K-Mean

Ngô Minh Nhựt

Bộ môn Công nghệ Tri thức

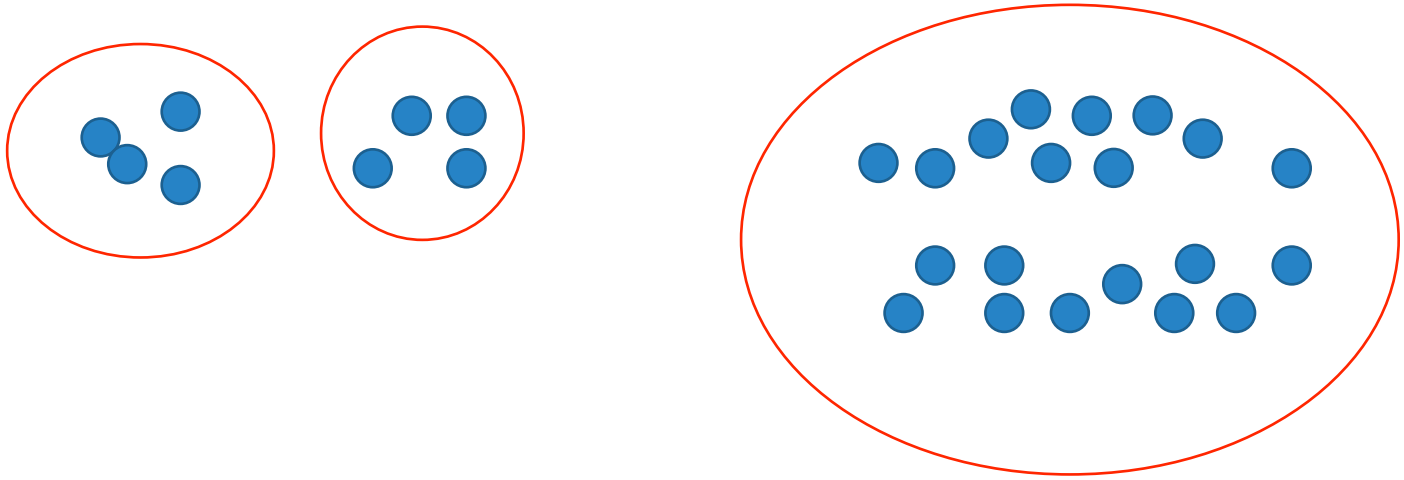
2021

Gom cụm

- Là thuật toán học không giám sát
- Dữ liệu học không cần gán nhãn
- Được dùng để nhận dạng các mẫu giống nhau. Ví dụ:
  - Kết quả tìm kiếm,
  - Thói quen mua sắm, …
- Thuật toán học hữu ích khi có ít thông tin về dữ liệu

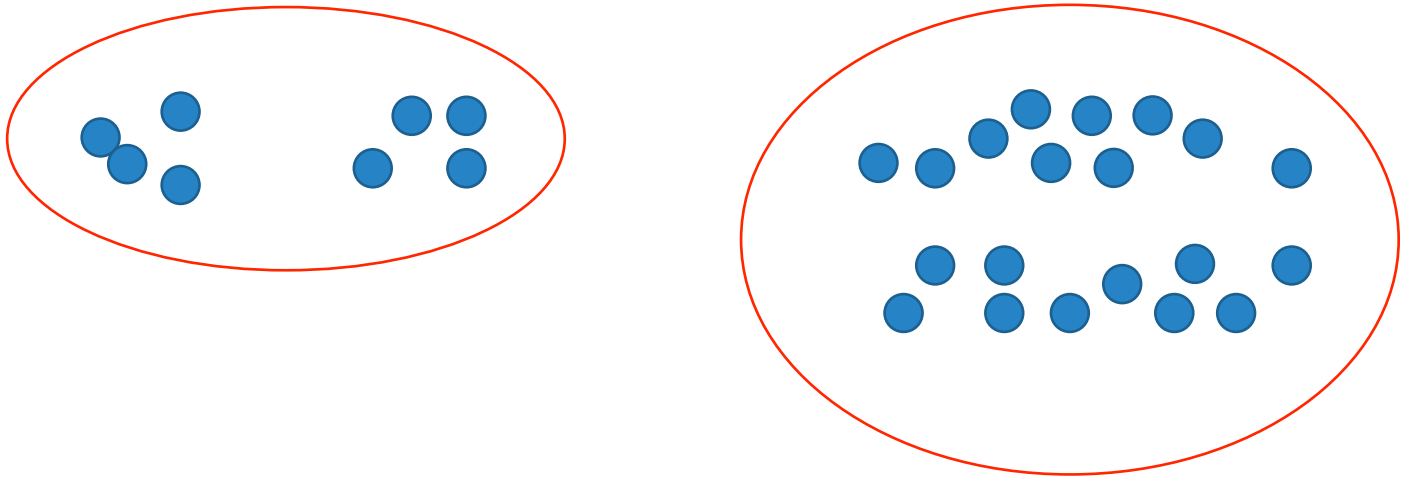
Giới thiệu

- Ý tưởng về gom cụm:
  - Gom những mẫu giống nhau vào cùng nhóm
  - Ví dụ: xét những mẫu dữ liệu 2 chiều sau



