**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HCM  
KHOA ĐÀO TẠO CHẤT LƯỢNG CAO  
NGÀNH CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO ĐỒ ÁN 3**

**TÌM HIỂU VÀ DEMO MỘT HỆ THỐNG GỢI Ý SỬ DỤNG THUẬT TOÁN DBSCAN**

**GVHD: Ts. Lê Văn Vinh**

**Sinh viên thực hiện MSSV**

**Luyện Ngọc Thanh 17110221**

**Nguyễn Thanh Lập 17110169**

Tp. Hồ Chí Minh, tháng 10 năm 2020

MỤC LỤC

[**CHƯƠNG 1: GIỚI THIỆU** 1](#_Toc60513905)

[**1.** **Machine learning là gì?** 1](#_Toc60513906)

[**2.** **Các loại hệ thống Machine learning** 1](#_Toc60513907)

[**3.** **Giải pháp** 1](#_Toc60513908)

[**CHƯƠNG 2: TÌM HIỂU THUẬT TOÁN DBSCAN** 4](#_Toc60513909)

[**1.** **Sơ lược về DBSCAN** 4](#_Toc60513910)

[**2.** **Giải thuật.** 5](#_Toc60513911)

[**CHƯƠNG 3. DEMO 7**](#_Toc60513912)

[**3.1. Demo gom cụm với thuật toán dbscan dựa trên ngôn ngữ python 7**](#_Toc60513913)

[**CHƯƠNG 4: KẾT LUẬN 11**](#_Toc60513914)

**4.**[**1. Tổng kết 11**](#_Toc60513915)

**4.**[**2. Hạn chế 11**](#_Toc60513916)

[Hình 1. Mô phỏng quy trình gom cụm 5](#_Toc60513918)

[Hình 2. Giải thuật dbscan 6](#_Toc60513919)

[Hình 3. Đọc dữ liệu từ file mall\_customers.csv 7](#_Toc60513920)

[Hình 4. Đọc và xử lý dữ liệu từ dataset 7](#_Toc60513921)

[Hình 5. Xác định eps và min\_samples 8](#_Toc60513922)

[Hình 6. Heatplot xác định số cluster 9](#_Toc60513923)

[Hình 7. Heatplot xác định eps và min\_samples 9](#_Toc60513924)

[Hình 8. Hàm phân cụm với 2 tham số eps và min\_samples 10](#_Toc60513925)

[Hình 9. Tiến hành phân cụm và visualize kết quả 10](#_Toc60513926)

[Hình 10. Visualize kết quả phân cụm với thuật toán DBSCAN 10](#_Toc60513927)

# **CHƯƠNG 1: GIỚI THIỆU**

1. **Machine learning là gì?**

Machine learning là khoa học (và nghệ thuật) lập trình máy tính để chúng có thể

học từ dữ liệu.

1. **Các loại hệ thống Machine learning**

Có rất nhiều loại hệ thống Machine learning khác nhau có thể được phân loại dựa trên:

* Có được đào tạo với sự giám sát của con người hay không:
* Supervised learning (Học có giám sát)
* Unsupervised learning (Học không giám sát)
* Semi-supervised learning (Học bán giám sát)
* Reinforcement learning (Học tăng cường)
* Có thể học tăng dần một cách nhanh chóng hay không (online versus batch learning).
* Có hoạt động đơn giản bằng cách so sánh các điểm dữ liệu mới với các điểm dữ liệu đã biết hay thay vào đó là phát hiện các mẫu trong dữ liệu đào tạo và xây dựng mô hình dự đoán, giống như các nhà khoa học làm (instance-based versus model-based learning).

