**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HCM**

**KHOA ĐÀO TẠO CHẤT LƯỢNG CAO**

-------\*\*\*-------

****

**ĐỒ ÁN 2**

**THUẬT TOÁN DBSCAN**

**Giảng viên hướng dẫn: Trần Nhật Quang**

**Nhóm thực hiện:**

**Đỗ Đào Anh Tuấn 16110241**

**Vũ Văn Đồng 16110054**

Mục Lục

**[THUẬT TOÁN DBSCAN](#_Toc9533489)** [1](#_Toc9533489)

[**1.** **Giới thiệu** 4](#_Toc9533490)

[**2.** **DBSCAN** 4](#_Toc9533491)

[**2.1** **Định nghĩa** 4](#_Toc9533492)

[**2.2** **cách hoạt động** 8](#_Toc9533493)

[**3 Thuật toán** 9](#_Toc9533494)

[**3.1 Ước tính tham số** 10](#_Toc9533495)

[**3.2 cách vận hành** 10](#_Toc9533496)

[**3.3 Thuật toán heuristics** 10](#_Toc9533497)

[**4.Code của thuật toán** 12](#_Toc9533498)

[**4.1Thư viện sklearn** 12](#_Toc9533499)

[**5 Ưu điểm và nhược điểm** 15](#_Toc9533500)

[**5.1 Ưu điểm của DBSCAN:** 15](#_Toc9533501)

[**5.2Nhược điểm của DBSCAN:** 15](#_Toc9533502)

[**6 Nguồn Kham Khảo:** 16](#_Toc9533503)

**Bảng phân công công việc**

|  |  |  |
| --- | --- | --- |
| Họ Tên | Công việc | Phần trăm |
| Đỗ Đào Anh Tuấn | Tìm hiểu, nghiên cứu, làm báo cáo | 50% |
| Vũ Văn Đồng | Tìm hiểu, nghiên cứu,  Làm code | 50% |

**THUẬT TOÁN DBSCAN**

1. **Giới thiệu**

**DBSCAN** (Density-Based Spatial Clustering of Applications with Noise) là một phương pháp học unsupervised sử dụng trong các thuật toán xây dựng model building và machine learning algorithms.

**Unsupervised learning** nghĩa là chúng ta không biết được outcome mà chỉ có dữ liệu đầu vào. Thuật toán unsupervised learning sẽ dựa vào cấu trúc của dữ liệu để thực hiện một công việc nào đó nghĩa là chúng ta không biết câu trả lời chính xác cho mỗi dữ liệu đầu vào.

1. **DBSCAN**

DBSCAN là một phương pháp phân cụm được sử dụng trong học máy để tách các cụm có mật độ cao khỏi các cụm có mật độ thấp.Nó là một thuật toán phân cụm dựa trên mật độ, nó thực hiện rất tốt việc tìm kiếm các khu vực trong dữ liệu có mật độ quan sát cao, so với các khu vực của dữ liệu không quá dày đặc với các quan sát. DBSCAN có thể sắp xếp dữ liệu thành các cụm có hình dạng khác nhau đó là một lợi thế mạnh mẽ khác!

* 1. **Định nghĩa**

**Định nghĩa 1**

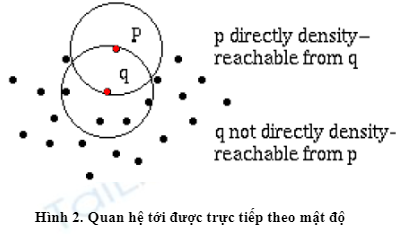
Môt vùng lân cận Eps của đối tượng p, kýhiệu NEps(p) là tập hợp các đối tượng q sao cho khoảng cách giữa p và q dist(p,q) nhỏ hơn Eps.NEps(p) = {q∈D| dist(p,q) ≤ Eps}.

**Tính chất**:-Nói chung vùng lân cận của đối tượng biên có số đối tượng ít hơn đáng kể so hơn đối tượng biên

**Định nghĩa 2** : Đối tượng p tới được trực tiếp theo mật độ (directlydensity-reachable) thỏa Eps, MinPts từ đối tượng q nếu p∈NEps(q) và |NEps(q)| ≥ MinPts.

**Tính chất**:-Nếu p, q đều là core point quan hệ directly density-reachable là đối xứng nghĩa là p tới được trực tiếp theo mật độ từ q và ngược lại.

Nếu trong p, q có một đối tượng lõi (corepoint), một đối tượng biên như hình dưới thì chỉ đối tượng biên cótới được trực tiếp theo mật độ từ đối tượng lõi mà không có chiều ngược lại (bất đối xứng).



**Định nghĩa 3:**

Đối tượng p tới được theo mật độ (density-reachable) thỏa Eps, MinPts từ đối tượng q nếu tồn tại một dãy p1,p2,...,pn(p1=q,pn=p)sao cho pi+1tới được theo mật độ trực tiếp từ pi.

