|  |
| --- |
| **TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI**  **KHOA CÔNG NGHỆ THÔNG TIN**  ---------\*\*\*--------  Trường Đại học Giao thông Vận tải – Wikipedia tiếng Việt  **BÁO CÁO BÀI TẬP LỚN**  **Ứng dụng thuật toán k-means vào bài toán phân loại rượu**  Lớp: CNTT Việt-Anh 1 K61  Giảng viên: Gv. Nguyễn Quốc Tuấn  *Nhóm sinh viên thực hiện: Nhóm 8 Mã sinh viên*  **Nguyễn Điển Đăng 202636244**  Phạm Anh Quân 202602952  Hoàng Thị Linh 202602945  Đinh Thị Hường 202602939  Kiều Hồng Quang 202605718 |

**MỤC LỤC**

**CHƯƠNG 1 TỔNG QUAN VỀ PHÂN CỤM 7**

1.1 Phân cụm là gì và mục đích của phân cụm 7

1.2. Các Phương pháp phân cụm 8  
 1.3. Phân cụm, phân hoạch 8  
[**CHƯƠNG 2 THUẬT TOÁN K-MEANS CLUSTERING TRONG BÀI TOÁN**](#_Toc14315)

[**PHÂN CỤM** 7](#_Toc14316)

[1.1 Tổng quan về thuật toán K-Means Clustering 7](#_Toc14317)

[1.2. Thuật toán K-Means Clustering: 8](#_Toc14318)

[1.2.1. Mô hình toán học: 8](#_Toc14319)

[1.2.2. Độ chính xác của thuật toán: 9](#_Toc14320)

[1.2.3. Nghiệm của thuật toán K-Means Clustering: 9](#_Toc14321)

[1.2.4. Tóm tắt thuật toán: 10](#_Toc14322)

[**CHƯƠNG 3 ỨNG DỤNG THUẬT TOÁN K-MEANS CLUSTERING** 11](#_Toc14323)

[**2.1. Giới thiệu về bài toán:** 11](#_Toc14324)

[2.2. Dữ liệu wine\_red\_quality: 14](#_Toc14325)

[2.2.1. Phát biểu bài toán: 14](#_Toc14326)

[2.2.2. Yêu cầu: 14](#_Toc14327)

[2.2.3. Bộ dữ liệu: 14](#_Toc14328)

[2.2.4. Tiến hành phân cụm 16](#_Toc14329)

**2.3. Ví dụ trên python [11](#_Toc14324)**

[2.3.1 Giới thiệu bài toán: 14](#_Toc14325)

[2.3.2. Hiển thị dữ liệu trên đồ thị 14](#_Toc14326)

[2.3.3. Các hàm số cần thiết cho K-means clustering: 14](#_Toc14327)

[2.3.4. Kết quả tìm được bằng thư viện scikit-learn: 14](#_Toc14328)

[**CHƯƠNG 4 ƯU ĐIỂM & HẠN CHẾ CỦA THUẬT TOÁN** ….………………14](#_Toc14323)  
**TÀI LIỆU THAM KHẢO**.....................................................................................19

**LỜI MỞ ĐẦU**

Công nghệ ngày càng phổ biến và không ai có thể phủ nhận được tầm quan trọng và những hiệu quả mà nó đem lại cho cuộc sống chúng ta. Bất kỳ trong lĩnh vực nào, sự góp mặt của trí tuệ nhân tạo sẽ giúp con người làm việc và hoàn thành tốt công việc hơn. Và gần đây, một thuật ngữ “machine learning” rất được nhiều người quan tâm.Thay vì phải code phần mềm với cách thức thủ công theo một bộ hướng dẫn cụ thể nhằm hoàn thành một nhiệm vụ đề ra thì máy sẽ tự “học hỏi” bằng cách sử dụng một lượng lớn dữ liệu cùng những thuật toán cho phép nó thực hiện các tác vụ.

Đây là một lĩnh vực khoa học tuy không mới, nhưng cho thấy lĩnh vực trí tuệ nhân tạo đang ngày càng phát triển và có thể tiến xa hơn trong tương lai. Đồng thời, thời điểm này nó được xem là một lĩnh vực “nóng” và dành rất nhiều mối quan tâm để phát triển nó một cách mạnh mẽ, bùng nổ hơn.

Hiện nay, việc quan tâm machine learning càng ngày càng tăng lên là vì nhờ có machine learning giúp gia tăng dung lượng lưu trữ các loại dữ liệu sẵn, việc xử lý tính toán có chi phí thấp và hiệu quả hơn rất nhiều.

Những điều trên được hiểu là nó có thể thực hiện tự động, nhanh chóng để tạo ra những mô hình cho phép phân tích các dữ liệu có quy mô lớn hơn và phức tạp hơn đồng thời đưa ra những kết quả một cách nhanh và chính xác hơn.

Chính sự hiệu quả trong công việc và các lợi ích vượt bậc mà nó đem lại cho chúng ta khiến machine learning ngày càng được chú trọng và quan tâm nhiều hơn. Vì vậy chúng em đã chọn đề tài ”Ứng dụng thuật toán K-Means Clustering để dự đoán chất lượng của rượu vang đỏ.

Chúng em xin chân thành gửi lời cảm ơn tới thầy Vũ Văn Định đã tận tình giảng dạy, truyền đạt cho chúng em những kiến thức cũng như kinh nghiệm quý báu trong suốt quá trình học. Thầy đã tận tình theo sát giúp đỡ, trực tiếp chỉ bảo, hướng dẫn trong suốt quá trình nghiên cứu và học tập của chúng em.