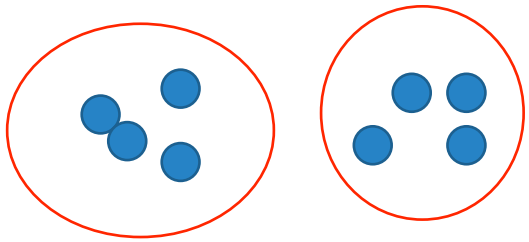
Giới thiệu

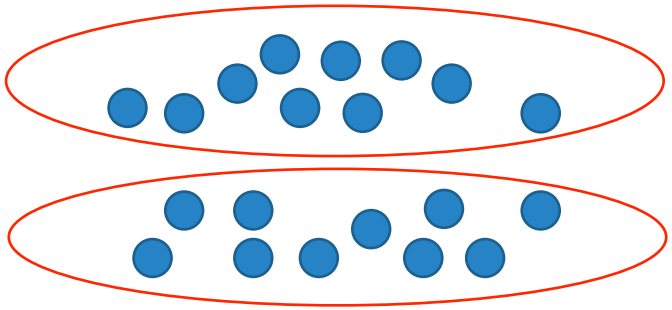
- Ý tưởng về gom cụm:
  - Gom những mẫu giống nhau vào cùng nhóm
  - Ví dụ: xét những mẫu dữ liệu 2 chiều sau



Giới thiệu

- Ý tưởng về gom cụm:
  - Gom những mẫu giống nhau vào cùng nhóm
  - Ví dụ: xét những mẫu dữ liệu 2 chiều sau





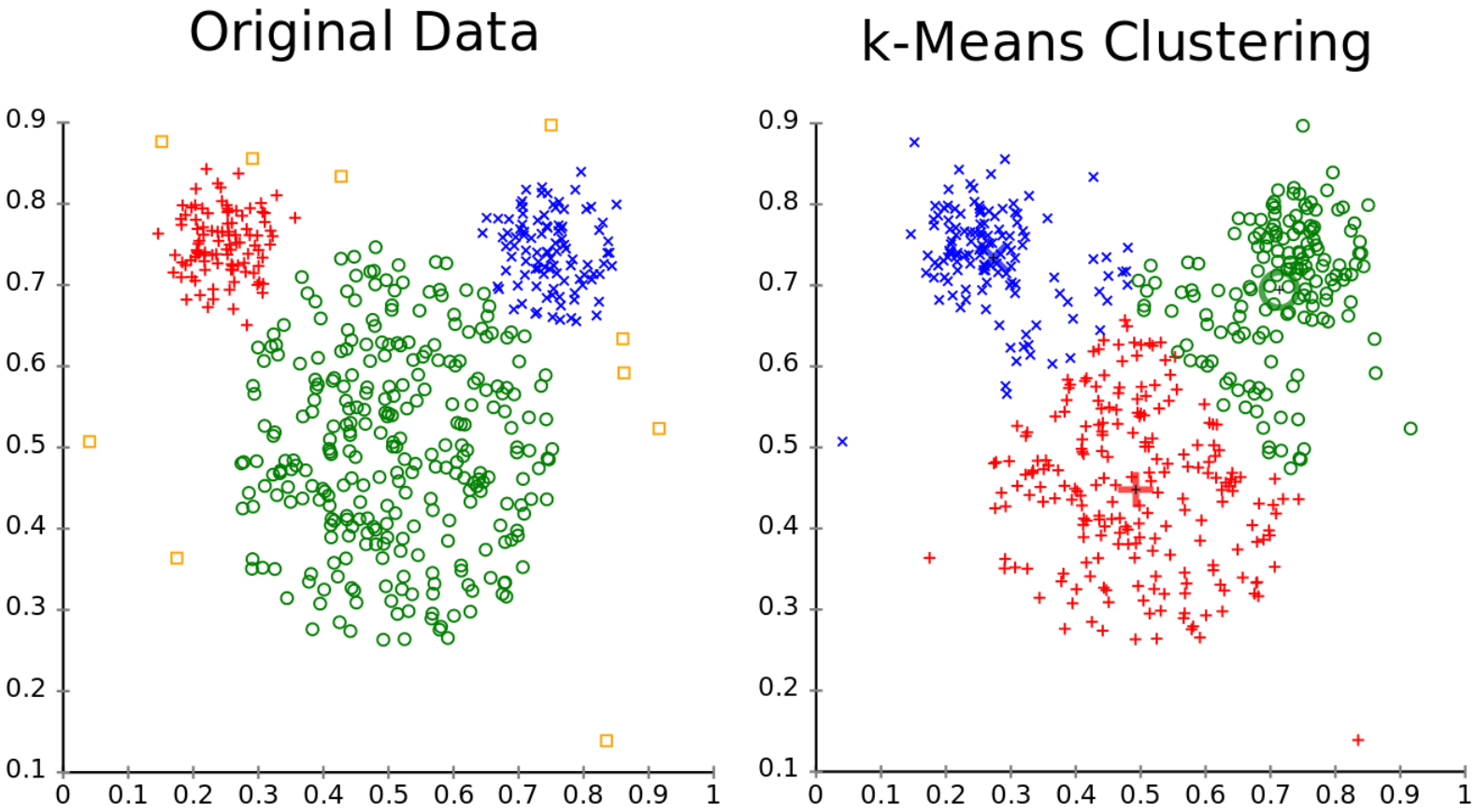
Độ giống nhau (similarity)?