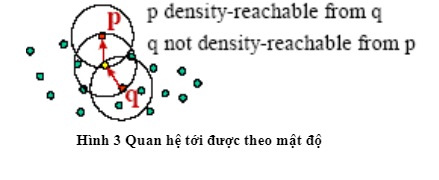
**Tính chất:**

-Quan hệ density-reachable là sự mở rộng của directly density-reachable.-Quan hệ density-reachable có tính bắt cầu

-Nếu p , q đều là đối tượng lõi (corepoint) thì quan hệ density-reachable là đối xứng nghĩa là p tới được theo mật độ từ q và ngược lại.

-Nếu p , q đều là đối tượng biên (borderpoint) thì p không tới được theo mật độ từ q và ngược lại.

-Nếu trong p, q có một đối tượng lõi(corepoint),một đối tượng biên như hình dưới thì chỉ đối tượng biên cótới được theo mật độ từ đối tượng lõi mà không có chiều ngược lại (bất đối xứng)

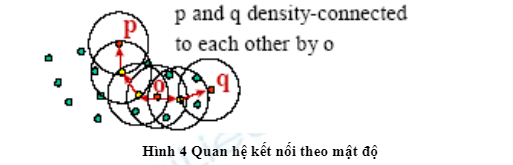


**Định nghĩa 4**

Đối tượng p kết nối theo mật độ (density-connected) thỏa Eps, MinPts với đối tượng q tồn tại đối tượng o sao cho cả p và q đều tới được theo mật độ từ o.

**Tính chất:**

-Đối với các đối tượng tới được theo mật độ với đối tượng khác, quan hệ density-connected có tính phản xạ

-Quan hệ density-connected có tính đối xứng

**Định nghĩa 5:**

Cho cơ sở dữ liệu D, cụm C thỏa Eps và MinPts là tập con khác rỗng của D thỏa 2 điều kiện sau:

1) ∀p,q: nếu p ∈C và q liên hệ theo mật độ từ p thỏa Eps và MinPts thì q∈C.

2)∀p,q∈C: p kết nối theo mật độ với p thỏa Eps và MinPts .Cụm C thỏa định nghĩa trên sẽ có ít nhất MinPts đối tượng vì lý do sau :C phải có ít nhất một đối tượngp (C khácrỗng), p phải liên hệ mật độ với bản thân nó thông qua một đối tượng o ( điều kiện 2 của định nghĩa5).

Vì vậy,o là đối tượng lõi và vùng lân cận Eps của o có ít nhất MinPts đối tượng (do p có liên hệ mật độ từ o).

**Định nghĩa 6 :**

Cho các cụm C1,..., Ck của cơ sở dữ liệu D với các tham số Epsi va MinPtsi,(i=1,...,k).

Tập nhiễu là tập các đối tượng thuộc D nhưng không thuộc bất kỳ cụmCinào.noise = {p∈D| ∀i: p ∉Ci}

**Bổ đề 1**

Cho p là một đối tượng trong D và|NEps(p)|≥MinPts.TậpO={o|o∈D vào tới được theo mật độ từ p thỏa Eps and MinPts} là một cụm thỏa Eps and MinPts.

**Bổ đề 2** Gọi C là một cụm thỏa Eps và MinPts,p là một đối tượng thuộc C sao cho|NEps(p)|≥MinPts. Khi đó tập Ctương đương với tập O ={o|o tới được theo mật độ từ p thỏa Eps and MinPts}.

* 1. **cách hoạt động**

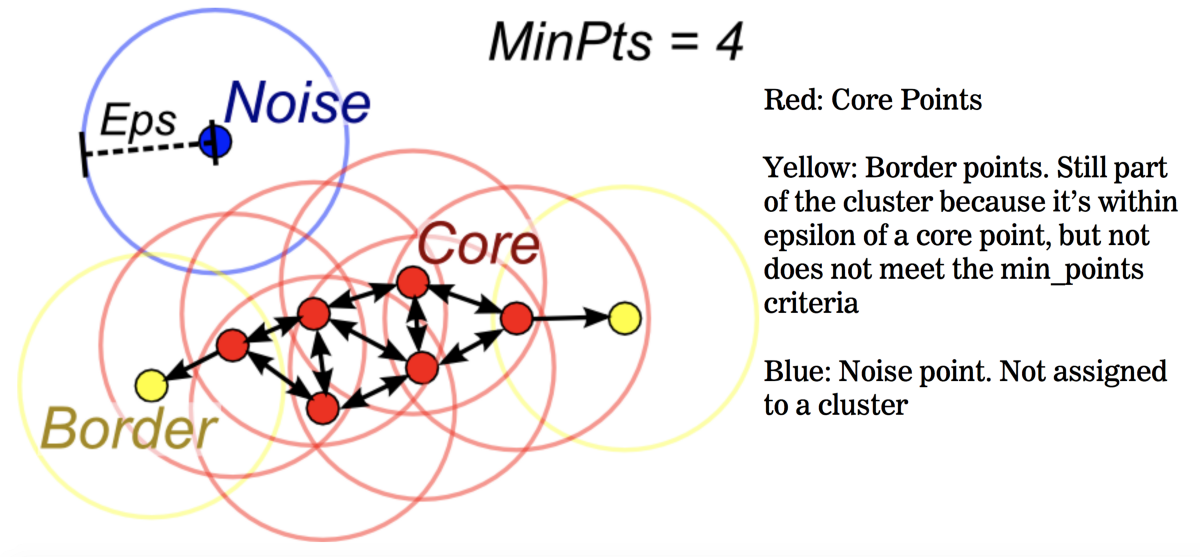
\* Chia tập dữ liệu thành n chiều.

