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

**KHOA ĐÀO TẠO CHẤT LƯỢNG CAO**

**🕯✡🕮🕮✡🕯**



**BÁO CÁO ĐỒ ÁN 2**

**TÌM HIỂU THUẬT TOÁN**

**HIERARCHICAL CLUSTERING**

**GVHD: Thầy Trần Nhật Quang**

**SVTH:**

**Nguyễn Thanh Tùng 15119154**

**Trương Minh Luân 16110560**

**Tp. Hồ Chí Minh, tháng 5 năm 2019**

# 

MỤC LỤC

[**1.**](#_heading=h.30j0zll) TÌM HIỂU VỀ ĐẠO VĂN

[**Lời cam kết** 1](#_heading=h.1fob9te)

[**2.**](#_heading=h.3znysh7) GIỚI THIỆU THƯ VIỆN SCIKIT-LEARN

[**2.1.1. Giới thiệu** 2](#_heading=h.2et92p0)

[**2.1.2. Lý do dùng Scikit-learn** 2](#_heading=h.tyjcwt)

[**2.1.3. Cài đặt Scikit-learn** 2](#_heading=h.3dy6vkm)

[**2.1.4. Ưu điểm thư viện sklearn** 2](#_heading=h.1t3h5sf)

[**3.** TÌM HIỂU THUẬT TOÁN HIERARCHICAL CLUSTERING](#_heading=h.4d34og8)

[**3.1.1. Khái niệm phân cụm (clustering)** 2](#_heading=h.2s8eyo1)

[**3.1.2. Khái niệm tính khoảng cách giữa các điểm dữ liệu** 3](#_heading=h.17dp8vu)

[**3.1.3. Giới thiệu về phân cụm thứ bậc (hierarchical clustering)** 3](#_heading=h.3rdcrjn)

[**3.1.4. Đánh giá thuật toán** 3](#_heading=h.26in1rg)

[**3.1.5. Project Demo** 3](#_heading=h.lnxbz9)

[**4.** TÀI LIỆU THAM KHẢO 3](#_heading=h.35nkun2)

# 

# TÌM HIỂU VỀ ĐẠO VĂN

Đạo văn không chỉ đơn thuần là việc copy thành quả của một ai đó hay vay mượn các ý tưởng gốc, mà còn rất nhiều biến hóa khác của đạo văn. Tại các quốc gia phát triển họ coi ý tưởng hay sản phẩm trí thức là một tài sản được Pháp luật bảo vệ nên có thể nói đạo văn là một "tội".

Để hiểu rõ về đạo văn một cách rõ ràng hơn ta sẽ tìm hiểu những hình thức đạo văn khác nhau :

* “The Ghost Writer”: Sử dụng toàn bộ công trình của người khác thành của mình
* “The Photocopy”: Sao chép bố cục của các đoạn văn từ một nguồn duy nhất, không hề sửa đổi lại.
* “The Potluck Paper”: Cố gắng “trá hình” việc đạo văn của mình bằng cách sao chép từ nhiều nguồn khác nhau, biên tập lại các câu sao cho nội dung thật hợp lí.
* “The Poor Disguise”: Giữ lại các nội dung quan trọng của nguồn, nhưng người đó vẫn sửa lại một chút về “diện mạo” của bài viết đó bằng cách thay đổi từ khóa hay câu cú.
* “The Labor of Laziness”: Chú giải các nguồn khác nhau và nối chúng lại với nhau, thay vì dành nỗ lực tương tự cho công việc của mình.
* “The Self-Stealer”: Sao chép đáng kể các thành quả trước đó của chính mình để phục vụ cho bài viết/nghiên cứu mới.

Vậy để tránh việc đạo văn ta có những cách sau :

* Tìm hiểu kỹ về vấn đề mà bạn đang muốn nói tới
* Diễn đạt lại nhiều lần bằng các cách khác nhau.
* Nắm được những gì không cần trích dẫn.
* Nêu tên tác giả của ý tưởng đó.
* Trích dẫn nguồn bất cứ khi nào sử dụng lời trích, chú giải một cách chi tiết và cụ thể.