**CHƯƠNG 1: TỔNG QUAN PHƯƠNG PHÁP PHÂN CỤM**

* 1. **Phân cụm là gì và mục đích của phân cụm**

Phân cụm là quá trình tổ chức các đối tượng hoặc dữ liệu vào các nhóm (hoặc cụm) dựa trên sự tương tự giữa chúng. Mục tiêu chính của phân cụm là tạo ra các nhóm có tính tương đồng cao bên trong và khác biệt lớn giữa các nhóm. Dưới đây là một số mục đích quan trọng của phân cụm:

Tóm tắt dữ liệu: Phân cụm giúp biểu diễn một tập dữ liệu lớn dưới dạng các nhóm hoặc cụm nhỏ, từ đó giúp đơn giản hóa và tóm tắt thông tin phức tạp.

Khám phá tri thức: Phân cụm có thể giúp bạn khám phá các mối quan hệ, sự tương đồng và sự khác biệt giữa các đối tượng hoặc dữ liệu. Điều này có thể dẫn đến hiểu biết sâu hơn về dữ liệu và dự đoán xu hướng hoặc kết quả.

Phát hiện bất thường: Phân cụm có thể giúp xác định các nhóm không phù hợp hoặc bất thường trong dữ liệu, đặc biệt trong việc phát hiện gian lận, lỗi hoặc các tình huống đặc biệt.

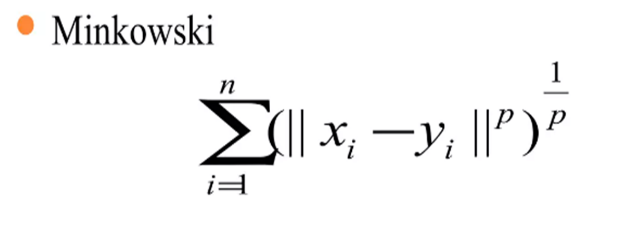
Hỗ trợ quyết định: Phân cụm có thể hỗ trợ quyết định bằng cách chia dữ liệu thành các nhóm có tính chất tương tự, giúp người ra quyết định hiểu rõ hơn về các biến thể và tùy chọn có sẵn.

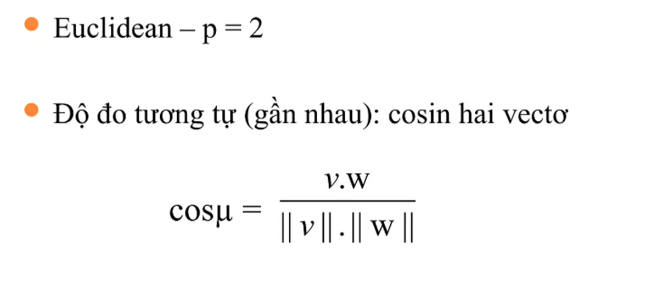
Tiết kiệm thời gian và nguồn lực: Khi dữ liệu được phân cụm, bạn có thể tập trung vào từng nhóm một thay vì phải xử lý toàn bộ tập dữ liệu, giúp tiết kiệm thời gian và nguồn lực.

Tạo ra hệ thống tự động: Phân cụm có thể được sử dụng để xây dựng các hệ thống tự động, chẳng hạn như hệ thống gợi ý sản phẩm dựa trên sở thích của người dùng.

Những mục đích này thể hiện sự quan trọng của phân cụm trong phân tích dữ liệu

và đưa ra những quyết định thông minh dựa trên dữ liệu.





* 1. **Các Phương pháp phân cụm**
* **Phương pháp phân cụm phân vùng**

Nhằm phân 1 tập dữ liệu có n phần tử cho trước thành k nhóm dữ liệu sao cho: mỗi phần tử dữ liệu chỉ thuộc về một nhóm dữ liệu và mỗi nhóm dữ liệu có tối thiểu ít nhất 1 phần tử dữ liệu. Các thuật toán phân vùng cố gắng cải tiến tiêu chuẩn phân cụm bằng cách tính các giá trị đo độ tương tự giữa các đối tượng dữ liệu và sắp xếp các giá trị này, sau đó thuật toán lựa chọn một giá trị trong dãy sắp xếp sao cho hàm tiêu chuẩn đạt giá trị tối thiểu (ý tưởng chính là sử dụng chiến lược ăn tham Greedy để tìm kiếm nghiệm)

Một số thuật toán: K-means, Pam, Clara,..

* **Phương pháp phân cụm phân cấp**

Phương pháp này xây dựng 1 phân cấp trên cơ sở các đối tượng dữ liệu đang xem xét. Nghĩa là sắp xếp một tập dữ liệu đã cho thành một cấu trúc có dạng hình cây, cây phân cấp này được xây dựng theo kỹ thuật đệ quy.

* **Phương pháp phân cụm dựa trên mật độ**

Phương pháp này nhóm các đối tượng theo hàm mật độ xác định. Mật độ được định nghĩa như là số các đối tượng lân cận của 1 đối tượng dữ liệu theo 1 ngưỡng nào đó. Một số thuật toán dựa trên mật độ: OPTICS, DBSCAN,..

* **Phương pháp phân cụm dựa trên lưới**

Đây là phương pháp dựa trên cấu trúc dữ liệu lưới để phân cụm dữ liệu, phương pháp này chủ yếu tập trung và áp dụng cho lớp dữ liệu không gian.

Các phương pháp dựa trên lưới lượng tử hóa không gian đối tượng thành một không gian hữu hạn số lượng ô tạo thành cấu trúc lưới. Tất cả các hoạt động phân cụm được thực hiện trên cấu trúc lưới.Một số thuật toán phân cụm dữ liệu dựa trên cấu trúc lưới: STING, WaveCluster,..

