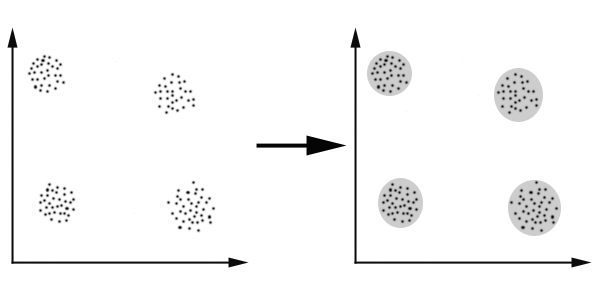
BTL Nhận Dạng

# CHƯƠNG 1. TỔNG QUAN

## Phân cụm là gì?

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.

Mục đích của phân cụm là tìm ra bản chất bên trong các nhóm của dữ liệu. Các thuật toán phân cụm (Clustering Algorithms) đều sinh ra các cụm (clusters). Tuy nhiên, không có tiêu chí nào là được xem là tốt nhất để đánh hiệu của của phân tích phân cụm, điều này phụ thuộc vào mục đích của phân cụm là gì.



Nếu X : 1 tập các điểm dữ liệu

Ci : cụm thứ i

X = C1 ∪ … ∪ Ck … ∪ Cngoại lai

Ci ∩ Cj = ∅

## Một số phương pháp phân cụm điển hình

* + Phân cụm phân hoạch (\*)
  + Phân cụm phân cấp
  + Phân cụm dựa trên mật độ
  + Phân cụm dựa trên lưới
  + Phân cụm dựa trên mô hình
  + Phân cụm có ràng buộc

(\*) Trong giới hạn đề tài nhóm chỉ đề cập đến phân cụm phân hoạch và thuật toán Kmean trong phương pháp đó.

* Phân cụm phân hoạch:

Phân 1 tập dữ liệu có n phần tử cho trước thành k tập con dữ liệu (k ≤ n), mỗi tập con biểu diễn 1 cụm.

Các cụm hình thành trên cơ sở làm tối ưu giá trị hàm đo độ tương tự sao cho:

* + Các đối tượng trong 1 cụm là tương tự.
  + Các đối tượng trong các cụm khác nhau là không tương tự nhau.

Đặc điểm:

* Mỗi đối tượng chỉ thuộc về 1 cụm.
* Mỗi cụm có tối thiểu 1 đối tượng.
* Một số thuật toán điển hình : K-mean, PAM, CLARA,…

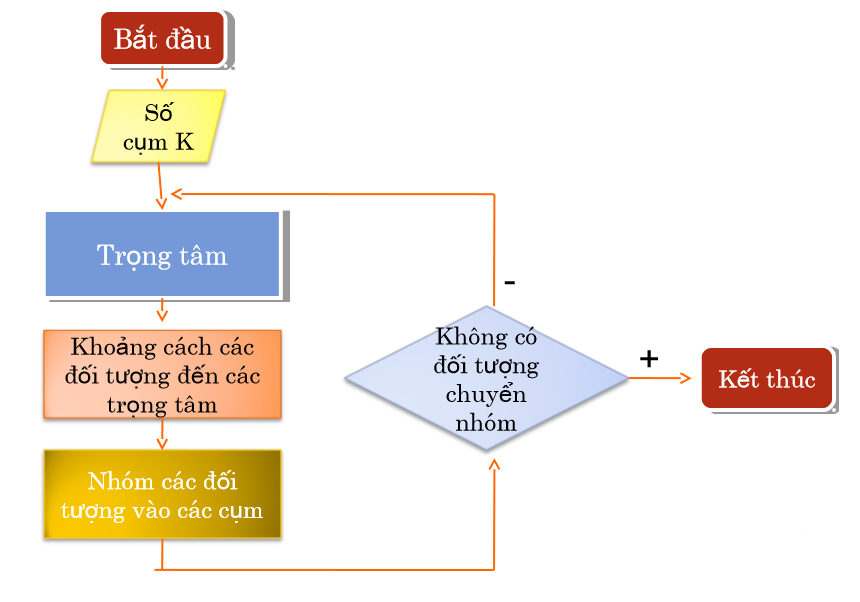
## Thuật toán K- Mean

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 đinh trước, K nguyên dương) sao cho tổng bình phương khoảng cách giữa các đối tượng đến tâm nhóm (centroid) là nhỏ nhất.

