THUẬT TOÁN KNN

KNN (K-Nearest Neighbors) là một trong những thuật toán học có giám sát đơn giản nhất được sử dụng nhiều trong khai phá dữ liệu và học máy. Ý tưởng của thuật toán này là nó không học một điều gì từ tập dữ liệu học (nên KNN được xếp vào loại lazy learning), mọi tính toán được thực hiện khi nó cần dự đoán nhãn của dữ liệu mới. Lớp (nhãn) của một đối tượng dữ liệu mới có thể dự đoán từ các lớp (nhãn) của k hàng xóm gần nó nhất.

Các bước trong KNN

1. Ta có D là tập các điểm dữ liệu đã được gắn nhãn và A là dữ liệu chưa được phân loại.
2. Đo khoảng cách (Euclidian, Manhattan, Minkowski, Minkowski hoặc Trọng số) từ dữ liệu mới A đến tất cả các dữ liệu khác đã được phân loại trong D.
3. Chọn K (K là tham số mà bạn định nghĩa) khoảng cách nhỏ nhất tính được ở bước 2.
4. Trong K khoảng cách được chọn, xác định số lần xuất hiện của các loại lớp (nhãn) có trong tập D. Ví dụ tập D có 2 lớp (nhãn) là **trắng** và **đen** với K = 5 thu được [trắng, đen, đen, trắng, đen] => lớp (nhãn) đen xuất hiện 3 lần, lớp (nhãn) trắng xuất hiện 2 lần.
5. Phân loại dữ liệu A theo lớp (nhãn) xuất hiện nhiều nhất trong bước 4. Trong ví dụ trên là lớp (nhãn) **đen**.
6. Lớp của dữ liệu mới là lớp mà bạn đã nhận được ở bước 5.

Dựa vào mô tả trên về thuật toán KNN, viết 1 chương trình sử dụng KNN để thực hiện yêu cầu sau:

Xác định phân nhóm (Group) cho 1 điểm chưa xác định trong không gian hai chiều

Mô tả: Cho trước một tập hợp **n** điểm trong không gian hai chiều. Các điểm trong tập hợp này được phân loại vào 2 nhóm (nhóm 0 và nhóm 1); Ví dụ: Điểm [1, 12] thuộc nhóm 0, Điểm [5, 3] thuộc nhóm 1… Sử dụng thuật toán KNN, **xác định phân nhóm cho một điểm đã biết tọa độ nhưng chưa có trong tập hợp** dựa vào khoảng cách Euclidean từ điểm đó tới các điểm đã biết trong tập hợp.

Gợi ý làm:

1. Xây dựng một cấu trúc (struct) lưu thông tin các điểm. Cấu trúc bao gồm: Tọa độ x; Tọa độ y; Phân nhóm.
2. Viết hàm tính khoảng cách Euclidean giữa 2 điểm.
3. Viết hàm sắp xếp phần tử của mảng theo thứ tự tăng dần.
4. Khởi tạo thông tin cho tập hợp **n** điểm.

5. Nhập thông tin Tọa độ x và Tọa độ y cho **điểm cần xác định phân nhóm.**
6. Sử dụng thuật toán KNN (với k = 1,3,5,7,9) để xác định phân nhóm cho điểm khởi tạo ở bước 3.

Mở rộng: Tăng số lượng phân nhóm lên 3 (vuông, tròn, tam giác) và thực hiện lại yêu cầu nêu trên.