\* Đối với mỗi điểm trong tập dữ liệu, DBSCAN tạo thành hình dạng n chiều xung quanh điểm dữ liệu đó và sau đó đếm xem có bao nhiêu điểm dữ liệu nằm trong hình dạng đó.

\* DBSCAN tính hình dạng này là một cụm. DBSCAN lặp lại mở rộng cụm, bằng cách đi qua từng điểm riêng lẻ trong cụm và đếm số lượng điểm dữ liệu khác gần đó.

\*ví dụ:

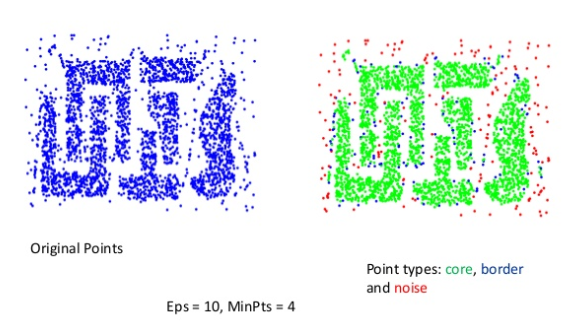
Đầu tiên DBSCAN sẽ bắt đầu bằng cách chia dữ liệu thành n chiều nó sẽ bắt đầu tại một điểm ngẫu nhiên (trong trường hợp này giả sử đó là một trong những điểm đỏ) và nó sẽ tính xem có bao nhiêu điểm khác ở gần. DBSCAN sẽ tiếp tục quá trình này cho đến khi không có điểm dữ liệu nào khác ở gần, và sau đó nó sẽ tìm cách tạo thành cụm thứ hai.

Vậy số lượng điểm cần thiết để xác định một cụm là bao nhiêu ? và khoảng cách là bao xa để một điểm có thể là điểm tiếp theo trong cụm ?

-Số điểm tối thiểu cần thiết được đặt thành 4 Khi đi qua từng điểm dữ liệu, miễn là DBSCAN tìm thấy 4 điểm trong khoảng cách epsilon của nhau, một cụm được hình thành.

-Epsilon là bán kính được đưa ra để kiểm tra khoảng cách giữa các điểm dữ liệu. Nếu một điểm nằm trong khoảng cách epsilon của một điểm khác, hai điểm đó sẽ nằm trong cùng một cụm

-Các đối tượng trong mỗi cụm được phần làm 2 loại : đối tượng bên trong cụm ( core point : đối tượng lõi ) và đối tượng nằm trên đường biên của cụm (border point : đối tượng biên ).



# **3 Thuật toán**

Thuật toán DBSCAN về cơ bản yêu cầu 2 tham số:

**Eps** : khoảng cách tối thiểu giữa hai điểm. Điều đó có nghĩa là nếu khoảng cách giữa hai điểm thấp hơn hoặc bằng giá trị này (eps), những điểm này được coi là neighbour.

**MinPoints** : số điểm tối thiểu để tạo thành một khu vực dày đặc.

Ví dụ: nếu chúng ta đặt tham số minPoints là 5, thì chúng ta cần ít nhất 5 điểm để tạo thành một vùng dày đặc.

## **3.1 Ước tính tham số**

**Eps** : nếu giá trị eps được chọn quá nhỏ, một phần lớn dữ liệu sẽ không được phân cụm. Nó sẽ được coi là ngoại lệ vì không thỏa mãn số điểm để tạo ra một khu vực dày đặc. Mặt khác, nếu giá trị được chọn quá cao, các cụm sẽ hợp nhất và phần lớn các đối tượng sẽ nằm trong cùng một cụm. Nên chọn eps dựa trên khoảng cách của tập dữ liệu, nhưng nói chung các giá trị eps nhỏ là thích hợp hơn.

**MinPoints** : Theo quy tắc chung, một minPoints tối thiểu có thể được lấy từ một số thứ nguyên (D) trong tập dữ liệu, vì minPoints ≥ D + 1. Các giá trị lớn hơn thường tốt hơn cho các tập dữ liệu có nhiễu và sẽ tạo thành các cụm đáng kể hơn. Giá trị tối thiểu cho các minPoints phải là 4, nhưng tập dữ liệu càng lớn thì giá trị minPoints nên được chọn càng lớn.