* Phát biểu bài toán:
* Input
* Tập các đối tượng X = {xi| i = 1, 2, …, N},
* Số cụm: K
* Output

Các cụm Ci ( i = 1 ÷ K) tách rời và hàm tiêu chuẩn E đạt giá trị tối thiểu.

* Mô tả:



**Thuật toán K-Means thực hiện qua các bước chính sau:**

1.    Chọn ngẫu nhiên K tâm (centroid) cho K cụm (cluster). Mỗi cụm được đại diện bằng các tâm của cụm.

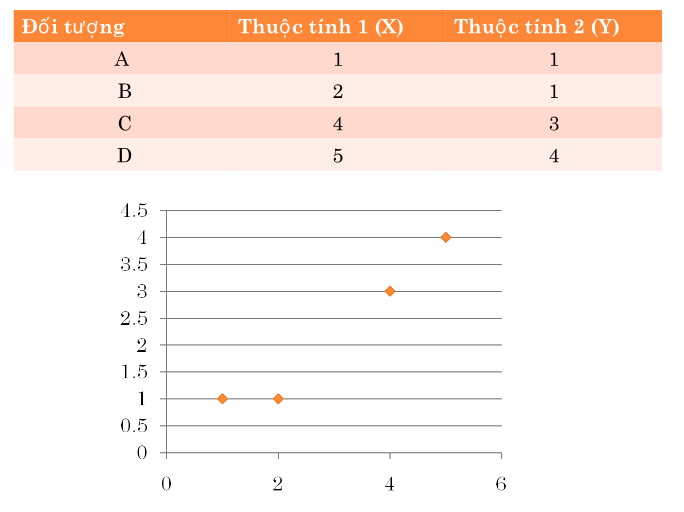
2.    Tính khoảng cách giữa các đối tượng (objects) đến K tâm (thường dùng khoảng cách Euclidean)

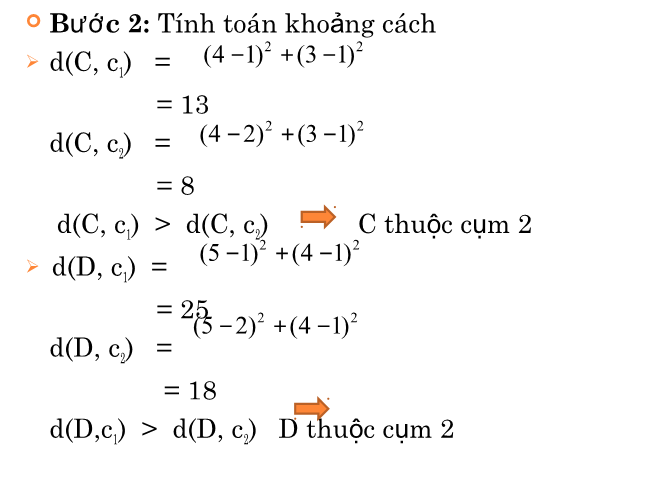
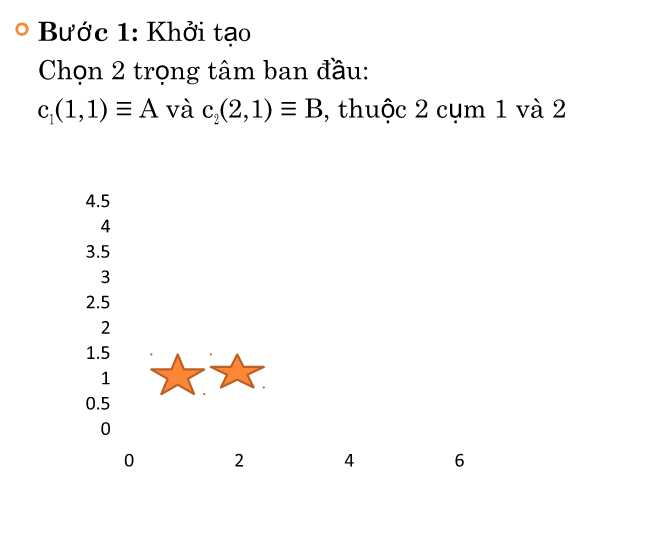
3.    Nhóm các đối tượng vào nhóm gần nhất

