



# Khai phá dữ liệu

Áp dụng K-Means và công cụ Weka

---

Nhóm Undefined

Giảng viên: Lê Hồng Trang

26th May 2021

Thành viên

---

## Thành viên

1771726 - Bùi Quốc Khải

1711400 - Nguyễn Minh Hoàng

1710779 - Huỳnh Thi Trường Duy

1711130 - Võ Quý Giang

## Nội dung

---

# Nội dung

1. Mô tả kỹ thuật K-Means
  - 1.1 Phân cụm dữ liệu
  - 1.2 Kỹ thuật K-Means
  - 1.3 Mô tả hoạt động
  - 1.4 Phương pháp xác định số cụm K
  - 1.5 Phương pháp xác định khoảng cách
2. K-Means++ và K-Medoids
3. Công cụ Weka
4. Mô tả thực nghiệm
  - 4.1 Khởi tạo
  - 4.2 Kết quả
5. Phân tích

## Mô tả kỹ thuật K-Means

---

# Mô tả kỹ thuật K-Means

---

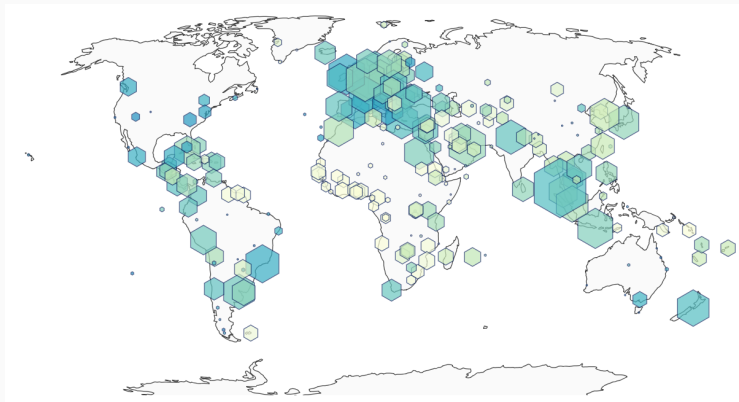
Phân cụm dữ liệu

## Phân cụm dữ liệu

Phân cụm là kỹ thuật rất quan trọng trong khai phá dữ liệu, nó thuộc lớp các phương pháp **Unsupervised Learning** trong Machine Learning. Có rất nhiều định nghĩa khác nhau về kỹ thuật này, nhưng về bản chất ta có thể hiểu phân cụm là các qui trình tìm cách nhóm các đối tượng đã cho vào các cụm (clusters), sao cho các đối tượng trong cùng 1 cụm tương tự (similar) nhau và các đối tượng khác cụm thì không tương tự (dissimilar) nhau.