## **3.2 cách vận hành**

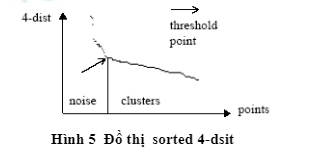
Ứng vỡi mỗi số Eps, MinPts cho trước, DBSCAN xác định một cụm thông qua 2 bước :

1. chọn đối tượng bất kỳ thỏa mãn điều kiện đối tượng lõi làm đối tượng hạt giống
2. tìm các đối tượng tới được theo mật độ từ đối tượng hạt giống . Trong trường hợp ít lý tưởng thì ứng với mỗi cụm , cần phải xác định được thông số Eps, MintPts ít nhất một đối tượng thuộc cụm ; sau đó tìm tất cả các đối tượng cho từng cụm. Tuy nhiên không dễ gì xác định được thông tin trên nhanh chóng và chính xác mêm DBSCAN sử dụng thông số Eps, MinPts của cụm có mật độ ít dày đặc nhất làm thông số chung cho tất cả các cụm và các thông số này được xác định thông qua thuật toán heuristic

## **3.3 Thuật toán heuristics**

Thông số esp và MinPts cho thuật toán DBSCAN có thể xác định bằng tay hoặc thông qua thuật toán heuistic xác định thông số Eps và MinPts cho cụm mật độ ít dày đặc nhất. Thuật toán này dựa trên 2 quan sát sau : gọi d là khoảng cách giữa đối tượng p và đối tượng gần nhất thứ k thì vùng lân cận d của đối tượng p chứa k+1 đối tượng ( hoặc nhiều hơn k+1 khi đối tượng có cùng khoảng cách đến p). Thay đổi giá trị k không dẫn đến thay đổi lớn giá trị của d trừ khi k đối tượng này cùng nằm xấp xỉ trên một đường thẳng.

Với giá trị k cho trước ,hàm k- dist là khoảng cách từ một đối tượng đến đối tượng gần nhất thứ k. Tạo đồ thị sorted k-dist bằng cáchsắp xếp các đối tượng theo giá trị k-dist giảm dần . Nếu chọn một đối tượng bất kỳ p, đặt thông số Eps là k- dist (p) và MinPts là k ,các đối tượng có khoảng cách với p nhỏ hơn hoặc bằng giá trị k dist sẽ thuộc về cụm tạo bởi đối tượng p.Nếu tìm được đối tượng ngưỡng với giá trị k-dist lớn nhất ở trong cụm mỏng nhất của D , ta sẽ tìm được giá trị thông số mong muốn.Đối tượng ngưỡng này là đối tượng đầu tiên trong vùng lõm đầu tiên của đồ thị sortedk-dist (xem hình5).Tất cả các đối tượng với giá trị k-dist cao hơn ( bên trái đối tượng ngưỡng) được xem là nhiễu. Các đối tượng còn lại (bên phải đối tượng ngưỡng ) sẽ thuộc về một cụm nào đó.



Tóm lại, thông số Eps và MinPts cho thuật toán DBSCAN có thể xác định qua cácbước sau:

-Hệ thống tính toán và hiển thị đồ thịsorted 4-dist.

-Nếu người dùng có thể ước tính số phần trăm nhiễu thì hệ thống sẽ đề nghịđối tượng ngưỡng theo số phần trăm nhiễu do người dùng nhập vào

.-Người dùng có thể chấp nhận đối tượng ngưỡng được đề nghị hoặc chọn đối tượng khác làm đối tượng ngưỡng. Giátrị 4 -dist của đối tượng ngưỡng được sử dụng làm thông số Eps cho thuật toán DBSCAN.

**4.Code của thuật toán**

**4.1Thư viện sklearn**

Thư viện Scikit-learn (viết tắt là sklearn) là một thư viện mã nguồn mở trong ngành Machine Learning, được sử dụng rộng rãi trong cộng đồng Python, được thiết kế trên nền NumPy và SciPy. Sklearn chứa hầu hết các thuật toán Machine learning hiện đại nhất, vì tính chất các thuật toán đã được viết sẵn, chúng ta chỉ cần đưa dữ liệu vào là nó có thể tự tính toán rồi đưa ra kết quả. Điểm mạnh của thư viện này là do nó được sử dụng nhiều trong giáo dục và công nghiệp cho nên nó luôn được cập nhật.

Lý do mà chúng ta nên dùng thư viện sklearn là vì nó cho phép ta sử dụng ngay các thuật toán một cách nhanh chóng và hiệu quả, sklearn là một trong những sự lựa chọn hàng đầu của các Dev

**print**(\_\_doc\_\_)

**import** **numpy** **as** **np**

**from** **sklearn.cluster** **import** [DBSCAN](https://word-edit.officeapps.live.com/we/wordeditorframe.aspx?ui=en-US&rs=vi-VN&hid=Tg1zeqE/nk+hq/rp2sqIrQ.0&WOPISrc=https%3A%2F%2Fwopi.onedrive.com%2Fwopi%2Ffiles%2F2E8DD5B4B01A6C46%21754&wdnd=1&wdPreviousSession=4a94ee68-cfba-4b0a-9e91-ca9f055a581f&wdNewAndOpenCt=1558592019997&wdo=7&wdPreviousCorrelation=fbd3f029-3820-48e0-af37-9e296c9f7a59&wde=docx&sc=host%3D%26qt%3DFolders&mscc=1&wdp=0#sklearn.cluster.DBSCAN)