1. **Giải pháp**

* Sau khi tìm hiểu các dạng Machine learning khác nhau, nhóm nghiên cứu quyết định tìm hiểu sâu về hai dạng của Machine learning là Supervised learning và Unsupervised learning để tìm ra giải pháp cho bài toán. Từ đó sẽ tìm hiểu một số thuật toán phù hợp để giải quyết vấn đề đặt ra ứng với mỗi dạng.
* Cụ thể với Supervised learning nhóm sẽ tìm hiểu về thuật toán DBSCAN.
  1. **Supervised learning**
* Trong học tập có giám sát, dữ liệu training bạn cung cấp cho thuật toán bao gồm các giải pháp mong muốn, được gọi là nhãn.
* Một ví dụ về học tập có giám sát điển hình là phân loại. Bộ lọc thư rác là một ví dụ điển hình về điều này: nó được đào tạo với nhiều email mẫu cùng với lớp của chúng (spam hoặc ham) và nó phải học cách để phân loại các email mới.
* Dưới đây là một số thuật toán học có giám sát quan trọng:
* k-Nearest Neighbors
* Linear Regression
* Logistic Regression
* Support Vector Machines (SVMs)
* Decision Trees and Random Forests
* Neural networks
  1. **Unsupervised learning**
* Theo như tên gọi của thuật toán, việc học máy máy sẽ không có sự hướng dẫn rõ ràng và dữ liệu được đưa vào sẽ không được tiền xử lý bởi con người.
* Một số thuật toán không giám sát quan trọng:
* Clustering
* k-Means
* Hierarchical Cluster Analysis (HCA)
* Expectation Maximization
* Visualization and dimensionality reduction
* Principal Component Analysis (PCA)
* Kernel PCA
* Locally - Linear Embedding (LLE)
* t-distributed Stochastic Neighbor Embedding (t-SNE)
* Association rule learning
* Apriori
* Eclat
* Một ví dụ điển hình học tập không giám sát là giả sử bạn có nhiều dữ liệu về khách truy cập trang web đọc sách của mình. Bạn có thể muốn chạy một thuật toán phân cụm để cố gắng phát hiện các nhóm đọc giả truy cập tương tự (như nhóm yêu thích thể loại khoa học viễn tưởng, hay nhóm thích sách về kinh tế, …). Bạn không cho thuật toán biết trước là khách truy cập thuộc nhóm nào: nó sẽ phải tìm thấy các kết nối đó mà không cần sự trợ giúp của bạn.

# **CHƯƠNG 2: TÌM HIỂU THUẬT TOÁN DBSCAN**

1. **Sơ lược về DBSCAN**

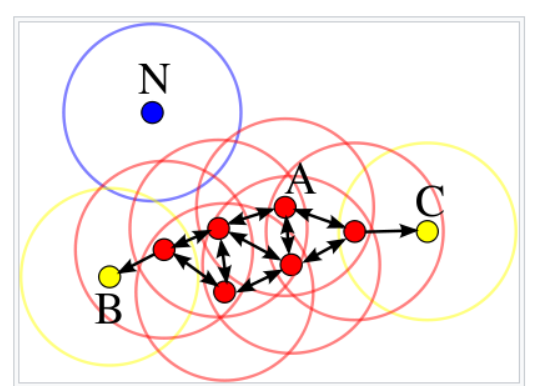
DBSCAN hay Density Base Spatial Clustering of Application with Noise, ý tưởng chính của thuật toán là một điểm sẽ thuộc về 1 cluster nếu nó gần các điểm thuộc cluster đó.

Có 2 tham số chính của DBSCAN:

* eps: Là khoảng cách của các “hàng xóm”. Hai điểm được coi là “hàng xóm” nếu khoảng cách giữa chúng nhỏ hơn hoặc bằng eps.
* minPts: Số điểm tối thiểu để tạo thành 1 cluster.

Dựa vào 2 tham số trên, các điểm được phân loại thành điểm lõi, điểm biên hoặc điểm ngoài.

* Điểm lõi: Một điểm sẽ là điểm lõi nếu có tối thiểu minPts số điểm tính cả chính nó được bao quanh bởi 1 khu vực với bán kính là eps.
* Điểm biên: Một điểm sẽ là điểm biên nếu nó có thể được “chạm tới” từ điểm lõi và có ít hơn minPts số điểm ở trong khu vực của nó.
* Điểm ngoài: Một điểm sẽ là điểm ngoài nếu nó không phải là điểm lõi cũng như là điểm biên.



Hình 1. Mô phỏng quy trình gom cụm

Ở trường hợp này minPts sẽ là 4, điểm A sẽ là điểm lõi vì quanh nó có tối thiểu 4 điểm với bán kính là eps, B và C sẽ là điểm biên vì nó có thể “chạm tới” từ điểm A và có ít hơn 4 điểm bao quanh nó, cuối cùng điểm N sẽ là điểm ngoài bởi vì N không thể được “chạm tới” bởi bất cứ điểm nào nằm trong cluster.

1. **Giải thuật.**

* Xác định minPts và eps.
* Điểm bắt đầu được chọn ngẫu nhiên ở một khu vực được xác định bởi bán kính eps. Nếu ở đó có tối thiểu minPts số điểm trong vùng phụ cận thì điểm đó được đánh dấu là điểm lõi nếu không nó sẽ được coi là điểm gây nhiễu (noise point). Một khi thông tin về 1 cluser được hình thành (tạm gọi là cluster A), tất cả các điểm ở trong khu vực của điểm khởi tạo sẽ trở thành một phần của cluster A. Nếu có một điểm mới mà điểm đó cũng là điểm lõi thì các điểm trong khu vực của điểm đó cũng được thêm vào cluster A.