* 1. **Phân cụm, phân hoạch**

Phân cụm và phân hoạch là hai khái niệm liên quan đến việc tổ chức và nhóm hóa dữ liệu, nhưng chúng có mục tiêu và phương pháp khác nhau. Dưới đây là sự khác biệt giữa phân cụm và phân hoạch:

* Phân cụm (Clustering):

Mục tiêu: Phân cụm tập dữ liệu thành các nhóm hoặc cụm sao cho các đối tượng trong cùng một nhóm tương tự nhau, và các đối tượng ở các nhóm khác biệt lớn nhau.

Phương pháp: Phân cụm không yêu cầu trước thông tin về số lượng nhóm và không cần nhãn (label) dữ liệu. Phương pháp này thường sử dụng độ đo tương tự (similarity measure) như khoảng cách để xác định sự tương đồng giữa các đối tượng.

* Phân hoạch (Partitioning):

Mục tiêu: Phân chia tập dữ liệu thành các nhóm, nhưng trong trường hợp này, số lượng nhóm cần được xác định trước.

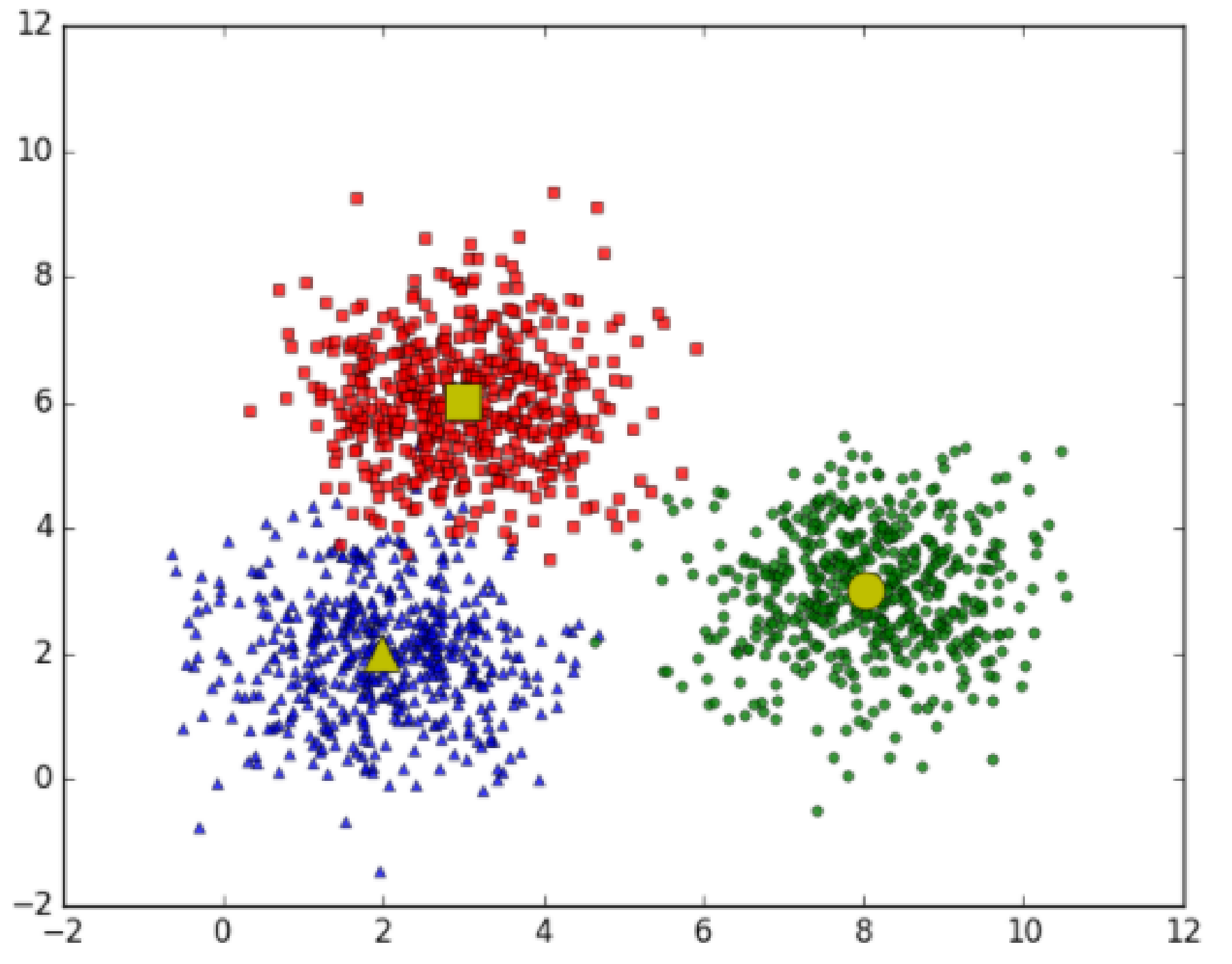
Phương pháp: Phân hoạch yêu cầu xác định số lượng nhóm (k) trước khi thực hiện phân chia. Phương pháp phân hoạch phổ biến nhất là K-Means, trong đó mỗi đối tượng thuộc vào một nhóm dựa trên khoảng cách tới trọng tâm (centroid) của nhóm.

Tóm lại, phân cụm tập trung vào việc tìm các nhóm dựa trên sự tương đồng tự nhiên giữa các đối tượng mà không cần biết trước số lượng nhóm, trong khi phân hoạch đặt ra một số lượng nhóm cố định và phân chia dữ liệu thành chúng dựa trên khoảng cách.