**from** **sklearn** **import** metrics

**from** **sklearn.datasets.samples\_generator** **import** [make\_blobs](https://word-edit.officeapps.live.com/we/wordeditorframe.aspx?ui=en-US&rs=vi-VN&hid=Tg1zeqE/nk+hq/rp2sqIrQ.0&WOPISrc=https%3A%2F%2Fwopi.onedrive.com%2Fwopi%2Ffiles%2F2E8DD5B4B01A6C46%21754&wdnd=1&wdPreviousSession=4a94ee68-cfba-4b0a-9e91-ca9f055a581f&wdNewAndOpenCt=1558592019997&wdo=7&wdPreviousCorrelation=fbd3f029-3820-48e0-af37-9e296c9f7a59&wde=docx&sc=host%3D%26qt%3DFolders&mscc=1&wdp=0" \l "sklearn.datasets.make_blobs" \t "_blank)

**from** **sklearn.preprocessing** **import** [StandardScaler](https://word-edit.officeapps.live.com/we/wordeditorframe.aspx?ui=en-US&rs=vi-VN&hid=Tg1zeqE/nk+hq/rp2sqIrQ.0&WOPISrc=https%3A%2F%2Fwopi.onedrive.com%2Fwopi%2Ffiles%2F2E8DD5B4B01A6C46%21754&wdnd=1&wdPreviousSession=4a94ee68-cfba-4b0a-9e91-ca9f055a581f&wdNewAndOpenCt=1558592019997&wdo=7&wdPreviousCorrelation=fbd3f029-3820-48e0-af37-9e296c9f7a59&wde=docx&sc=host%3D%26qt%3DFolders&mscc=1&wdp=0" \l "sklearn.preprocessing.StandardScaler" \t "_blank)

*# khởi tạo dữ liệu trung tâm*

centers = [[1, 1], [-1, -1], [1, -1]]

X, labels\_true = [make\_blobs](https://word-edit.officeapps.live.com/we/wordeditorframe.aspx?ui=en-US&rs=vi-VN&hid=Tg1zeqE/nk+hq/rp2sqIrQ.0&WOPISrc=https%3A%2F%2Fwopi.onedrive.com%2Fwopi%2Ffiles%2F2E8DD5B4B01A6C46%21754&wdnd=1&wdPreviousSession=4a94ee68-cfba-4b0a-9e91-ca9f055a581f&wdNewAndOpenCt=1558592019997&wdo=7&wdPreviousCorrelation=fbd3f029-3820-48e0-af37-9e296c9f7a59&wde=docx&sc=host%3D%26qt%3DFolders&mscc=1&wdp=0" \l "sklearn.datasets.make_blobs" \t "_blank)(n\_samples=750, centers=centers, cluster\_std=0.4,

                            random\_state=0)

X = [StandardScaler](https://word-edit.officeapps.live.com/we/wordeditorframe.aspx?ui=en-US&rs=vi-VN&hid=Tg1zeqE/nk+hq/rp2sqIrQ.0&WOPISrc=https%3A%2F%2Fwopi.onedrive.com%2Fwopi%2Ffiles%2F2E8DD5B4B01A6C46%21754&wdnd=1&wdPreviousSession=4a94ee68-cfba-4b0a-9e91-ca9f055a581f&wdNewAndOpenCt=1558592019997&wdo=7&wdPreviousCorrelation=fbd3f029-3820-48e0-af37-9e296c9f7a59&wde=docx&sc=host%3D%26qt%3DFolders&mscc=1&wdp=0" \l "sklearn.preprocessing.StandardScaler" \t "_blank)().fit\_transform(X)

*# #############################################################################*

*# Compute DBSCAN*

db = [DBSCAN](https://word-edit.officeapps.live.com/we/wordeditorframe.aspx?ui=en-US&rs=vi-VN&hid=Tg1zeqE/nk+hq/rp2sqIrQ.0&WOPISrc=https%3A%2F%2Fwopi.onedrive.com%2Fwopi%2Ffiles%2F2E8DD5B4B01A6C46%21754&wdnd=1&wdPreviousSession=4a94ee68-cfba-4b0a-9e91-ca9f055a581f&wdNewAndOpenCt=1558592019997&wdo=7&wdPreviousCorrelation=fbd3f029-3820-48e0-af37-9e296c9f7a59&wde=docx&sc=host%3D%26qt%3DFolders&mscc=1&wdp=0#sklearn.cluster.DBSCAN)(eps=0.3, min\_samples=10).fit(X) //Gán epsilon là 0,3,minpoints là 10

core\_samples\_mask = [np.zeros\_like](https://word-edit.officeapps.live.com/we/wordeditorframe.aspx?ui=en-US&rs=vi-VN&hid=Tg1zeqE/nk+hq/rp2sqIrQ.0&WOPISrc=https%3A%2F%2Fwopi.onedrive.com%2Fwopi%2Ffiles%2F2E8DD5B4B01A6C46%21754&wdnd=1&wdPreviousSession=4a94ee68-cfba-4b0a-9e91-ca9f055a581f&wdNewAndOpenCt=1558592019997&wdo=7&wdPreviousCorrelation=fbd3f029-3820-48e0-af37-9e296c9f7a59&wde=docx&sc=host%3D%26qt%3DFolders&mscc=1&wdp=0" \l "numpy.zeros_like" \t "_blank)(db.labels\_, dtype=bool)

core\_samples\_mask[db.core\_sample\_indices\_] = True

labels = db.labels\_

*# Số cụm trên biểu đồ khi bỏ qua điểm nhiễu*

n\_clusters\_ = len(set(labels)) - (1 **if** -1 **in** labels **else** 0).