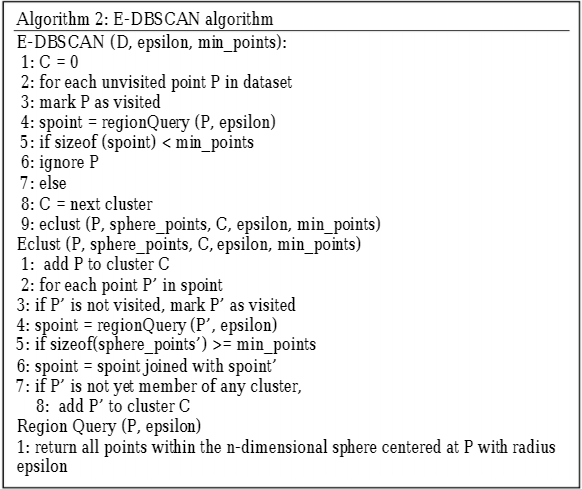
Chú ý: Điểm được coi là điểm gây nhiễu cũng có thể được kiểm tra lại và trở thành một phần của cluster nếu thỏa mãn điều kiện.

* Bước tiếp theo là chọn ngẫu nhiên một điểm ở giữa các điểm mà điểm đó chưa được “ghé thăm” ở bước trước đó. Và vẫn là quy trình kiểm tra được áp dụng.
* Tiến trình này sẽ kết thúc khi tất cả các điểm đã được “ghé thăm”.

Chú ý: Khoảng cách giữa các điểm được xác định bằng cách sử dụng phương pháp tính khoảng cách được đề cập ở giải thuật k-means.

Bằng cách áp dụng các bước trên, giải thuật DBSCAN phù hợp để tìm các khu vực có tỉ trọng cao và tách nó ra khỏi khu vực có tỉ trọng thấp.

Điều kiện cần để hình thành một cluster đó là phải có tối thiểu một điểm lõi. Trường hợp xấu nhất cluster sẽ có chỉ một điểm lõi và các điểm biên của nó.

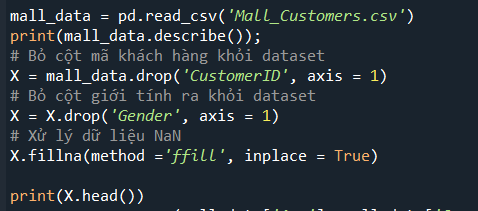


Hình 2. Giải thuật dbscan

CHƯƠNG 3. DEMO

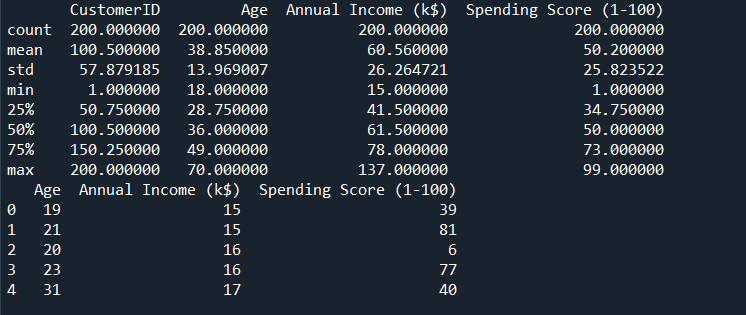
3.1. Demo gom cụm với thuật toán dbscan dựa trên ngôn ngữ python

Đầu tiên chúng ta tiến hành đọc dữ liệu từ dataset



Hình 3. Đọc dữ liệu từ file mall\_customers.csv

Kết quả thu được



Hình 4. Đọc và xử lý dữ liệu từ dataset

Để xác định được tham số eps và min\_samples (min Pts) trong thuật toán DBSCAN, ta phải dựa vào tham số silhouette score để đưa ra lựa chọn tốt nhất.

