TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC MÁY TÍNH

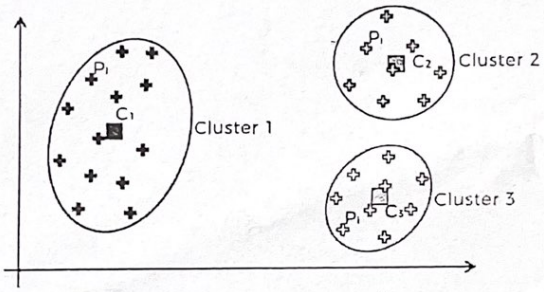
ĐỀ THI CUỐI KỲ
HỌC KỲ 2 – NĂM HỌC 2021-2022

Môn thi: CẤU TRÚC DỮ LIỆU & GIẢI THUẬT
Mã môn/lớp: IT003.M21.KHTN, IT003.M21.ANTN
Thời gian làm bài: 90 phút
(*Sinh viên được sử dụng tài liệu*)

Gom cụm – Clustering

Gom cụm là một phương thức tự gán nhãn cho các cá thể được sử dụng trong ***Máy học – Machine Learning***.

Xét quần thể gồm n cá thể, mỗi cá thể P được xác định bằng vectơ đặc trưng m thành phần $(p_1, p_2, \ldots, p_m)$. Các cá thể có tính chất tương tự nhau được gom vào cùng 1 cụm và gán cùng 1 nhãn. 2 cá thể được gọi là có tính chất tương tự nhau nếu khoảng cách giữa chúng không vượt quá giới hạn D.



Có nhiều cách để xác định khoảng cách giữa 2 cá thể như khoảng cách Euclidean (d_{euc}), khoảng cách Manhattan (d_{man}), ... Ở đây, khoảng cách giữa P và Q là khoảng cách IT003 (d_{003}) được định nghĩa:

$$d_{003}(P,Q) = \left| \sum_{i=1}^{m} (P_i - Q_i) \right|$$

Hãy gom các cá thể vào các cụm sao cho:

1. Khoảng cách giữa 2 cá thể bất trong cùng một cụm không vượt quá D
2. Số lượng cụm gom được là ít nhất.

Dữ liệu:

- Dòng thứ nhất ghi 3 số nguyên dương m, n, D ($m * n \leq 10^6$)
- n dòng tiếp theo, mỗi dòng ghi m số nguyên có trị tuyệt đối không vượt quá 10^6 mô tả vectơ đặc trưng của các cá thể

Kết quả: số cụm gom được

Ví dụ:

INPUT
3 3 2 1 1 1 2 2 2 1 2 2

OUTPUT
2

Hãy lựa chọn cấu trúc dữ liệu và thiết kế giải thuật để giải quyết bài toán trên. Các nội dung cần trình bày:

1. Ý tưởng chung để giải quyết bài toán;
2. Các cấu trúc dữ liệu được chọn: tên, kiểu, ý nghĩa; cài đặt trong ngôn ngữ lập trình C/C++ hoặc Python;
3. Thuật toán (mã giả hoặc lưu đồ);
4. Ước lượng độ phức tạp thuật toán;

HẾT

Duyệt đề **Giảng viên ra đề**