n\_noise\_ = list(labels).count(-1)

**print**('Estimated number of clusters: *%d*' % n\_clusters\_)

**print**('Estimated number of noise points: *%d*' % n\_noise\_)

**print**("Homogeneity: *%0.3f*" % [metrics.homogeneity\_score](https://word-edit.officeapps.live.com/we/wordeditorframe.aspx?ui=en-US&rs=vi-VN&hid=Tg1zeqE/nk+hq/rp2sqIrQ.0&WOPISrc=https%3A%2F%2Fwopi.onedrive.com%2Fwopi%2Ffiles%2F2E8DD5B4B01A6C46%21754&wdnd=1&wdPreviousSession=4a94ee68-cfba-4b0a-9e91-ca9f055a581f&wdNewAndOpenCt=1558592019997&wdo=7&wdPreviousCorrelation=fbd3f029-3820-48e0-af37-9e296c9f7a59&wde=docx&sc=host%3D%26qt%3DFolders&mscc=1&wdp=0#sklearn.metrics.homogeneity_score)(labels\_true, labels))

**print**("Completeness: *%0.3f*" % [metrics.completeness\_score](https://word-edit.officeapps.live.com/we/wordeditorframe.aspx?ui=en-US&rs=vi-VN&hid=Tg1zeqE/nk+hq/rp2sqIrQ.0&WOPISrc=https%3A%2F%2Fwopi.onedrive.com%2Fwopi%2Ffiles%2F2E8DD5B4B01A6C46%21754&wdnd=1&wdPreviousSession=4a94ee68-cfba-4b0a-9e91-ca9f055a581f&wdNewAndOpenCt=1558592019997&wdo=7&wdPreviousCorrelation=fbd3f029-3820-48e0-af37-9e296c9f7a59&wde=docx&sc=host%3D%26qt%3DFolders&mscc=1&wdp=0#sklearn.metrics.completeness_score)(labels\_true, labels))

**print**("V-measure: *%0.3f*" % [metrics.v\_measure\_score](https://word-edit.officeapps.live.com/we/wordeditorframe.aspx?ui=en-US&rs=vi-VN&hid=Tg1zeqE/nk+hq/rp2sqIrQ.0&WOPISrc=https%3A%2F%2Fwopi.onedrive.com%2Fwopi%2Ffiles%2F2E8DD5B4B01A6C46%21754&wdnd=1&wdPreviousSession=4a94ee68-cfba-4b0a-9e91-ca9f055a581f&wdNewAndOpenCt=1558592019997&wdo=7&wdPreviousCorrelation=fbd3f029-3820-48e0-af37-9e296c9f7a59&wde=docx&sc=host%3D%26qt%3DFolders&mscc=1&wdp=0" \l "sklearn.metrics.v_measure_score" \t "_blank)(labels\_true, labels))

**print**("Adjusted Rand Index: *%0.3f*"

      % [metrics.adjusted\_rand\_score](https://word-edit.officeapps.live.com/we/wordeditorframe.aspx?ui=en-US&rs=vi-VN&hid=Tg1zeqE/nk+hq/rp2sqIrQ.0&WOPISrc=https%3A%2F%2Fwopi.onedrive.com%2Fwopi%2Ffiles%2F2E8DD5B4B01A6C46%21754&wdnd=1&wdPreviousSession=4a94ee68-cfba-4b0a-9e91-ca9f055a581f&wdNewAndOpenCt=1558592019997&wdo=7&wdPreviousCorrelation=fbd3f029-3820-48e0-af37-9e296c9f7a59&wde=docx&sc=host%3D%26qt%3DFolders&mscc=1&wdp=0" \l "sklearn.metrics.adjusted_rand_score" \t "_blank)(labels\_true, labels))

**print**("Adjusted Mutual Information: *%0.3f*"

      % [metrics.adjusted\_mutual\_info\_score](https://word-edit.officeapps.live.com/we/wordeditorframe.aspx?ui=en-US&rs=vi-VN&hid=Tg1zeqE/nk+hq/rp2sqIrQ.0&WOPISrc=https%3A%2F%2Fwopi.onedrive.com%2Fwopi%2Ffiles%2F2E8DD5B4B01A6C46%21754&wdnd=1&wdPreviousSession=4a94ee68-cfba-4b0a-9e91-ca9f055a581f&wdNewAndOpenCt=1558592019997&wdo=7&wdPreviousCorrelation=fbd3f029-3820-48e0-af37-9e296c9f7a59&wde=docx&sc=host%3D%26qt%3DFolders&mscc=1&wdp=0" \l "sklearn.metrics.adjusted_mutual_info_score" \t "_blank)(labels\_true, labels,

                                           average\_method='arithmetic'))