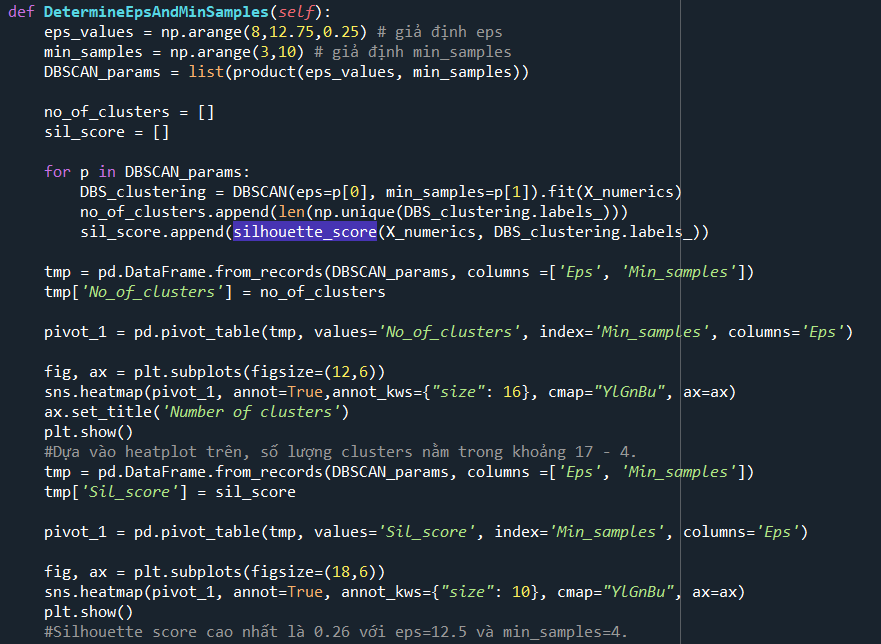
*Silhouette score* làmột thông số đánh giá hiệu năng của việc phân cụm. Cụ thể chúng ta sử dụng silhouette score. Với mỗi ví dụ, giá trị silhouette score của nó được tính như sau:

Silhouette score =

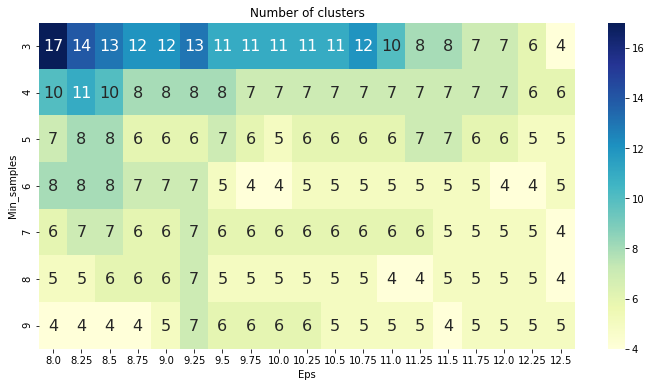
Trong đó:

* a là khoảng cách trung bình tới các ví dụ khác ở trong cùng một cụm
* b là khoảng cách trung bình tới các ví dụ trong cụm gần nhất  
  Giá trị silhouette score nằm trong khoảng từ -1 đến 1
* silhouette score nằm gần 1 nghĩa là ví dụ đang được phân cụm chính xác, xa các cụm khác
* silhouette score gần 0 nghĩa là ví dụ đang nằm gần đường bao của cụm
* silhouette score gần -1 nghĩa là ví dụ đang bị phân sai cụm  
  Chúng ta có thể lấy giá trị silhouette score thông qua hàm silhouette\_score của thư viện scikit-learn.

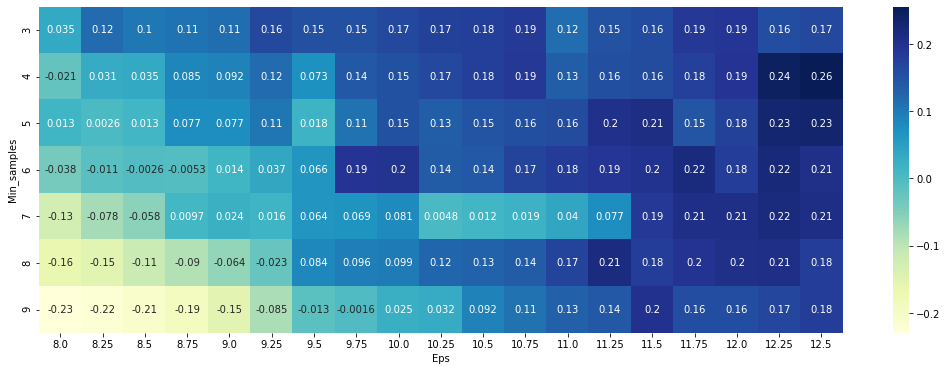
Tiến hành giả định eps và min\_samples sẽ nằm trong một khoảng giá trị từ đó thực hiện tính toán tham số silhouette score. Dựa vào kết quả tính toán, xây dựng heatplot để đưa ra kết luận về eps và min\_samples.