## **Lời cam kết**

“Chúng tôi xin cam đoan đồ án này do chính chúng tôi thực hiện. Chúng tôi không sao chép, sử dụng bất kỳ tài liệu, mã nguồn của người khác mà không ghi rõ nguồn gốc. Chúng tôi xin chịu hoàn toàn trách nhiệm nếu vi phạm.”

# GIỚI THIỆU THƯ VIỆN SCIKIT-LEARN

### **Giới thiệu**

Scikit-learn (sklearn) là thư viện mã nguồn mở về machine learning phổ biến nhất trong cộng đồng Python.

Sklearn chứa hầu hết các thuật toán machine learning từ cơ bản cho đến phức tạp , ta chỉ việc nhét dữ liệu vào, rồi lấy kết quả.

### **Lý do dùng Scikit-learn**

Scikit-learn cho phép ta sử dụng ngay các thuật toán quan trọng một cách đơn giản và hiệu quả.

Scikit-learn còn là một trong những lựa chọn hàng đầu của các researchers và deverlopers.

Trang chủ của thư viện: <http://scikit-learn.org/>

### **Cài đặt Scikit-learn**

Thư viện scikit-learn yêu cầu chúng ta phải cài đặt những module như dưới đây:

* Python
* NumPy
* SciPy

Đầu tiên ta phải cài đặt Python ở trên máy.

Sau khi đã có Python trên máy của mình, ta có thể cài đặt 2 gói NumPy và SciPy bằng lệnh :

**$ pip install numpy scipy matplotlib scikit-learn pandas**

### **Ưu điểm thư viện sklearn**

⬥ Hàm ngắn gọn, dễ hiểu

⬥ Xử lý được lượng dữ liệu lớn

⬥ Độ chính xác và tin cậy cao

⬥ Tốc độ tính toán rất nhanh

⬥ Hỗ trợ rất nhiều thuật toán Machine Learning đơn giản và phức tạp

# TÌM HIỂU THUẬT TOÁN HIERARCHICAL CLUSTERING

* + 1. **Khái niệm phân tích cụm (clustering)**

Trong tự nhiên, dữ liệu thường không được gán nhãn sẵn.

Phân tích cụm là phân loại các đối tượng không có nhãn sao cho các đối tượng có các đặc tính tương tự nhau vào cùng một cụm với nhau.

Các đặc tính để phân cụm là các biến có khả năng mô tả được sự giống nhau giữa các đối tượng tùy theo lĩnh vực của dữ liệu mà ta phân tích.

Vì các thuật toán phân tích cụm thường dựa trên cơ sở khoảng cách nên sau đây ta tìm hiểu về một số khái niệm thước đo khoảng cách.

* + 1. **Khái niệm tính khoảng cách giữa các điểm dữ liệu**

Khoảng cách Euclid :

Là khoảng cách được tính bằng công thức Pytago. Ví dụ khoảng cách giữa 2 điểm p và q.



Khoảng cách Manhattan:

Là một dạng khoảng cách giữa hai điểm trong không gian Euclid với hệ tọa độ Descartes. Đại lượng này được tính bằng tổng chiều dài của hình chiếu của đường thẳng nối hai điểm này trong hệ trục tọa độ Descartes.

Ví dụ, khoảng giữa hai điểm p có tọa độ (x1,y1) và điểm q có tọa độ (x2,y2) là |x1-x2|+|y1-y2|.

Khoảng cách Chebychev:

Là khoảng cách giữa hai vector là lớn nhất so với bất kì hiệu tọa độ thành phần của chúng.

Ví dụ khoảng cách giữa hai véc tơ p và q, với tọa độ p và q là:

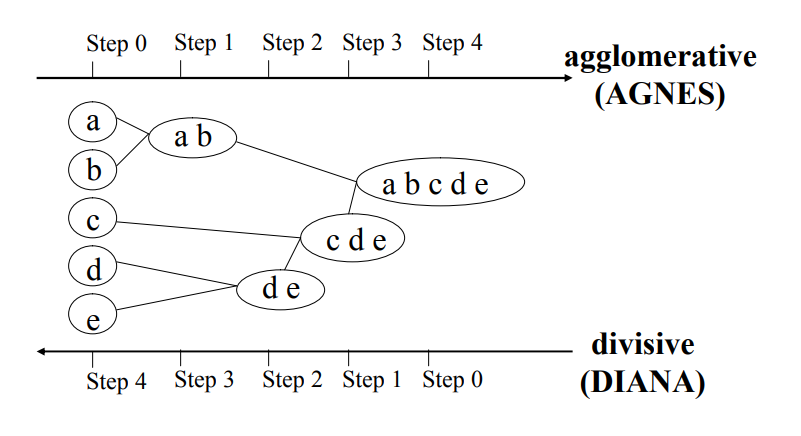


* + 1. **Giới thiệu về phân cụm thứ bậc (hierarchical clustering)**

Phân cụm thứ bậc là thuật toán được xây dựng theo cấu trúc thứ bậc dạng hình cây. Có 2 phương pháp chính đó là: Agglomerative Approach và Divisive Approach.

Agglomerative Aapproach (bottom up approach): chúng ta xem mỗi đối tượng ban đầu là 1 cụm (cluster) và gom 2 đối tượng gần nhất lại với nhau thành 1 cluster.

Divisive Approach (top down approach): Ngược lại với Agglomerative Approach, là ta xem tất cả các đối tượng thuộc cùng 1 cluster, sau đó tách thành 2 nhóm con (thường dựa vào khoảng cách lớn nhất). Quá trình này được thực hiện cho đến khi mỗi cụm chỉ còn 1 đối tượng.



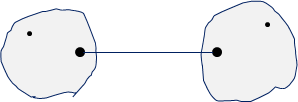
Các bước trong phương pháp phân cụm Agglomerative Approach như sau:

1. Chuyển đổi các đặc trưng (thuộc tính - Features) của đối tượng (objects) vào ma trận khoảng cách
2. Xem mỗi đối tượng ban đầu là một cluster
3. Lặp lại 2 bước sau cho đến khi chỉ còn 1 cluster duy nhất

* Gộp 2 cluster gần nhất
* Cập nhật ma trận khoảng cách

Tính liên kết giữa các đối tượng (linkages):

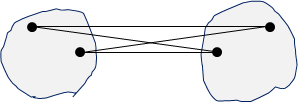
Single Linkage: Khoảng cách giữa 2 cluster Là khoảng cách giữa 2 đối tượng gần nhau nhất trong 2 clusters.



Complete Linkage: Là khoảng cách giữa 2 đối tượng xa nhau nhất trong 2 clusters.



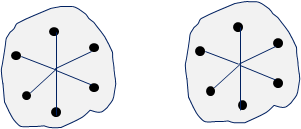
Average Group: Là khoảng cách trung bình giữa các đối tượng trong 2 cluster.



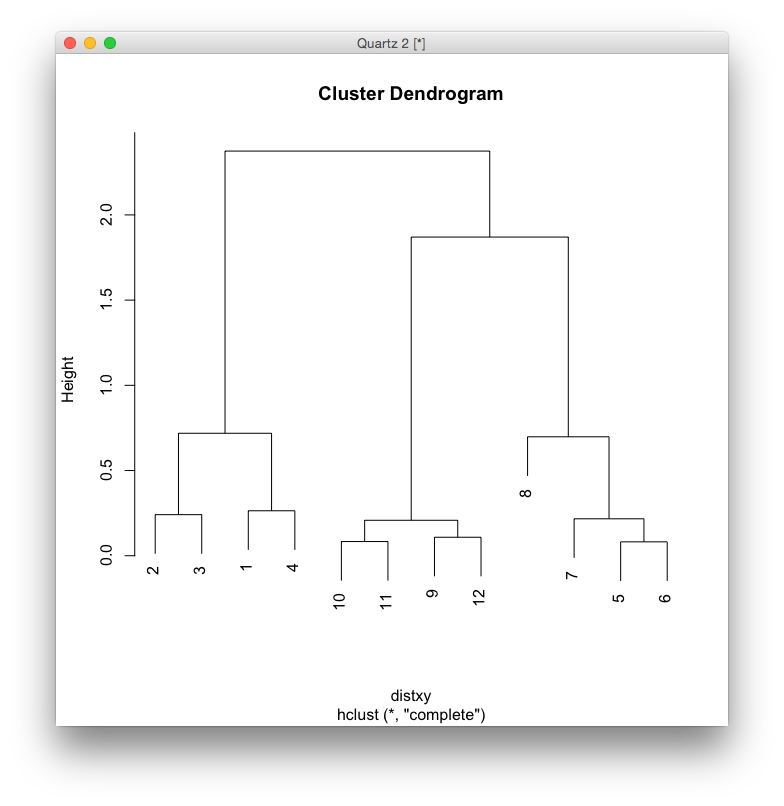
Phương sai theo thủ tục Ward:

Tính giá trị trung bình tất cả các đối tượng trong từng cụm.

