ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN
BỘ MÔN KHOA HỌC MÁY TÍNH

Bài tập thực hành:
Phân lớp dữ liệu

Thông tin nhóm:

Đàm Thiệu Quang	1241393
Nguyễn Thị Yến	1241444

Bài 2: Trình bày cải tiến của K-mean++:

K-Means++ là một cải tiến của K-means trong đó có sự cải tiến việc khởi tạo tâm ban đầu theo tiêu chí chọn các đặc trưng cách xa nhau và tránh được tình trạng bị ảnh hưởng bởi outlier.

Cải tiến như sau:

Xét tập đặc trưng T: (x_1,x_2,x_3,..,x_n)

- Gọi Dist(x) là khoảng cách từ đặc trưng x (thuộc tập T) đến tâm gần nhất đã được xác định trước đó.
- Đầu tiên, chọn ngẫu nhiên tâm t_1 trong tập T
- Ta chọn tâm tiếp theo $t_i = x' \in T$ với xác suất chọn là:

$$\frac{Dist(x')^2}{\sum_{x \in T} Dist(x)^2}$$

- Chọn tiếp các tâm tiếp theo cho đến khi đủ k tâm