- Ví dụ: khoảng cách Euclide
- Kết quả gom cụm phụ thuộc vào cách tính độ giống nhau

K-mean

- Là thuật toán học không giám sát
- Được dùng để gom cụm dữ liệu: học cấu trúc
- Dựa vào khoảng cách Euclide: 2 mẫu có khoảng cách nhỏ thì thuộc cùng một cụm

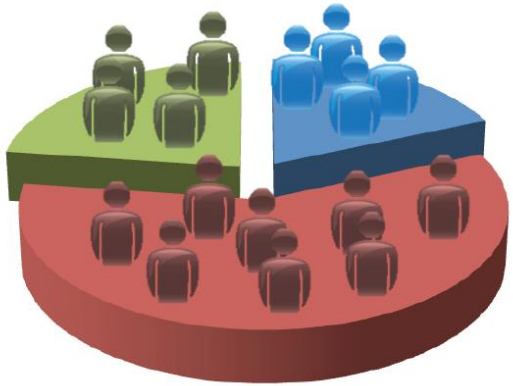
Giới thiệu



Source: Wikipedia

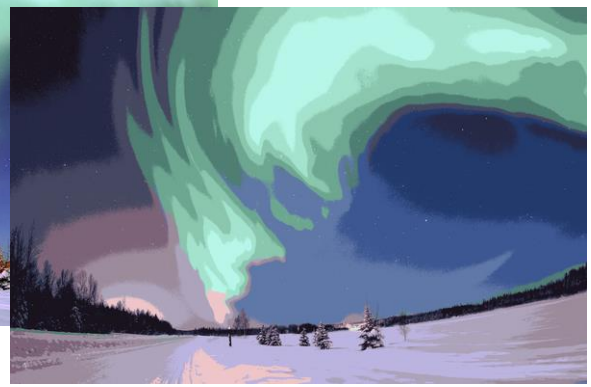
Ứng dụng của gom cụm

- Computer science: image segmentation, recommender system, anomaly detection
- Social network analysis: clustering community, search result grouping
- Business marketing: dividing consumers into market segments



Original

After clustering



Source: Andrew Ng, Wikipedia

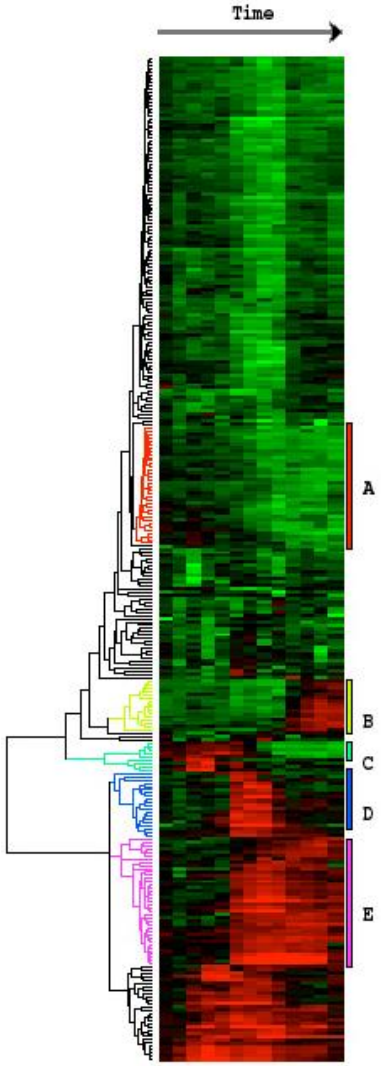
Ứng dụng của gom cụm

- Image segmentation
  - Mục tiêu: phân chia ảnh thành các vùng có ý nghĩa hoặc giống nhau trực quan