## Phân cụm dữ liệu



**Figure 1:** Ứng dụng phân cụm ([miro.medium.com/max](https://miro.medium.com/max))

# Mô tả kỹ thuật K-Means

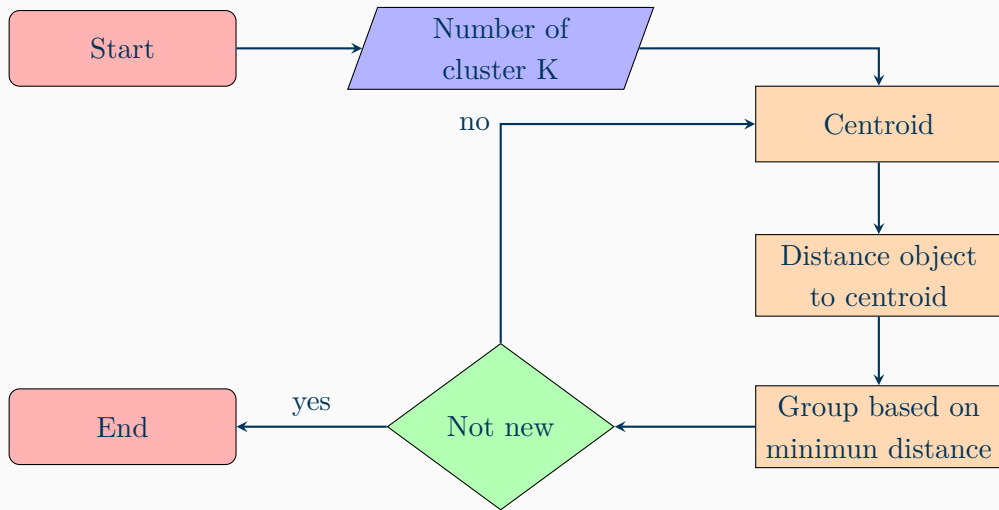
---

## Kỹ thuật K-Means

## Kỹ thuật K-Means

K-Means là thuật toán rất quan trọng và được sử dụng phổ biến trong kỹ thuật phân cụm. Tư tưởng chính của thuật toán K-Means là tìm cách phân nhóm các đối tượng (objects) đã cho vào K cụm (K là số các cụm được xác định trước, K nguyên dương) sao cho tổng bình phương khoảng cách giữa các đối tượng đến tâm nhóm (centroid ) là nhỏ nhất.

## Kỹ thuật K-Means



# Mô tả kỹ thuật K-Means

---

Mô tả hoạt động

## Mô tả hoạt động

Hoạt động của thuật toán K-Means tuân theo các bước:

1. Chọn số K để quyết định số lượng cụm.
2. Chọn K điểm hoặc trọng tâm ngẫu nhiên.
3. Gán mỗi điểm dữ liệu cho trung tâm gần nhất của chúng, sẽ tạo thành các cụm K được xác định trước.
4. Tính toán phương sai và đặt một trung tâm mới của mỗi cụm.
5. Lặp lại các bước thứ ba, có nghĩa là chỉ định lại mỗi điểm dữ liệu cho trung tâm gần nhất mới của mỗi cụm.
6. Nếu có bất kỳ sự phân công lại nào xảy ra, hãy chuyển sang bước-4, sau đó hoàn tất quá trình.

# Mô tả kỹ thuật K-Means

---

Phương pháp xác định số cụm K

## Phương pháp xác định số cụm K

1. Thực hiện phân cụm K-Means trên một tập dữ liệu nhất định cho các giá trị K khác nhau (phạm vi từ 1-10).
2. Đối với mỗi giá trị của K tính giá trị WCSS:

$$WWCS(k) = \sum_{j=1}^k \sum_{i \in cluster j} \|x_i - \bar{x}_j\|^2$$

3. Vẽ đồ thị một đường cong giữa các giá trị WCSS được tính toán và số lượng các cụm K.
4. Điểm uốn cong hoặc một điểm của đồ thị có dạng như một cánh tay thì điểm đó được coi là giá trị tốt nhất của K.



## Phương pháp xác định số cụm K

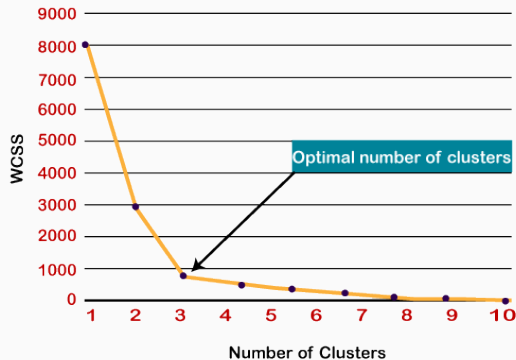


Figure 2: Đồ thị xác định K trong K-Means

# Mô tả kỹ thuật K-Means

---

Phương pháp xác định khoảng cách

## Phương pháp xác định khoảng cách

- Phương pháp thước đo Euclidean:

$$d = \sqrt{\sum_{i=1}^n (q_j - p_i)^2}$$

- Phương pháp thước đo Manhattan:

$$d = \sum_{i=1}^n (q_j - p_i)^2$$

## Phương pháp xác định khoảng cách

- Phương pháp thước đo Euclidean về bình phương khoảng cách:

$$d = \sum_{i=1}^n |q_x - p_x| + |q_y - p_y|$$

- Phương pháp thước đo Cosine:

$$d = \frac{\sum_{i=1}^{n-1} q_i - p_x}{\sum_{i=1}^{n-1} (q_i)^2 \times \sum_{i=1}^{n-1} (p_i)^2}$$

## K-Means++ và K-Medoids

---

# K-Means++ và K-Medoids

---

K-Means++

## K-Means++

Thuật toán K-Means++ chỉ định một thủ tục để khởi tạo các k-centroid trước khi tiếp tục với các lần lặp tối ưu hóa K-Means tiêu chuẩn. Với việc khởi tạo K-Means++, thuật toán được đảm bảo tìm độ tối ưu  $O(\log k)$  so với giải pháp K-Means tối ưu.