**CHƯƠNG 2: THUẬT TOÁN K-MEANS CLUSTERING TRONG BÀI TOÁN** **PHÂN CỤM**

**2.1 Tổng quan về thuật toán K-Means Clustering**

Với thuật toán K-Means Clustering, chúng ta không biết nhãn (label) của từng điểm dữ liệu. Mục đích là làm thể nào để phân dữ liệu thành các cụm (cluster) khác nhau sao cho dữ liệu trong cùng một cụm có tính chất giống nhau. Ý tưởng đơn giản nhất về cluster (cụm) là tập hợp các điểm ở gần nhau trong một không gian nào đó (không gian này có thể có rất nhiều chiều trong trường hợp thông tin về một điểm dữ liệu là rất lớn). Hình bên dưới là một ví dụ về 3 cụm dữ liệu (từ giờ tôi sẽ viết gọn là cluster).



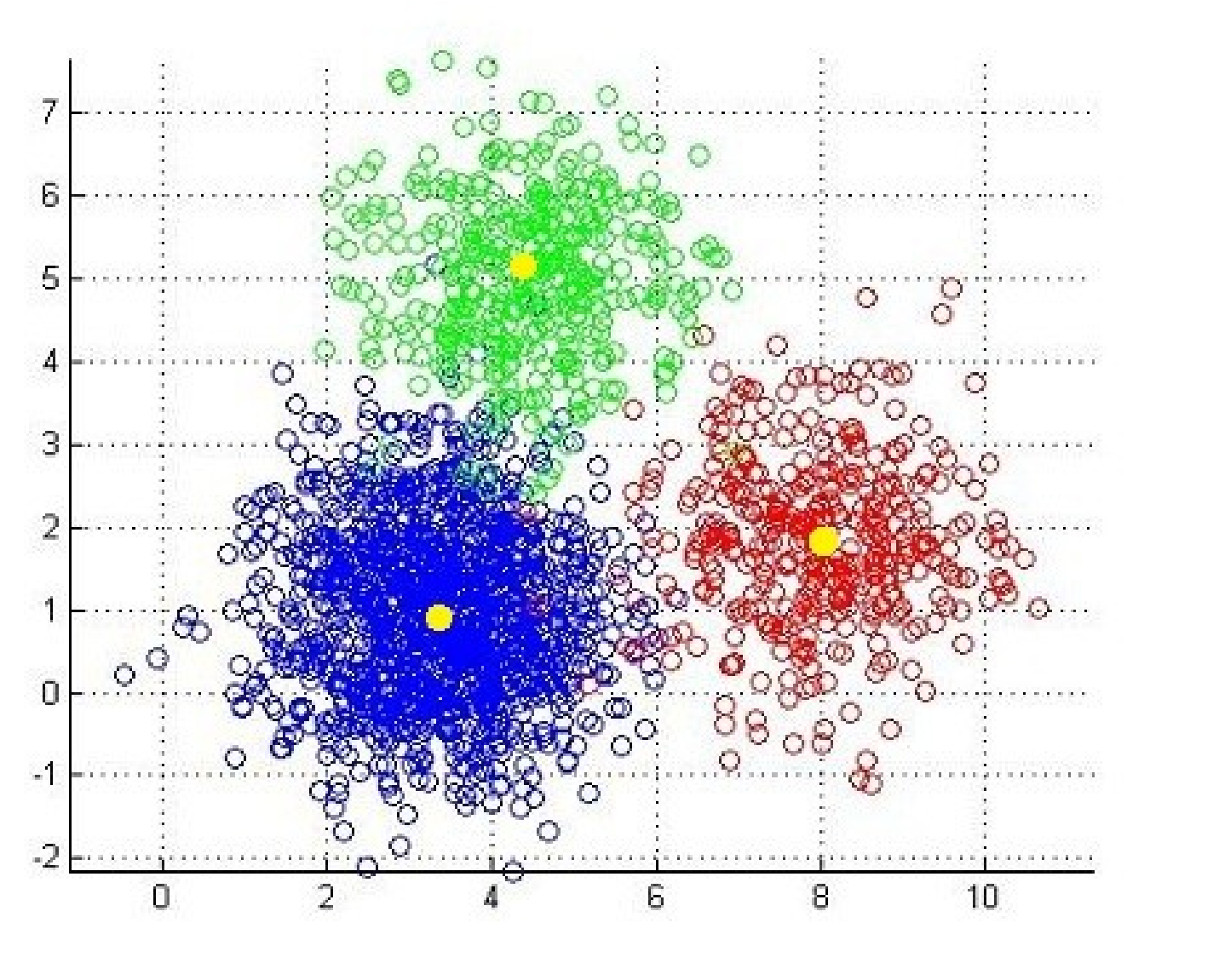
*Hình 2.1: Bài toán với 3 clusters*

Giả sử mỗi cluster có một điểm đại diện (*center*) màu vàng. Và những điểm xung quanh mỗi center thuộc vào cùng nhóm với center đó. Một cách đơn giản nhất, xét một điểm bất kỳ, ta xét xem điểm đó gần với center nào nhất thì nó thuộc về cùng nhóm với center đó.

**2.2. Thuật toán K-Means Clustering:**

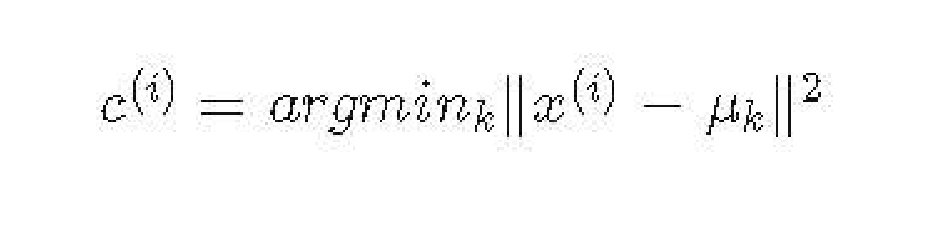
*2.2.1. Mô hình toán học:*

Ta gọi điểm tại vị trí trung bình của tất cả các điểm dữ liệu trong một cụm là **trung tâm cụm**. Như vậy, nếu có K cụm thì sẽ có K trung tâm cụm và mỗi trung tâm cụm sẽ nằm gần các điểm dữ liệu trong cụm tương ứng hơn các trung tâm cụm khác. Trong hình dưới đây, K = 3 và ta có 3 trung tâm cụm là các điểm màu vàng.