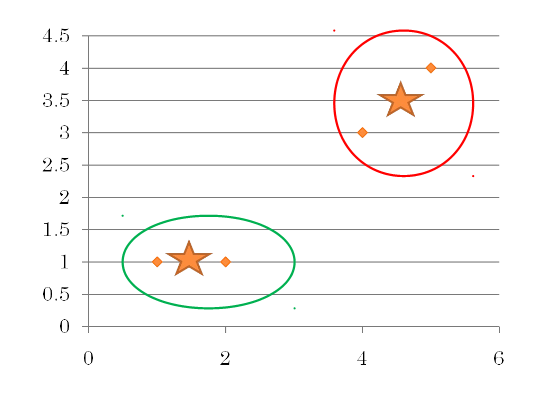
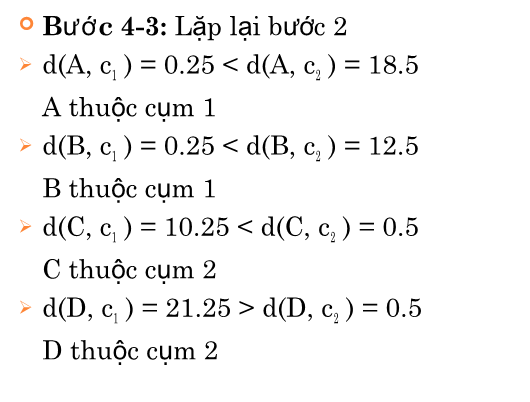
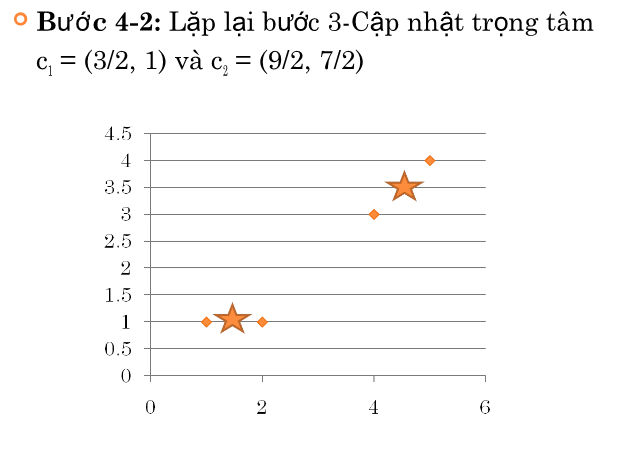
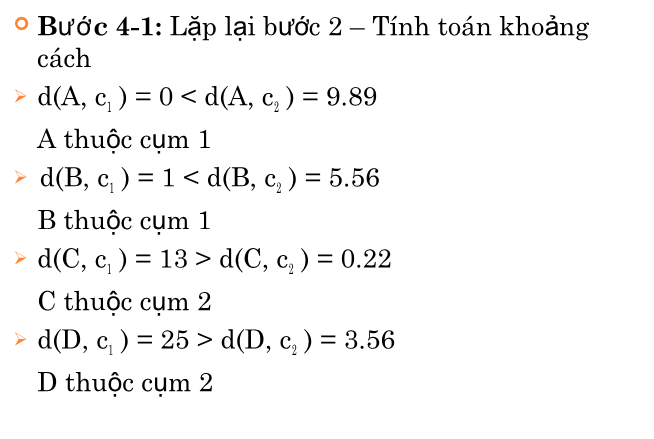
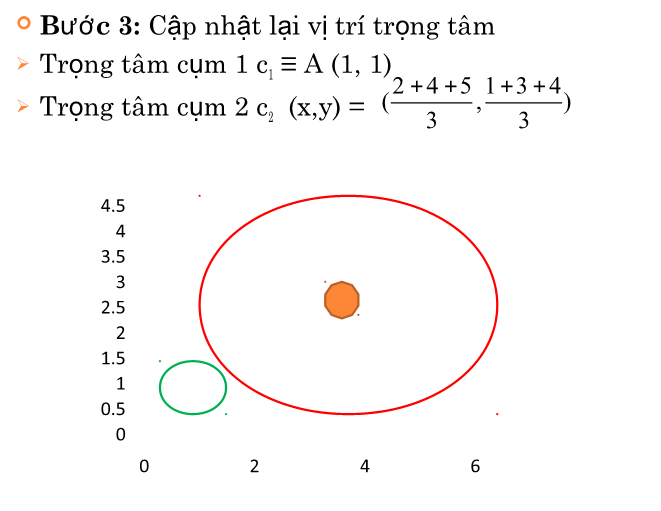
4.    Xác định lại tâm mới cho các nhóm