Tính khoảng cách theo thước đo Euclid giữa các phần tử trong cụm với giá trị trung bình của cụm.



Quá trình phân cụm theo phương pháp Agglomerative Aapproach ta có sơ đồ dạng cây như hình sau :



**Đánh giá thuật toán**

**Ưu điểm**

* Giải thuật đơn giản, dễ hiểu
* Không cần tham số k cluster. Xác định số nhóm cần thiết bằng việc cắt ngang sơ đồ hình cây tại mức thích hợp.

**Nhược điểm**

Tính co dãn thấp : Độ phức tạp là O(n^2) với n – số đối tượng.

* Không thể quay lui về bước trước.
* Khó xác định phương pháp tích tụ hay chia nhỏ.
* Nhạy cảm với nhiễu, cá biệt.
* Gặp vấn đề khi các nhóm có kích thước khác nhau.
* Có xu hướng phân chia các nhóm dữ liệu lớn.

* + 1. **Project Demo**

**Bước 1** : Import thư viện

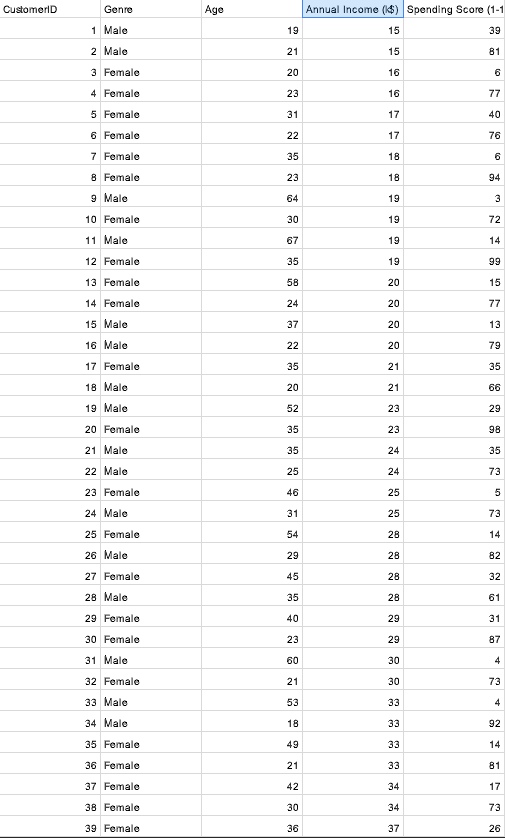
**import** matplotlib.pyplot **as** plt  
**import** matplotlib.cm **as** cm  
**import** pandas **as** pd  
**import** matplotlib  
**import** numpy **as** np  
**import** scipy.cluster.hierarchy **as** shc *#thư viện phân cấp thứ bậc***from** sklearn.cluster **import** AgglomerativeClustering *# dùng sklean để gom nhóm các cluster*

**Bước 2**: Đọc file dữ liệu

dataset = pd.read\_csv(**"shopping\_data.csv"**)  
  
dataset.shape *#trả về ma trận (200, 5) nghĩa là 200 dòng dữ liệu và 5 feature.*

dataset = pd.read\_csv(**"shopping\_data.csv"**)  
  