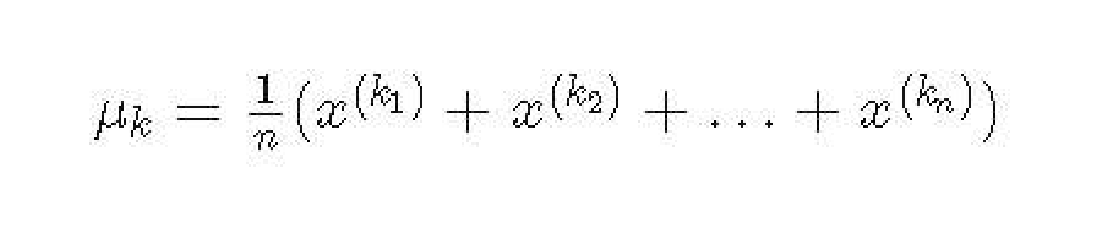


*Hình 2.2: Mô hình dữ liệu được phân cụm*

Để phân cụm dữ liệu bằng K-Means Clustering, trước hết ta chọn K là số cụm để phân chia và chọn ngẫu nhiên K trong số m dữ liệu ban đầu làm trung tâm cụmμ1, μ2, …, μK. Sau đó, với điểm dữ liệu x(i) ta sẽ gán nó cho cụm c(i) là cụm có trung tâm cụm gần nó nhất.



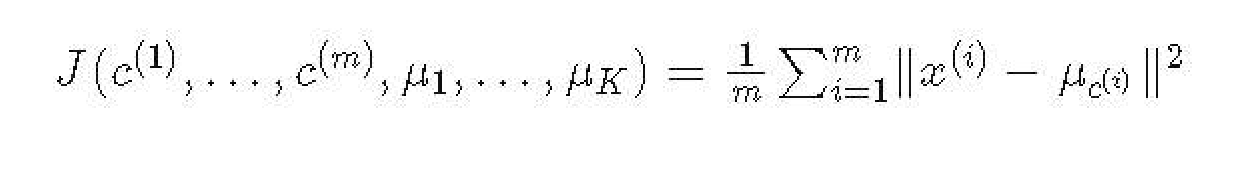
Khi tất cả các điểm dữ liệu đã được gán về các cụm, bước tiếp theo là tính toán lại vị trí các trung tâm cụm bằng trung bình tọa độ các điểm dữ liệu trong cụm đó.



với k1, k2, …, kn là chỉ số các dữ liệu thuộc cụm thứ k. Các bước trên được lặp lại cho tới khi vị trí các trung tâm cụm không đổi sau một bước lặp nào đó.

*2.2.2. Độ chính xác của thuật toán:*

Hàm mất mát của thuật toán K-Means Clustering đặc trưng cho độ chính xác của nó sẽ càng lớn khi khoảng cách từ mỗi điểm dữ liệu tới trung tâm cụm càng lớn.



*2.2.3. Nghiệm của thuật toán K-Means Clustering:*

Trong các bước của thuật toán, thực chất bước gán các điểm dữ liệu về trung tâm cụm gần nhất và bước thay đổi trung tâm cụm về vị trí trung bình của các điểm dữ liệu trong cụm đều nhằm mục đích giảm hàm mất mát. Thuật toán kết thúc khi vị trí các trung tâm cụm không đổi sau một bước lặp nào đó. Khi đó hàm mất mát đạt giá trị nhỏ nhất.

Khi K càng nhỏ so với m, thuật toán càng dễ đi đến kết quả chưa phải tối ưu. Điều này phụ thuộc vào cách chọn K trung tâm cụm ban đầu.

Để khắc phục điều này, ta cần lặp lại thuật toán nhiều lần và chọn phương án có giá trị hàm mất mát nhỏ nhất.

*2.2.4. Tóm tắt thuật toán:*

**Đầu vào:** Dữ liệu XX và số lượng cluster cần tìm KK.

**Đầu ra:** Các center M và label vector cho từng điểm dữ liệu Y.

1. Chọn K điểm bất kỳ làm các center ban đầu.
2. Phân mỗi điểm dữ liệu vào cluster có center gần nó nhất.
3. Nếu việc gán dữ liệu vào từng cluster ở bước 2 không thay đổi so với vòng lặp trước nó thì ta dừng thuật toán.
4. Cập nhật center cho từng cluster bằng cách lấy trung bình cộng của tất các các điểm dữ liệu đã được gán vào cluster đó sau bước 2.
5. Quay lại bước 2.

2.2.5. Ví dụ

Giả sử ta có 4 loại thuốc A,B,C,D, mỗi loại thuộc được biểu diễn bởi 2 đặc trưng X và Y như sau. Mục đích của ta là nhóm các thuốc đã cho vào 2 nhóm (K=2) dựa vào các đặc trưng của chúng

|  |  |  |
| --- | --- | --- |
| Object | Thuộc tính (X1): w index | Thuộc tính (X2): pH |
| Medicine A | 1 | 1 |
| Medicine B | 2 | 1 |
| Medicine C | 4 | 3 |
| Medicine D | 5 | 4 |

**Các bước thực hiện thuật toán :**

**B1*.*** Khởi tạo tâm (centroid) cho 2 nhóm. Giả sử ta chọn A là tâm của nhóm thứ nhất (tọa độ tâm nhóm thứ nhất A(1,1)) và B là tâm của nhóm thứ 2 (tạo độ tâm nhóm thứ hai

B(2,1)).