Hình 5. Xác định eps và min\_samples

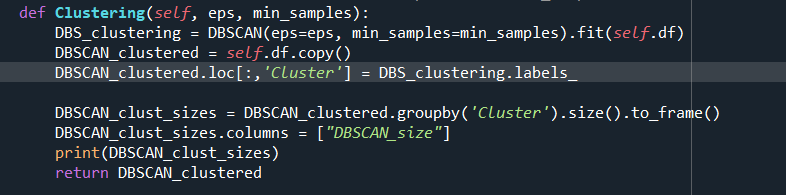


Hình 6. Heatplot xác định số cluster

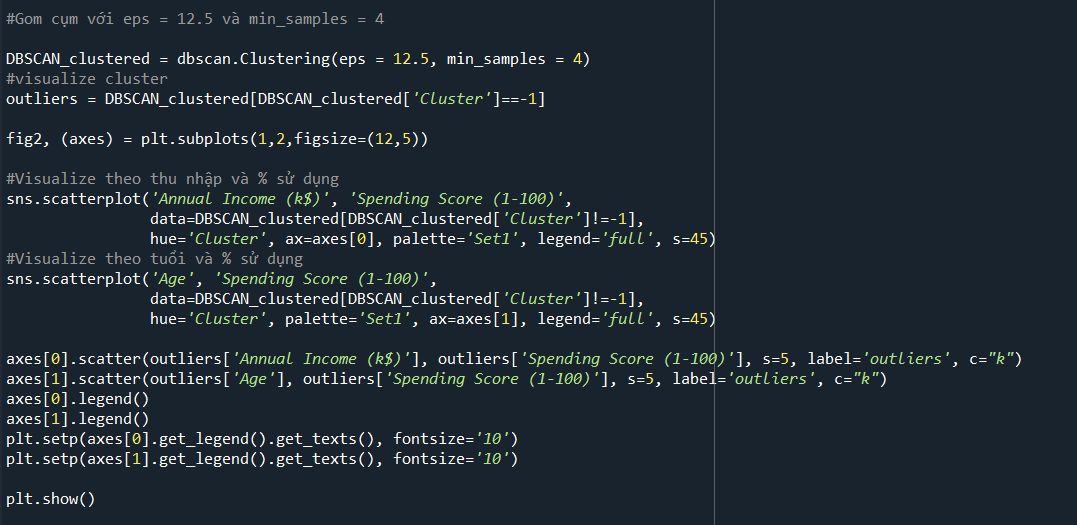


Hình 7. Heatplot xác định eps và min\_samples

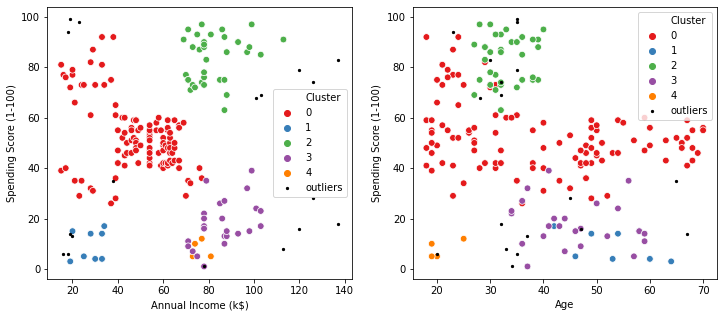
Dựa vào 2 heatplot trên ta thấy với eps = 12.5 và min\_samples = 4 thì tham số silhouette score đạt được giá trị lớn nhất là 0.26 chính vì vậy ta chọn eps = 12.5 và min\_samples = 4 để tiến hành phân cụm.



Hình 8. Hàm phân cụm với 2 tham số eps và min\_samples



Hình 9. Tiến hành phân cụm và visualize kết quả



Hình 10. Visualize kết quả phân cụm với thuật toán DBSCAN

CHƯƠNG 4: KẾT LUẬN

1. Tổng kết

Trong quá trình thực hiện đề tài nhóm đã tiến hành tìm hiểu sâu và rõ hơn về ngành học máy. Nhìn chung nhóm đã đạt được những mục tiêu đề ra ban đầu. Qua đó giúp nhóm em cải thiện được nhiều kỹ năng hơn và có thêm nhiều kinh nghiệm hơn.

Sau thời gian học tập và thực hiện nhóm đã hiểu về:

* Lập trình với ngôn ngữ python.
* Các thuật toán cơ bản trong ngành học máy.
* Biết cách.

1. Hạn chế