dataset.shape *#trả về ma trận (200, 5) nghĩa là 200 dòng dữ liệu và 5 feature.*

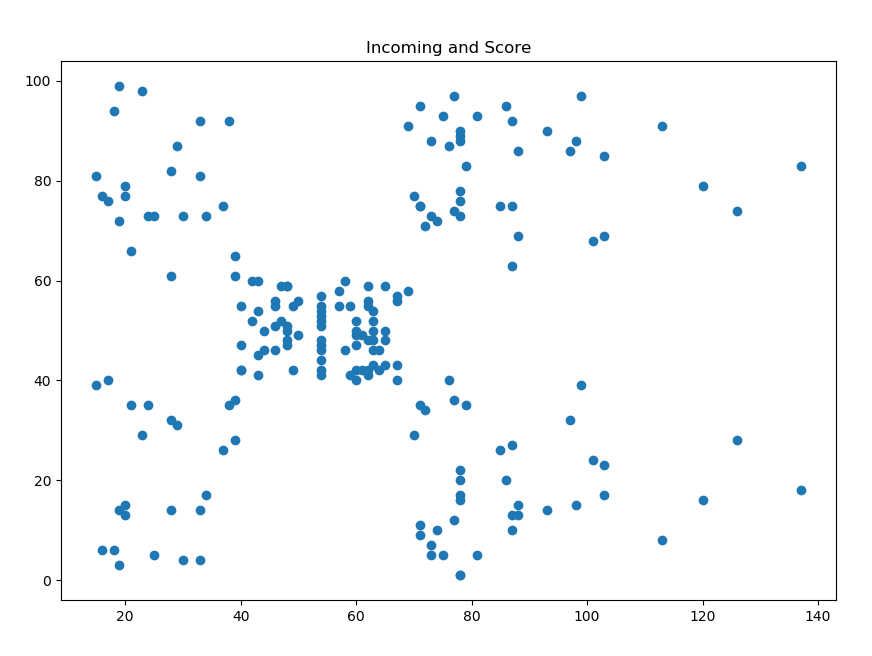
<https://drive.google.com/file/d/18Dsja5_1jRY1GnWoORXFFKGTJhylk6rJ/view>

Mô tả dataset : Khảo sát mua sắm của khách hàng dựa trên các feature : Giới tính, Thu nhập hằng năm, Điểm tích lũy mua sắm

Có tổng cộng 200 dòng dữ liệu và 5 feature

**Bước 3**: Chọn ra 2 feature để phân loại trong project này ta chọn Annual Icome và Spending Score và sau đó vẽ biểu đồ thể hiện 2 feature

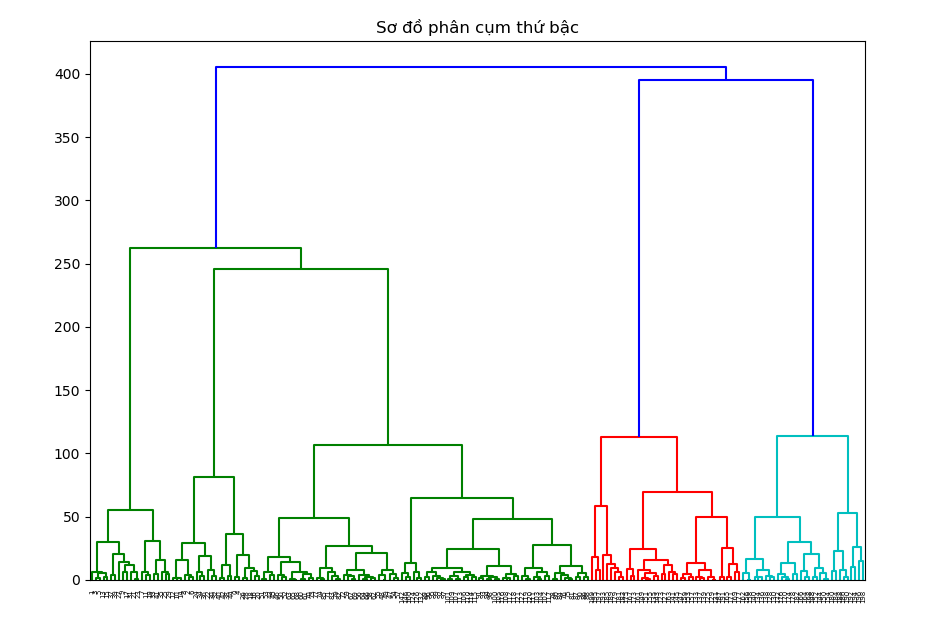
dataset.shape *#trả về ma trận (200, 5) nghĩa là 200 dòng dữ liệu và 5 feature.*data = dataset.iloc[:, 3:5].values *#Lấy từ cột index thứ 3 đến trước index 5  
#Phân loại khách hàng theo thu nhập hằng năm(Annual Income) và điểm mua sắm (spending score)*