B2. Tính khoảng cách từ các đối tượng đến tâm của các nhóm (Khoảng cách Euclidean)

**B3*.*** Nhóm các đối tượng vào nhóm gần nhất

**B4*.*** Tính lại tọa độ các tâm cho các nhóm mới dựa vào tọa độ của các đối tượng trong nhóm.

Lặp lại quá trình cho đến khi tâm đạt tới giá trị gần như không đổi.

Áp dụng các bước ta được bảng sau :

|  |  |  |  |
| --- | --- | --- | --- |
| Lặp 1 | A( 1,1) | B(2,1) | Cụm |
| A(1,1) | 0 | 1 | 1 |
| B(2,1) | 1 | 0 | 2 |
| C(4,3) | 3,61 | 2,83 | 2 |
| D(5,4) | 5 | 4,24 | 2 |

Cụm 1 : A(1,1) => T1(1,1).

Cụm 2 : B(2,1), C(4,3),D(5,4) => T2 =(

|  |  |  |  |
| --- | --- | --- | --- |
| Lặp 2 | T1( 1,1) | T2 | Cụm |
| A(1,1) | 0 | 3,14 | 1 |
| B(2,1) | 1 | 2,357 | 1 |
| C(4,3) | 3,61 | 0,47 | 2 |
| D(5,4) | 5 | 1,89 | 2 |

Tính lại tâm mới :

Cụm 1 : A(1,1), B(2,1) => T1 =( = ()

Cụm 2 : B(2,1), C(4,3),D(5,4) => T2 = () = ()

|  |  |  |  |
| --- | --- | --- | --- |
| Lặp 3 | T1() | T2() | Cụm |
| A(1,1) | 0,5 | 4,3 | 1 |
| B(2,1) | 0,5 | 3,5 | 1 |
| C(4,3) | 3,201 | 0,707 | 2 |
| D(5,4) | 4,609 | 0,707 | 2 |

Tính lại tâm mới :

Cụm 1 : A(1,1), B(2,1) => T1 =( = ()

Cụm 2 : B(2,1), C(4,3),D(5,4) => T2 = () = ()

Ta thấy giá trị T1 và T2 không đổi => dừng thuật toán .

Kết quả :

|  |  |  |  |
| --- | --- | --- | --- |
| Object | Thuộc tính (X1): w index | Thuộc tính (X2): pH | Class |
| Medicine A | 1 | 1 | 1 |
| Medicine B | 2 | 1 | 1 |
| Medicine C | 4 | 3 | 2 |
| Medicine D | 5 | 4 | 2 |

s

**CHƯƠNG 3: ỨNG DỤNG THUẬT TOÁN K-MEANS CLUSTERING**

***3.1. Phát biểu bài toán:***

Bài toán phân loại chất lượng của rượu vang đỏ đưa ra các tập dự liệu là các thông số của từng mẫu rượu vang đỏ để phân loại chất lượng của chúng

* Giá trị input: Thông tin, đặc tính của một mẫu rượu
* Giá trị output: tên của cụm chúng được phân vào

***3.1.1. Yêu cầu:***

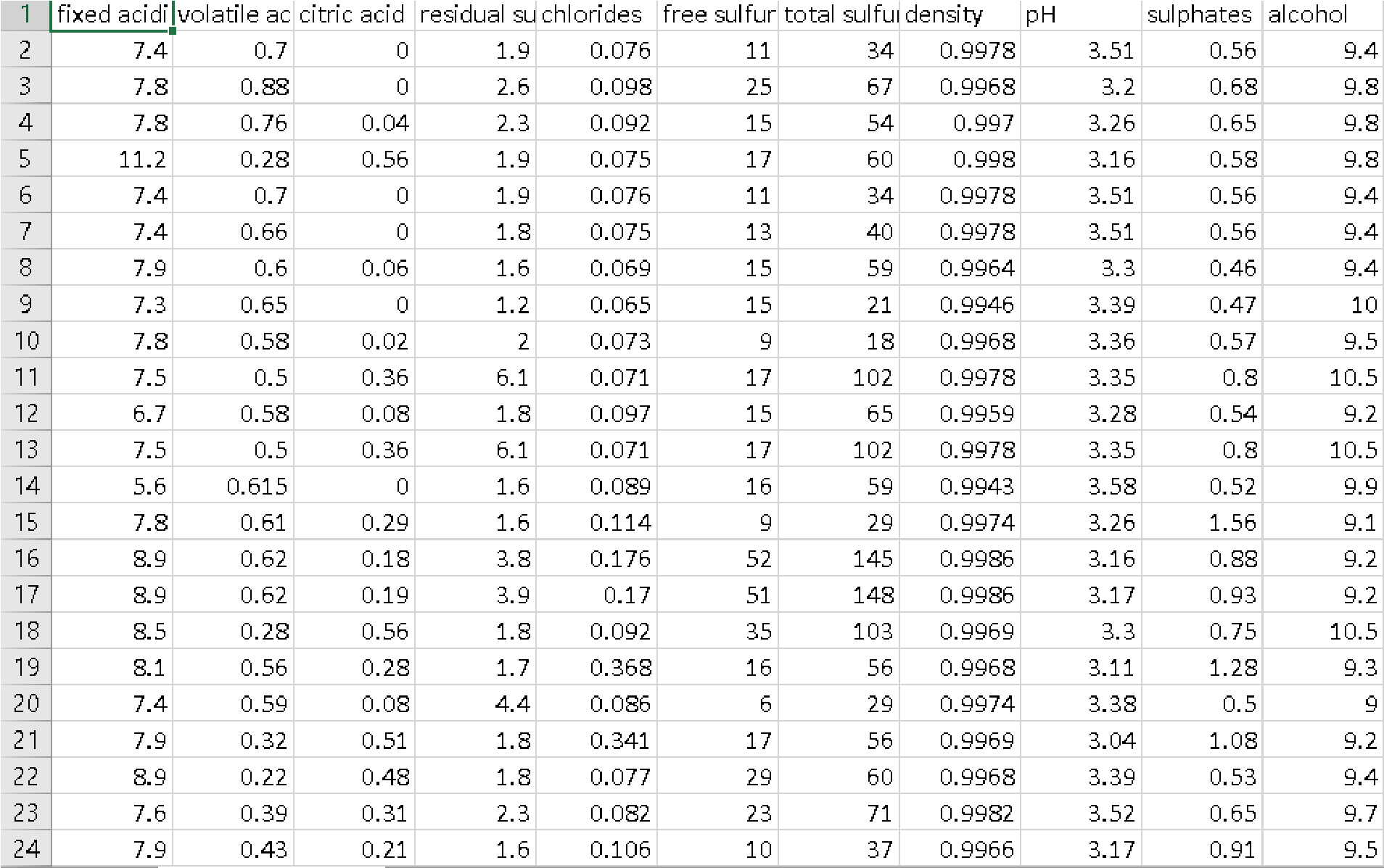
* Lấy dữ liệu mô tả đặc tính rượu
* Trích chọn đặc trưng từ tập dữ liệu lấy được
* Xử lý, làm sạch dữ liệu
* Tiến hành phân cụm
* Dữ liệu hóa đồ thị

