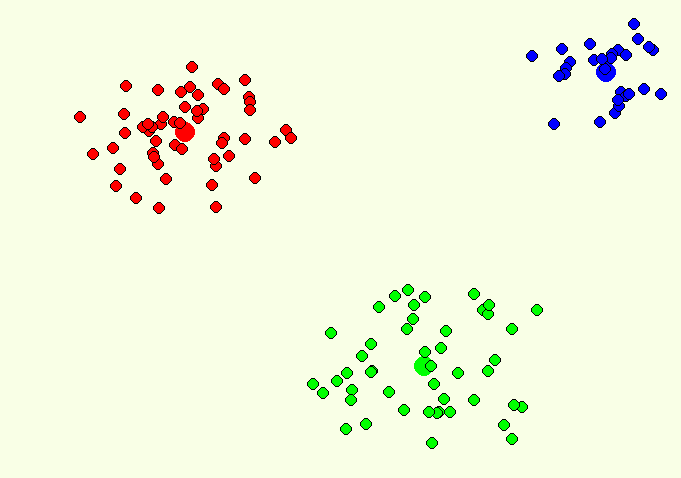
Báo cáo

Học phần: Trí tuệ nhân tạo

Giáo viên hướng dẫn: Nguyễn Xuân Phúc

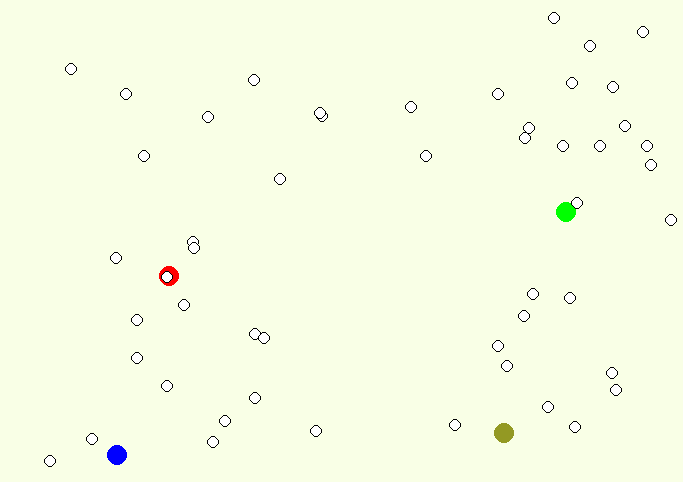
1. Giới thiệu về thuật toán phân cụm K-means ( K-means clustering )

* Là 1 thuật toán cơ bản và thông dụng trong unsupervised learning.
* Trong thuật toán K-means, chúng ta không biết nhãn (label) của từng điểm dữ liệu. Mục đích là làm thế nào để phân các dữ liệu thành các cụm (cluster) khác nhau sao cho dữ liệu trong cùng 1 cụm có tính chất tương tự nhau.
* Ví dụ đơn giản về cụm (cluster) là tập hợp các điểm gần nhau trong một không gian nào đó. Hình dưới đây biểu diễn ví dụ về 3 cluster:

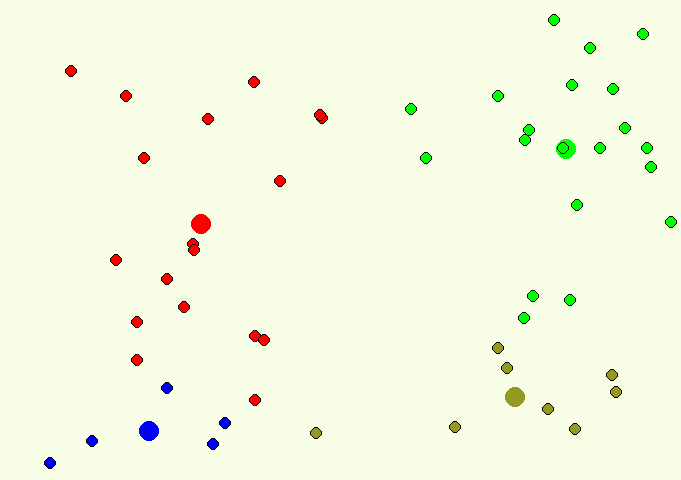


1. Tóm tắt thuật toán.

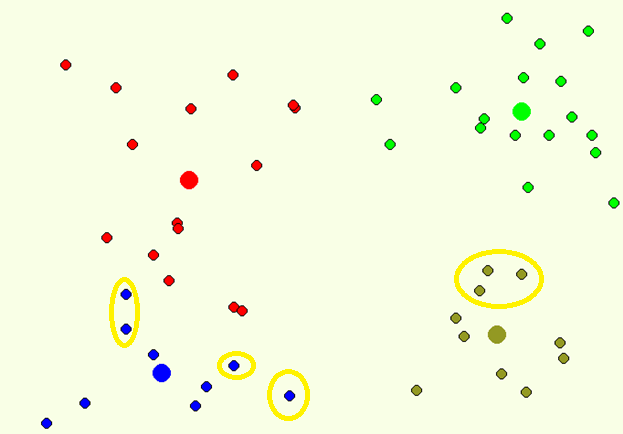
* Input : Dữ liệu X và số lượng cluster cần tìm K
* Output: Các center M và label cho từng điểm dữ liệu Y
* Các bước thực hiện:
  + Chọn K điểm bất kì làm các center ban đầu
  + Phân mỗi điểm dữ liệu vào cluster có center gần nó nhất và gán nhãn cho các điểm dữ liệu trên tương ứng với các center
  + Nếu việc gán nhãn vào từng cluster không thay đổi so với vòng lặp trước thì ta dừng thuật toán
  + Cập nhật center cho từng cluster bằng cách lấy trung bình cộng của tất cả các điểm dữ liệu đã được gán vào cluster đó sau bước 2
  + Quay lại bước 2
* Thuật toán sẽ dừng lại sau một số vòng lặp hữu hạn. Thật vậy , vì hàm mất mát là 1 số dương và sau bước 2 và bước 3, giá trị của hàm mất mát sẽ bị giảm đi nên sau hữu hạn bước thuật toán sẽ hội tụ. Hơn nữa, số lượng cách phân nhóm cũng là hữu hạn nên đến một lúc nào đó, hàm mất mát sẽ không thay đổi và thuật toán dừng lại.
* Sau đây là ví dụ với K =4
  + Bước 1: Khởi tạo dữ liệu và khởi tạo ngẫu nhiên 4 center



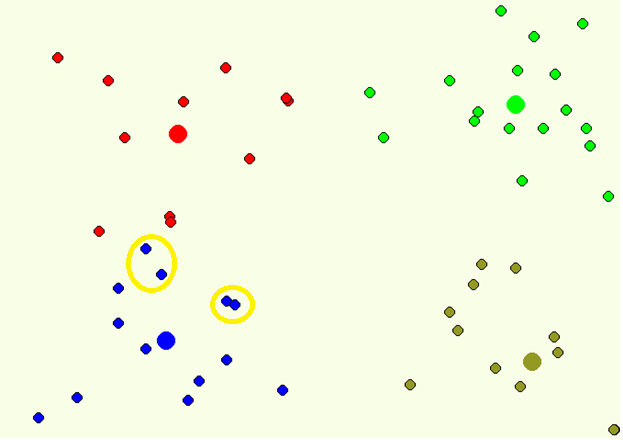
* + Bước 2: phân các điểm vào cluster và gán nhãn cho chúng



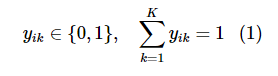
* + Bước 3: điều chỉnh lại các center và cập nhật lại nhãn cho các điểm dữ liệu (chạy lại bước 2)



* + Bước 4: Làm tương tự như bước 3 và sau đó đây cũng là bài toán khi hội tụ.



1. Phân tích toán học.

* Một số kí hiệu toán học
  + Giả sử có N điểm dữ liệu X = [x1, x2,…,xN] và K < N là số cluster chúng ta muốn phân chia. Chúng ta cần tìm các center m1 , m2 , …, mK và label của mỗi điểm dữ liệu
  + Với mỗi điểm dữ liệu xi đặt yi = [yi1, yi2,…, yiK] là label vector của nó, giả sử xi được phân vào cluster k thì yik = 1 và các yij = 0 với j != k
    - Ví dụ: nếu 1 điểm có label vector là [1,0,0,…,0] thì nó thuộc cluster 1
  + Ràng buộc của yi có thể được viết lại như sau:
    - 
* Hàm mất mát và tối ưu

1. Một số hạn chế.