụng	Dữ liệu đơn giản, phân phối gần tròn.	Phức tạp, dữ liệu có hình dạng bất kì và nhiều nhiễu.



- **Notes**
  - Hai tham số DBSCAN rất quan trọng, việc lựa chọn có thể ảnh hưởng đến kết quả phân cụm.
  - DBSCAN có độ phức tạp cao hơn, đặc biệt với dữ liệu phức tạp. Cần tối ưu hóa và có cấu trúc dữ liệu phù hợp.

2. DBSCAN

- DBSCAN (Density-Based Spatial Clustering of Applications with Noise), là thuật toán phân cụm dựa trên mật độ không gian với các dạng dữ liệu có *nhiễu*.
- Khi biểu diễn các điểm dữ liệu trong không gian chúng ta sẽ thấy rằng thông thường các *vùng không gian có mật độ cao sẽ xen kẽ bởi các vùng không gian có mật độ thấp*.
    - Nếu như phải dựa vào mật độ để phân chia thì khả năng rất cao những tâm cụm sẽ tập trung vào những vùng không gian có mật độ cao trong khi biên sẽ rơi vào những vùng không gian có mật độ thấp.
    - Trong lớp các mô hình phân cụm của học không giám sát tồn tại một kĩ thuật *phân cụm dựa trên mật độ* (Density-Based Clustering),
- **Ý tưởng**: một cụm trong không gian dữ liệu là một vùng có mật độ điểm cao được *ngăn cách với các cụm khác bằng các vùng liền kề có mật độ điểm thấp*.
- DBSCAN là một thuật toán cơ sở để phân nhóm dựa trên mật độ.
    - Nó có thể phát hiện ra các cụm có hình dạng và kích thước khác nhau từ một lượng lớn dữ liệu chứa nhiễu.

II. DBSCAN

1. Concepts

1.1 Eps-neighborhood

- **Vùng lân cận Epsilon** của một điểm dữ liệu P được định nghĩa là *tập tất cả các điểm nằm trong phạm vi bán kính* ε (epsilon).

$$N_{eps}(P) = \{Q \in \mathcal{D} : d(P, Q) \leq \varepsilon\}$$

1.2 Directly Density-Reachable

- **Khả năng tiếp cận trực tiếp mật độ** là việc một điểm có thể tiếp cận trực tiếp tới một điểm dữ liệu khác.

- Cụ thể là một điểm Q được coi là *có thể tiếp cận trực tiếp bởi* điểm P tương ứng với *tham số* ε và minPts nếu như nó thoả mãn hai điều kiện:
  1. $Q \in N_{eps}(P)$
     - Q nằm trong vùng lân cận của P.
  2. $|N_{eps}(Q)| \geq \mathrm{minPts}$
     - Số điểm trong vùng lận cận Q tối thiểu là minPts $\rightarrow$ không phải điểm ngoại biên (vùng mật độ thấp).
- Vậy một điểm *có thể tiếp cận* một điểm khác dựa vào:
  - Khoảng cách giữa các điểm;
  - Mật độ các điểm trung vùng lân cận epsilon phải tối thiểu bằng minPts $\rightarrow$ Vùng lân cận có mật độ cao $\rightarrow$ Phân vào cụm.
  - Các điểm thuộc vùng mật độ thấp $\rightarrow$ không có kết nối trực tiếp đến điểm trung tâm $\rightarrow$ biên cụm/nhiễu.