***3.2. Bộ dữ liệu:* wine\_red\_quality**

Nguồn: https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009

Bộ dữ liệu bao gồm các đặc trưng là các nồng độ các thành phần ảnh hưởng đển chất lượng của rượu, bao gồm:

* *fixed acidity*: độ acid cố định
* *volatile acidity*: độ acid dễ bay hơi
* *citric acid*: nồng độ citric acid – được dùng để thêm độ tươi và hương vị cho rượu vang
* *residual sugar*: lượng đường còn dư lại sau khi dừng lên men, thường lớn hơn 1 gr/liter
* *chlorides*: độ mặn của rượu
* *free sulfur dioxide*: lượng chất SO2 còn tồn tại ở trạng thái cân bằng giữa phân tử SO2 và bisulfite ion, nó ngăn cản vi sinh vật và sự oxi hóa ở rượu
* *total sulfur dioxide*: số lượng các dạng S02 tự do và ràng buộc; ở nồng độ thấp, SO2 hầu như không thể phát hiện được trong rượu vang, nhưng ở nồng độ SO2 tự do trên 50 ppm, SO2 trở nên rõ ràng trong mũi và mùi vị của rượu. - *density*: tỉ trọng của nước
* *pH*: là thước đo của tính axit và bazo trên thang điểm 0 -14
* *sulphates*: một chất phụ gia rượu vang có thể góp phần vào mức khí sulfur dioxide (S02), hoạt động như một chất chống vi khuẩn và chất chống oxy hóa
* *alcohol*: nồng độ cồn
* *quality*: chất lượng (là kết quả của một số mẫu rượu được đánh giá dựa trên vị giác của người thử rượu, tuy nhiên bài toán phân cụm không cần đến đặc trưng này)



***Hình 3.1: Dữ liệu cụ thể***

*3.2.4. Tiến hành phân cụm*

IMPORT THƯ VIỆN:

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

from sklearn.cluster import KMeans

from sklearn.decomposition import PCA

from sklearn.preprocessing import StandardScaler

ĐỌC FILE DỮ LIỆU CSV:

df = pd.read\_csv('winequality-red.csv')

BỎ NHỮNG DỮ LIỆU BỊ LẶP

df.drop(columns = ['quality','fixed acidity','volatile acidity'],inplace = True)

CHUẨN HÓA DỮ LIỆU (StandardScaler()):

X = df.values

X = StandardScaler().fit\_transform(X)

SỬ DỤNG PHƯƠNG PHÁP ELBOW ĐỂ ĐƯA RA SỐ CỤM K TỐI ƯU

* Inertia: Tổng khoảng cách của các điểm dữ liệu đến tâm trong cụm đó được gọi là inertia.
* Phương pháp Elbow: tìm điểm trên biểu đồ biểu diễn inertia sao cho tại điểm đó giá trị inertia giảm chậm lại

A graph with a line

Description automatically generated

**Hình 3.1. Biểu đồ inertia**

Trên biểu đồ có thể nhận thấy khi k =10 thì inertia giảm chậm lại. Chọn k = 10

HUẤN LUYỆN MÔ HÌNH

lr = KMeans(n\_clusters = 10,init = 'k-means++',random\_state = 6)

clr.fit(X)