# K-Means++ và K-Medoids

---

## K-Medoids



## K-Medoids

Ý tưởng của việc phân cụm K-Medoids là tạo ra các trung tâm cuối cùng dưới dạng các điểm dữ liệu thực tế. Kết quả này làm cho các trung tâm có thể giải thích được.

Thuật toán phân cụm K-Medoids được gọi là Phân vùng xung quanh Medoids (PAM) gần giống như thuật toán của Lloyd với một chút thay đổi trong bước cập nhật

## Công cụ Weka

---

## Công cụ Weka

**Weka** - một phần mềm mã nguồn mở cung cấp các công cụ để xử lý trước dữ liệu, triển khai một số thuật toán Máy học và trực quan hóa giúp phát triển các kỹ thuật máy học và áp dụng chúng vào các vấn đề khai thác dữ liệu trong thế giới thực.

# Công cụ Weka

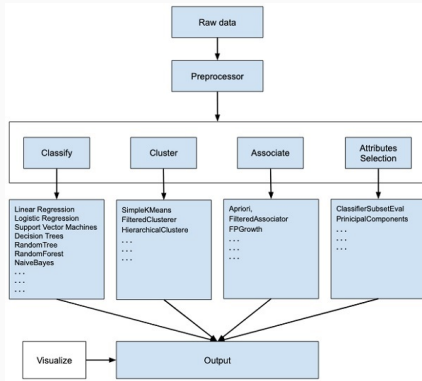


Figure 3: Ứng dụng Weka

## Mô tả thực nghiệm

---

# Mô tả thực nghiệm

---

Khởi tạo

## Khởi tạo

Lựa chọn tập dữ liệu phân tích đối tượng khách hàng của một ngân hàng chứa thông tin cơ bản của các khách hàng trong 600 đối tượng khảo sát thu được dataset: <http://bis.net.vn/files/storage/bank.rar>

Dữ liệu bao gồm các thuộc tính: **age, sex, region, income, married, children, car, save-act, current-act, mortgage, pep.**

Khởi chạy với công cụ Weka với công cụ gom cụm dữ liệu dựa trên K-Means với số cụm tăng dần từ **0** đến **10**

Mô tả thực nghiệm

---

Kết quả



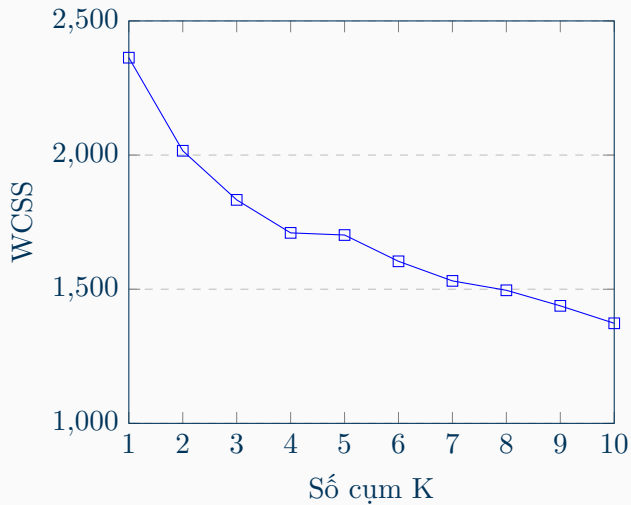
## Kết quả

Number of K	Iterations	WCSS
1	1	2363
2	4	2016
3	5	1833
4	7	1710
5	5	1702
6	6	1604
7	12	1531
8	7	1496
9	8	1438
10	7	1373

# Phân tích

---




## Phân tích



## Tài liệu tham khảo

---

## Tài liệu tham khảo

-  K-Means Clustering Algorithm: Applications, Types, Demos and Use Cases, nguồn tham khảo: <https://www.simplilearn.com/tutorials/machine-learning-tutorial/k-means-clustering-algorithm>, truy cập: 05/24/2021.
-  Hiểu các thuật toán phân cụm K-mean, K-means ++ và K-medoids, nguồn tham khảo: <https://ichi.pro/vi/hieu-cac-thuat-toan-phan-cum-k-mean-k-means-va-k-medoids>, truy cập: 05/24/2021.
-  Weka - Clustering, nguồn tham khảo: [https://www.tutorialspoint.com/weka/weka\\_clustering.htm](https://www.tutorialspoint.com/weka/weka_clustering.htm), truy cập: 05/24/2021.

Cảm ơn đã lắng nghe