1.3 Density-Reachable

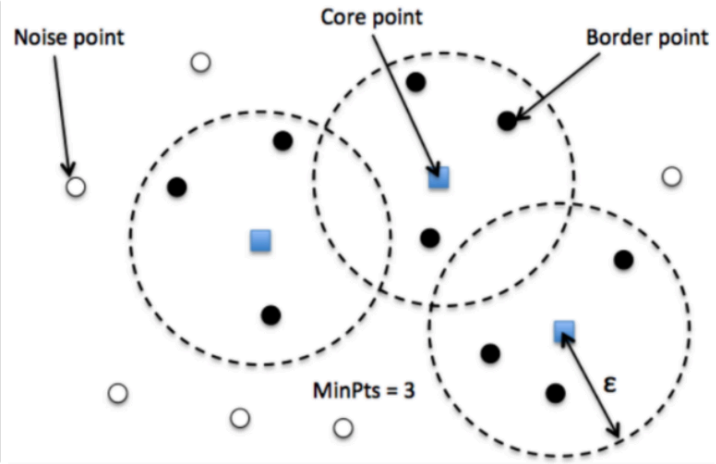
- **Khả năng tiếp cận mật độ** liên quan tới cách hình thành một chuỗi liên kết điểm trong cụm.
- Cụ thể, trong tập chuỗi điểm $\{P_i\}_{i=1}^{n} \subset \mathcal{D}$ nếu mà bất kì điểm P_i nào cũng có thể *tiếp cận trực tiếp mật độ* (Định nghĩa trên) bởi P_{i-1} theo tham số xác định $\rightarrow$ điểm $P = P_n$ *có khả năng kết nối mật độ* tới điểm $Q = P_1$.
- Từ đó suy ra, hai điểm $P_i, P_j \in \{P_i\}_{i=1}^{n}$ thỏa $i < j$ thì P_j có khả năng kết nối mật độ với P_i.
  - *Hai điểm này sẽ thuộc một cụm*.
  - Suy các điểm trong chuỗi trên đều thuộc về cùng 1 cụm.
- **Khả năng tiếp cận mật độ** thể hiện *sự mở rộng phạm vi của một cụm* dữ liệu dựa trên liên kết theo chuỗi.
  - Xuất phát từ một điểm dữ liệu ta có thể tìm được các điểm có khả năng *kết nối mật độ* tới nó theo *lan truyền chuỗi* để xác định cụm.

2. Point Classification in DBSCAN

Phân loại dạng điểm trong DBSCAN.

- Căn cứ vào vị trí các điểm so với cụm dữ liệu, ta *chia thành 3 loại* :

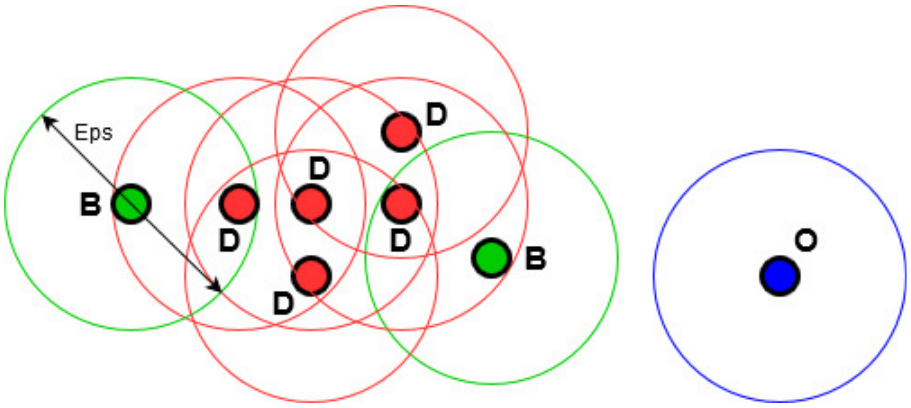
- Core (điểm lõi): sâu bên trong cụm.
- Border (điểm biên): phần ngoài cùng cụm.
- Noise (điểm nhiễu): không thuộc cụm nào.



- **Hai tham số chính**
  - minPts: ngưỡng số điểm tối thiểu được nhóm lại nhằm tạo nên vùng mật độ cao (không gồm điểm trung tâm).
  - ε (epsilon): khoảng cách xác định vùng lân cận epsilon.
- Hai giá trị trên giúp khả năng tiếp cận giữa các điểm lẫn nhau $\rightarrow$ kết nối chuỗi dữ liệu vào cụm.
- Từ đó, ta xác định 3 loại điểm nêu trên:
  - Core (điểm lõi): Đây là một điểm *có tối thiểu* minPts *điểm trong vùng lân cận* epsilon của chính nó.
  - Border (điểm biên): Đây là một điểm *có ít nhất một điểm lõi* nằm ở *vùng lân cận epsilon* nhưng *mật độ không đủ* minPts điểm.
  - Noise (điểm nhiễu): Đây là điểm *không phải* là điểm lõi hay điểm biên.
- Xét cặp điểm P và Q:
  - P, Q có khả năng kết nối mật độ với nhau: hai điểm thuộc chung 1 cụm.
  - $\begin{cases} P \text{ kết nối mật độ } Q \\ Q \text{ KHÔNG kết nối mật độ } P \end{cases}$: P điểm lõi, Q điểm biên.
  - P, Q không kết nối mật độ: hai cụm khác nhau hoặc hai điểm nhiễu.