**Bước 4:** Tính khoảng cách giữa các điểm dữ liệu theo thước đo Euclid và phương pháp xét khoảng cách giữa các cụm là ward linkage

dend = shc.dendrogram(shc.linkage(data, method=**'ward'**))

Sau đó ta được sơ đồ dạng cây

plt.figure(figsize=(10, 7))  
plt.title(**"Sơ đồ phân cụm thứ bậc"**)  
dend = shc.dendrogram(shc.linkage(data, method=**'ward'**))  
plt.show()

**Bước 5:** Dựa vào biểu đồ ta thiết kế mô hình thuật toán

cluster = AgglomerativeClustering(n\_clusters=5, affinity=**'euclidean'**, linkage=**'ward'**)

Dựa vào biểu đồ ta xác định có 5 nhóm cần phân loại bằng cách cắt ngang sơ đồ hình cây tại mức dài nhất của đường màu xanh.

Thước đo khoảng cách là euclidean

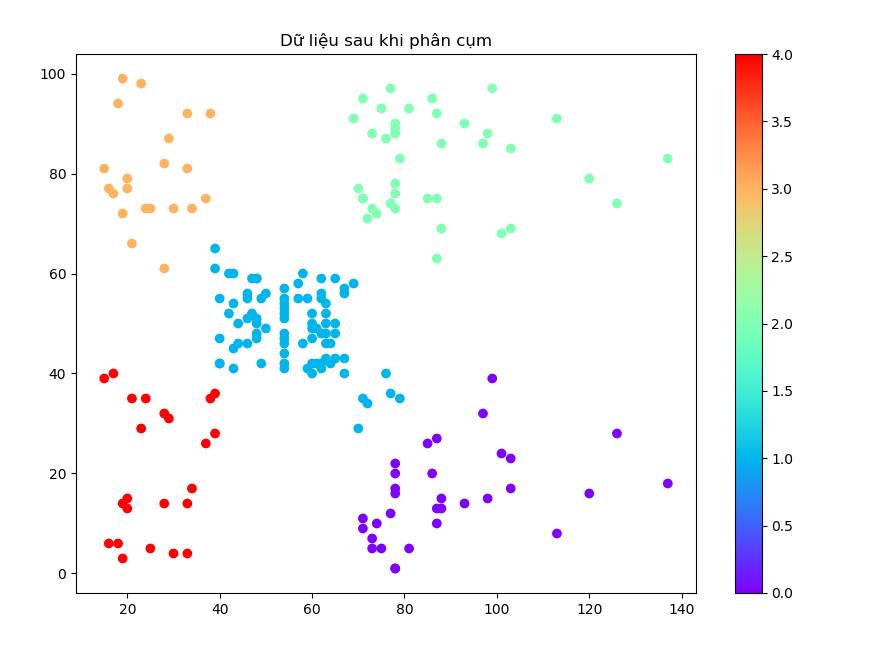
Phương pháp tính khoảng cách giữa các cluster : ward linkage

**Bước 6:** Đưa dữ liệu vào và tiến hành phân loại và dự đoán

cluster.fit\_predict(data)

**Bước 7:** Vẽ lại đồ thị biểu diễn dữ liệu sau khi phân loại

plt.figure(figsize=(10, 7))  
plt.title(**"Dữ liệu sau khi phân cụm"**)  
plt.scatter(data[:,0], data[:,1], c=cluster.labels\_, cmap=**'rainbow'**)  
plt.colorbar()  
plt.show()



# TÀI LIỆU THAM KHẢO

Tham khảo tại các trang web:

<https://www.datanovia.com/en/lessons/divisive-hierarchical-clustering/>

<http://bis.net.vn/forums/p/571/1485.aspx>

<https://ongxuanhong.wordpress.com/2015/08/27/gom-nhom-clustering-analysis-tap-du-lieu-labor/>