**print**("Silhouette Coefficient: *%0.3f*"

      % [metrics.silhouette\_score](https://word-edit.officeapps.live.com/we/wordeditorframe.aspx?ui=en-US&rs=vi-VN&hid=Tg1zeqE/nk+hq/rp2sqIrQ.0&WOPISrc=https%3A%2F%2Fwopi.onedrive.com%2Fwopi%2Ffiles%2F2E8DD5B4B01A6C46%21754&wdnd=1&wdPreviousSession=4a94ee68-cfba-4b0a-9e91-ca9f055a581f&wdNewAndOpenCt=1558592019997&wdo=7&wdPreviousCorrelation=fbd3f029-3820-48e0-af37-9e296c9f7a59&wde=docx&sc=host%3D%26qt%3DFolders&mscc=1&wdp=0" \l "sklearn.metrics.silhouette_score" \t "_blank)(X, labels))

*# #############################################################################*

*# import thư viện matlap*

**import** **matplotlib.pyplot** **as** **plt**

*# Hàm để xóa đi các điểm đen và lấy các điểm nhiễu để thay thế*

unique\_labels = set(labels)

colors = [plt.cm.Spectral(each)

**for** each **in** [np.linspace](https://word-edit.officeapps.live.com/we/wordeditorframe.aspx?ui=en-US&rs=vi-VN&hid=Tg1zeqE/nk+hq/rp2sqIrQ.0&WOPISrc=https%3A%2F%2Fwopi.onedrive.com%2Fwopi%2Ffiles%2F2E8DD5B4B01A6C46%21754&wdnd=1&wdPreviousSession=4a94ee68-cfba-4b0a-9e91-ca9f055a581f&wdNewAndOpenCt=1558592019997&wdo=7&wdPreviousCorrelation=fbd3f029-3820-48e0-af37-9e296c9f7a59&wde=docx&sc=host%3D%26qt%3DFolders&mscc=1&wdp=0" \l "numpy.linspace" \t "_blank)(0, 1, len(unique\_labels))]

**for** k, col **in** zip(unique\_labels, colors):

**if** k == -1:

*# Các điểm đen đổi thành điểm nhiễu*

        col = [0, 0, 0, 1]

    class\_member\_mask = (labels == k)

    xy = X[class\_member\_mask & core\_samples\_mask]

[plt.plot](https://word-edit.officeapps.live.com/we/wordeditorframe.aspx?ui=en-US&rs=vi-VN&hid=Tg1zeqE/nk+hq/rp2sqIrQ.0&WOPISrc=https%3A%2F%2Fwopi.onedrive.com%2Fwopi%2Ffiles%2F2E8DD5B4B01A6C46%21754&wdnd=1&wdPreviousSession=4a94ee68-cfba-4b0a-9e91-ca9f055a581f&wdNewAndOpenCt=1558592019997&wdo=7&wdPreviousCorrelation=fbd3f029-3820-48e0-af37-9e296c9f7a59&wde=docx&sc=host%3D%26qt%3DFolders&mscc=1&wdp=0#matplotlib.pyplot.plot)(xy[:, 0], xy[:, 1], 'o', markerfacecolor=tuple(col),

             markeredgecolor='k', markersize=14)

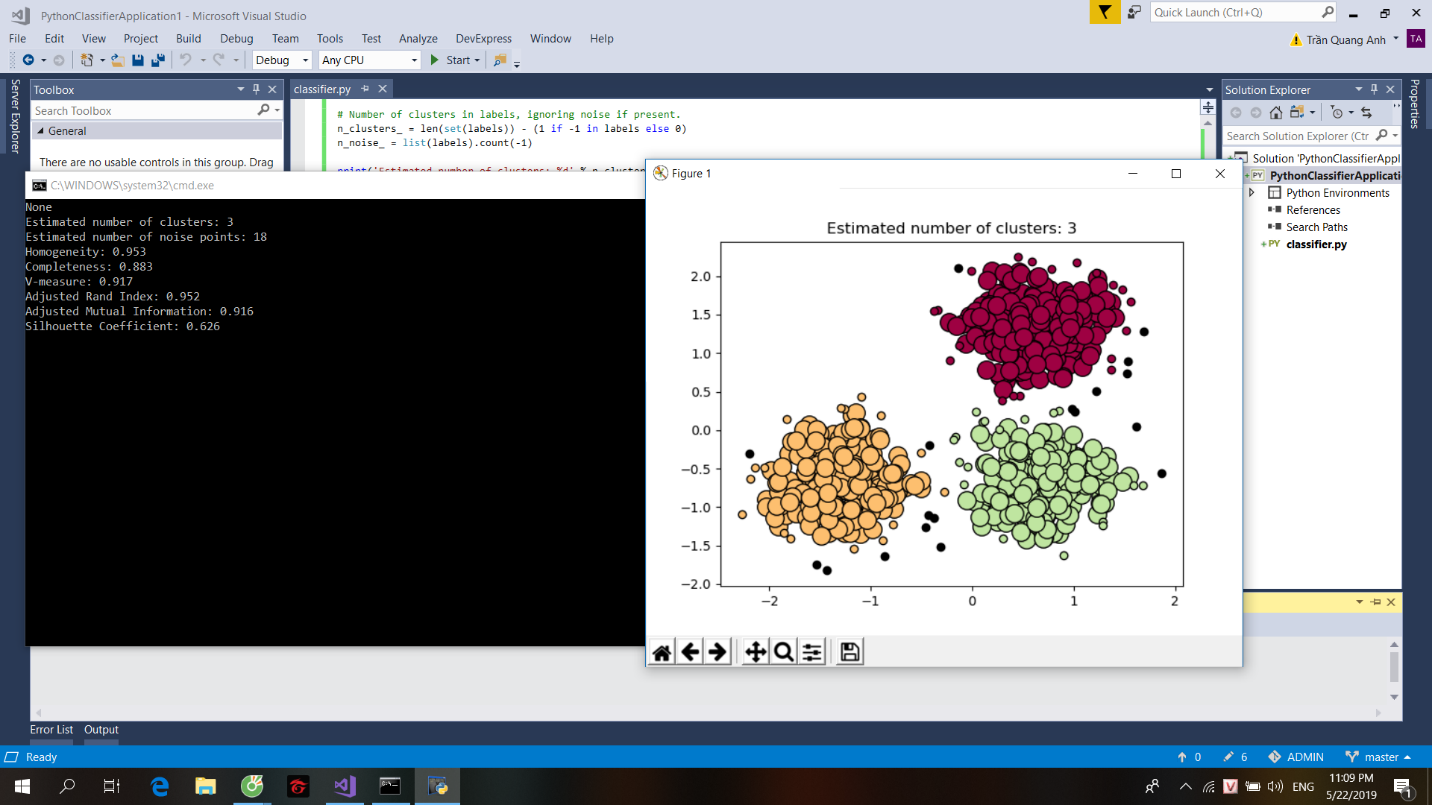
    xy = X[class\_member\_mask & ~core\_samples\_mask]