III. DBSCAN Algorithms

1. Algorithm



- Thuật toán thực hiện lan truyền mở rộng dần phạm vi cụm tới khi chạm tới các điểm biên thì sẽ chuyển sang cụm mới và lặp lại quá trình trên.
- Qui trình của thuật toán:
  - **Bước 1:** 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 epsilon* bằng cách lan truyền theo liên kết chuỗi các điểm thuộc cùng một cụm.
  - **Bước 2:** 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.
- **PseudoCode**:
  - Bắt đầu từ điểm dữ liệu p bất kì.
  - Xác định tất cả các điểm có khả năng kết nối mật độ với p dựa theo 2 tham số. Nếu p là:
    - Điểm lõi (core): một cụm được hình thành.
    - Điểm biên (border): không có điểm nào có thể tiếp cận theo mật độ từ p, và DBSCAN truy cập điểm tiếp theo của cơ sở dữ liệu.
  - Tiếp tục đến khi tất cả các điểm đã được duyệt qua.

2. Hyper-parameters

- Tuỳ theo đặc điểm và tính chất của phân phối của bộ dữ liệu, hai tham số cần lựa chọn trong *DBSCAN* đó chính là minPts và ε:

2.1 minPts:

- Quy tắc chung, tính theo số chiều D trong tập dữ liệu:

$$\mathrm{minPts} \geq D + 1$$

- Chú ý:
  - $\text{minPts} = 1$ thì vô nghĩa do mỗi điểm sẽ tự thân nó là 1 cụm.
  - $\text{minPts} \leq 2$, kết quả đạt được sẽ giống như phân cụm phân cấp (hierarchical clustering) với *single linkage* và biểu đồ *dendrogram* được cắt ở độ cao $y = \varepsilon$
- Do đó, giá trị ít nhất phải là 3. Tuy nhiên giá trị tốt hơn sẽ tốt cho các tập dữ liệu có nhiễu và kết quả phân cụm thường hợp lý hơn.
- Theo quy tắc chung, ta thường chọn:

  $$\text{minPts} = 2 \times \text{dim}$$

  trong trường hợp dữ liệu có nhiễu hoặc có nhiều mẫu lặp lại, ta cần lựa chọn giá trị lớn hơn nữa tương ứng với những bộ dữ liệu lớn.

2.2 ε (epsilon)

- Sử dụng biểu đồ **`k-distance`**, là biểu đồ thể hiện giá trị khoảng cách trong thuật toán K-Means Clustering đến $k = \text{minPts} - 1$ điểm láng giềng gần nhất. Ứng với mỗi điểm ta chỉ lựa chọn ra khoảng cách lớn nhất trong k khoảng cách đó (được sắp xếp giảm dần trên đồ thị).
- Giá trị tốt của ε chính là vị trí các điểm khuỷu tay (elbow point):
  - Quá nhỏ, phần lớn dữ liệu không được phân cụm (nhiễu).
  - Quá lớn, các cụm sẽ hợp nhất.
- Nói chung, các giá trị nhỏ của ε được ưu tiên hơn và theo quy tắc chung, chỉ một phần nhỏ các điểm nên nằm trong vùng lân cận epsilon.

2.3 Hàm khoảng cách

- Việc lựa chọn hàm khoảng cách có mối *liên hệ chặt chẽ* với lựa chọn ε và tạo ra *ảnh hưởng lớn tới kết quả*.
- Điểm quan trọng trước tiên đó là chúng ta cần xác định một thước đo hợp lý về *độ khác biệt* (*disimilarity*) cho tập dữ liệu trước khi có thể chọn tham số ε.
  - Khoảng cách được sử dụng phổ biến nhất là `euclidean distance`.

IV. Conclusion

- DBSCAN là một thuật toán đơn giản và hiệu quả.
- Hoạt động dựa trên cách *tiếp cận mật độ phân phối của dữ liệu*.
    - Ưu điểm của thuật toán đó là có thể tự động loại bỏ được các điểm dữ liệu nhiễu, hoạt động tốt đối với những dữ liệu có hình dạng phân phối đặc thù và có tốc độ tính toán nhanh.
    - Tuy nhiên DBSCAN thường không hiệu quả đối với những dữ liệu có phân phối đều khắp nơi.
- Khi huấn luyện DBSCAN thì các *tham số của mô hình* như khoảng cách ε, số lượng điểm lân cận tối thiểu minPts và hàm khoảng cách là những tham số *có ảnh hưởng rất lớn* đối với kết quả phân cụm.
- Thực tế cho thấy thuật toán *khá nhạy với tham số* ε và minPts nên chúng ta cần phải lựa chọn tham số cho mô hình trước khi tiến hành xây dựng mô hình.