5.    Thực hiện lại bước 2 cho đến khi không có sự thay đổi nhóm nào của các đối tượng

* **Minh họa**







# CHƯƠNG 2. ỨNG DỤNG PHÂN CỤM KẾT QUẢ HỌC TẬP CỦA SINH VIÊN

## 2.1. Phân tích hệ thống điểm

Nhóm chọn hệ thống tính điểm của trường ĐHCN HN để làm đề tài.

Kết quả học tập của sinh viện được đánh giá dựa theo 2 tiêu chí:

+ Điểm tích lũy

+ Điểm rèn luyện

Điểm tích lũy của sinh viên nằm trong khoảng từ 0 đến 4, điểm rèn luyện trong khoảng từ 0 đến 100

Đánh giá kết quả bằng tổng có trọng số theo công thức:

X = Điểm tích lũy \* 0.25 + Điểm rèn luyện \* 0.01

## 2.2. Đặc tả chương trình

- Chức năng:

+ Nhập danh sách điểm của sinh viện

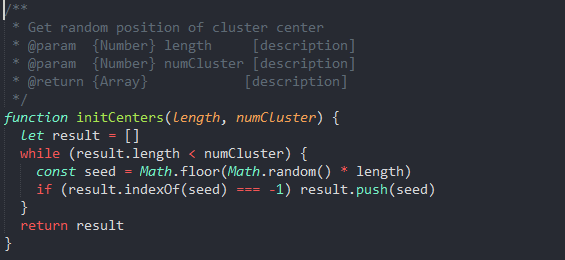
+ Chuẩn hóa dữ liệu đầu vào

+ Hiển thị dữ liệu dưới dạnh biểu đồ

+ Sắp xếp dữ liệu

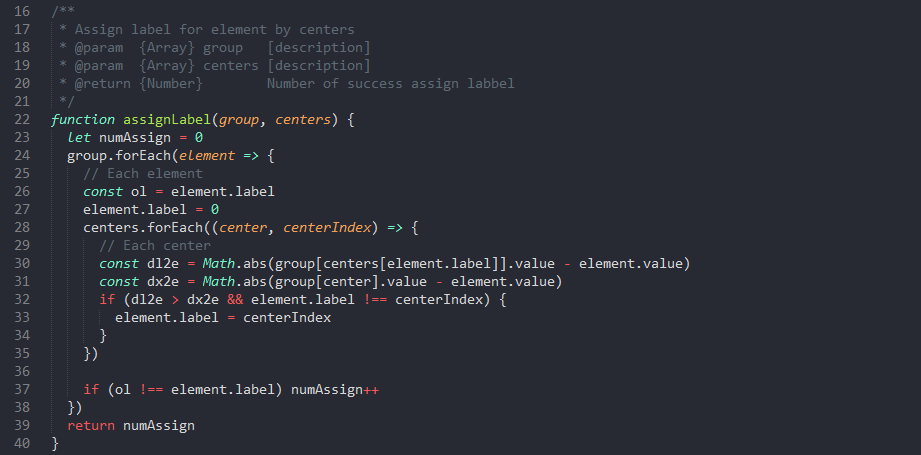
+ Phân cụm dữ liệu sử dụng thuật toán k-mean

-Chương trình gồm 3 hàm chinh:



initCenters: Hàm khởi tạo tâm ngẫu nhiên.

Nhận vào chiều dài tập hợp và số cụm cần chia. Hàm trả về vị trị tâm được gieo ngẫu nhiên.



assignLabel: Hàm gắn nhãn cho phần tử.

Dựa vào khoảng cách từ mỗi phần tử đến các tâm để xác định phần từ thuộc cụm nào.

Hàm trả về số lượng phần tử được gán nhãn thành công.



updateCenters: Hàm cập nhật lại tâm của các cụm

Xác định tâm mới của mỗi cũ dựa trên khoảng các từ tâm cũ đến từng phần tử trong cụm

# CHƯƠNG 3. CHƯƠNG TRÌNH DEMO