ĐƯA RA 10 TÂM CỦA BÀI TOÁN

array([[ 8.11129719e-01, -5.41370707e-02, -1.98640924e-01,

-5.69792516e-01, -7.13208981e-01, -6.61051488e-01,

-2.97880320e-01, 2.61456615e-01, 1.23129341e+00],

[ 8.63486515e-02, 3.57230106e-02, 4.33875908e-03,

8.78783181e-01, 1.42904068e+00, 3.03655032e-01,

-2.25790761e-01, -4.14965308e-01, -7.33431670e-01],

[-9.22782605e-01, 5.82506833e-02, -1.12320042e-01,

-3.66649582e-01, -3.91987389e-01, 1.49055707e-01,

1.07040905e+00, -1.83818907e-01, -2.32846039e-01],

[ 4.27003869e-01, -2.35747998e-01, 3.00573410e-01,

-5.66433594e-02, 5.23881423e-01, 5.20251741e-01,

-9.48116569e-01, 2.83713747e+00, -4.48057422e-01],

[-9.01186996e-01, -2.99601011e-01, -5.43476574e-01,

1.15088718e-03, -1.15235739e-01, -1.69644988e+00,

1.44865292e+00, -2.42575405e-01, 1.67878398e+00],

[ 1.39001109e+00, -3.91139088e-01, 6.63541465e+00,

-7.96848307e-02, 1.87234325e-01, 1.76664602e-01,

-1.59733554e+00, 3.14107620e+00, -9.40690430e-01],

[-1.68021671e-01, -2.53469419e-01, -2.60681136e-01,

1.45316192e+00, 3.95581669e-01, -3.91944049e-01,

3.96763762e-01, 3.89954546e-01, 4.55112412e-01],

[ 1.22370035e+00, 1.45523442e-01, 1.02848644e-02,

-5.51915593e-01, -4.94878532e-01, 1.20649076e+00,

-9.49626624e-01, 1.71482348e-01, -1.57441527e-01],

[ 4.14855742e-01, 4.96176756e+00, 2.96387895e-01,

1.75019116e+00, 1.69583213e+00, 1.22500051e+00,

-3.25459600e-01, -2.37893278e-02, -3.63912996e-01],

[-6.28981311e-01, -3.65957001e-01, -5.41943326e-02,

-5.06776562e-01, -4.24746140e-01, -1.53661525e-01,

-1.13600852e-01, -5.74288766e-01, -6.24902463e-01]])

TẠO MODEL PHÂN CỤM DỮ LIỆU VÀ BIỂU DIỄN DỮ LIỆU:

df = PCA(2).fit\_transform(df)

clr = KMeans(n\_clusters = 10,init = 'k-means++',random\_state = 6)

clr.fit(df)

labels = clr.predict(df)

centroid\_labels = clr.predict(clr.cluster\_centers\_)

predicted\_df = pd.DataFrame(data = df,columns = ['PCA1','PCA2'])

predicted\_df['Cluster'] = labels

centroid\_df = pd.DataFrame(data = clr.cluster\_centers\_,columns = ['PCA1','PCA2'])

centroid\_df['Cluster'] = centroid\_labels

plt.scatter(predicted\_df['PCA1'],predicted\_df['PCA2'],c = labels,alpha = 0.8)

plt.scatter(centroid\_df['PCA1'],centroid\_df['PCA2'],marker = '+',s = 100,c = 'red')

plt.show()

A colorful dots with red and blue crosses

Description automatically generated with medium confidence

***Hình 3.2: Hình ảnh dữ liệu sau khi được phân cụm trên đồ thị***

**CHƯƠNG 4 ƯU ĐIỂM & HẠN CHẾ CỦA THUẬT TOÁN**

Thuật toán K-Means clustering có nhiều ưu điểm và hạn chế trong bài toán phân cụm. Dưới đây là một số ưu điểm và hạn chế của thuật toán K-Means clustering:

**Ưu điểm của K-Means clustering:**

1. **Dễ triển khai và hiểu:** K-Means là một trong những thuật toán phân cụm đơn giản và dễ triển khai. Nó dễ hiểu và thường rất hiệu quả trong việc phân cụm dữ liệu.
2. **Tính hiệu quả tính toán:** K-Means có thể xử lý tập dữ liệu lớn một cách nhanh chóng và hiệu quả. Thuật toán hoạt động tốt với các tập dữ liệu có kích thước lớn.
3. **Phân cụm không giám sát:** K-Means là một phương pháp phân cụm không giám sát, nghĩa là không cần có thông tin nhãn trước cho dữ liệu. Nó tự động tạo ra các nhóm dựa trên tương đồng.
4. **Khả năng xử lý dữ liệu dạng số:** K-Means thích hợp cho dữ liệu số học, và có thể được sử dụng rộng rãi trong các ngành như khoa học dữ liệu và học máy.

**Hạn chế của K-Means clustering:**

1. **Nhạy cảm với lựa chọn số lượng nhóm (k):** K-Means đòi hỏi bạn phải xác định số lượng nhóm (k) trước khi thực hiện phân cụm, và lựa chọn k không phải lúc nào cũng dễ dàng. Sai lựa chọn k có thể dẫn đến kết quả phân cụm không tốt.
2. **Không hiệu quả đối với các cụm có hình dạng và kích thước khác nhau:** K-Means thường hoạt động tốt khi các cụm có hình dạng và kích thước tương tự. Nếu các cụm có kích thước và hình dạng khác nhau, K-Means có thể tạo ra kết quả không tốt.
3. **Nhạy cảm với các giá trị ngoại lệ:** K-Means dễ bị ảnh hưởng bởi các giá trị ngoại lệ (outliers) trong dữ liệu. Các giá trị ngoại lệ có thể làm thay đổi trọng tâm của cụm một cách nghiêm trọng.
4. **Không tìm ra cụm con:** K-Means tìm kiếm các cụm toàn cục, không phát hiện được cụm con trong trường hợp dữ liệu có cấu trúc phức tạp.
5. **Phụ thuộc vào điểm khởi tạo ban đầu:** Kết quả của K-Means có thể thay đổi dựa vào điểm ban đầu mà bạn chọn để khởi tạo cụm, và điều này có thể làm cho thuật toán không ổn định.

TÀI LIỆU THAM KHẢO

[1] Trang web: machinelearningcoban.com