[plt.plot](https://word-edit.officeapps.live.com/we/wordeditorframe.aspx?ui=en-US&rs=vi-VN&hid=Tg1zeqE/nk+hq/rp2sqIrQ.0&WOPISrc=https%3A%2F%2Fwopi.onedrive.com%2Fwopi%2Ffiles%2F2E8DD5B4B01A6C46%21754&wdnd=1&wdPreviousSession=4a94ee68-cfba-4b0a-9e91-ca9f055a581f&wdNewAndOpenCt=1558592019997&wdo=7&wdPreviousCorrelation=fbd3f029-3820-48e0-af37-9e296c9f7a59&wde=docx&sc=host%3D%26qt%3DFolders&mscc=1&wdp=0#matplotlib.pyplot.plot)(xy[:, 0], xy[:, 1], 'o', markerfacecolor=tuple(col),

             markeredgecolor='k', markersize=6)

[plt.title](https://word-edit.officeapps.live.com/we/wordeditorframe.aspx?ui=en-US&rs=vi-VN&hid=Tg1zeqE/nk+hq/rp2sqIrQ.0&WOPISrc=https%3A%2F%2Fwopi.onedrive.com%2Fwopi%2Ffiles%2F2E8DD5B4B01A6C46%21754&wdnd=1&wdPreviousSession=4a94ee68-cfba-4b0a-9e91-ca9f055a581f&wdNewAndOpenCt=1558592019997&wdo=7&wdPreviousCorrelation=fbd3f029-3820-48e0-af37-9e296c9f7a59&wde=docx&sc=host%3D%26qt%3DFolders&mscc=1&wdp=0#matplotlib.pyplot.title)('Estimated number of clusters: *%d*' % n\_clusters\_)

[plt.show](https://word-edit.officeapps.live.com/we/wordeditorframe.aspx?ui=en-US&rs=vi-VN&hid=Tg1zeqE/nk+hq/rp2sqIrQ.0&WOPISrc=https%3A%2F%2Fwopi.onedrive.com%2Fwopi%2Ffiles%2F2E8DD5B4B01A6C46%21754&wdnd=1&wdPreviousSession=4a94ee68-cfba-4b0a-9e91-ca9f055a581f&wdNewAndOpenCt=1558592019997&wdo=7&wdPreviousCorrelation=fbd3f029-3820-48e0-af37-9e296c9f7a59&wde=docx&sc=host%3D%26qt%3DFolders&mscc=1&wdp=0#matplotlib.pyplot.show)()



**5 Ưu điểm và nhược điểm**

**5.1 Ưu điểm của DBSCAN:**

Khám phá các cụm có hình dáng bất kỳ

-Chỉ cần một thông số đầu vào (số phần trăm nhiễu)

-Có thể thay đổi quy mô (scalability): hiệu quả trong cơ sở dữ liệu lớn

-Có khả năng xử lý nhiễu

**5.2Nhược điểm của DBSCAN:**

DBSCAN không thể hoàn toàn xác định các điểm biên, các điểm này có thể truy cập từ nhiều hơn một cụm và có thể là một phần của một trong hai cụm

DBSCAN không thể phân cụm dữ liệu tốt với sự khác biệt lớn về mật độ.

# **6 Nguồn Kham Khảo:**

<https://medium.com/@elutins/dbscan-what-is-it-when-to-use-it-how-to-use-it-8bd506293818>

<http://www.cit.ctu.edu.vn/~dtnghi/rech/dir_0/Duy-et-al-New.pdf>

<http://vai.org.vn/docs/Daotao/PtichDlieu/Thu6/SangThu6.pdf>

<https://en.wikipedia.org/wiki/DBSCAN>