Source: James Hayes

Ứng dụng của gom cụm

- Gom cụm dữ liệu biểu diễn gene
- Mục tiêu: tìm ra những mẫu gen tương tự nhau

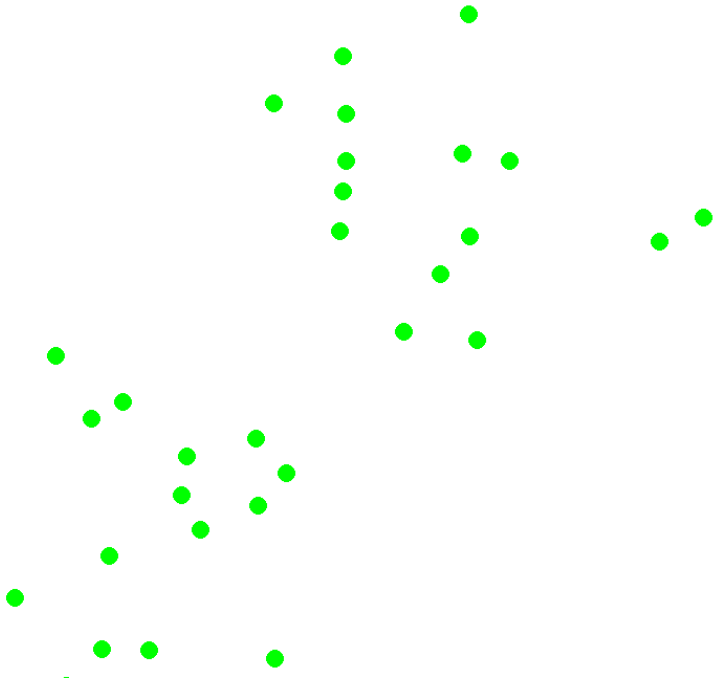


Source: Eisen et al, PNAS 1998

Thuật toán K-mean

- Input: số cụm K, m mẫu dữ liệu
- Mục tiêu: tìm các cụm sao cho khoảng cách giữa mẫu dữ liệu tới trung tâm là ngắn nhất

- Bước 1: khởi tạo K điểm trung tâm
- Bước 2: phân các điểm dữ liệu vào cụm gần nhất
- Bước 3: tính lại điểm trung tâm
- Lặp cho tới khi hội tụ



Thuật toán

- Khởi tạo ngẫu nhiên K điểm trung tâm: $\mu_1, \mu_2, \ldots, \mu_K$
- Lặp tới khi điểm trung tâm không đổi:
  - Lặp i = 1 tới m
    - $c^{(i)}$ = chỉ số của điểm trung tâm mà mẫu dữ liệu $x^{(i)}$ gần nhất
  - Lặp k = 1 tới K
    - μ_k = trung bình của các mẫu dữ liệu được phân vào cụm k

Hàm mục tiêu

- Giả sử:
  - $c^{(i)}$: cụm của mẫu x(i)
  - μ_k: điểm trung tâm của cụm k
  - $\mu_{c^{(i)}}$: điểm trung tâm của cụm mà x(i) được gán vào
- Hàm chi phí:

$$J(c^{(1)}, ..., c^{(m)}, \mu_1, ..., \mu_K) = \frac{1}{m} \sum_{i=1}^{m} \| x^{(i)} - \mu_{c^{(i)}} \|^2$$

- Mục tiêu:

$$\min_{c^{(1)}, ..., c^{(m)}, \mu_1, ..., \mu_K} J(c^{(1)}, ..., c^{(m)}, \mu_1, ..., \mu_K)$$

Thuật toán

- Khởi tạo ngẫu nhiên K điểm trung tâm: $\mu_1, \mu_2, \ldots, \mu_K$
- Lặp tới khi điểm trung tâm không đổi:
  - Lặp i = 1 tới m $\min_{c^{(i)}} J(\ldots)$
    - $c^{(i)}$ = chỉ số của điểm trung tâm mà mẫu dữ liệu $x^{(i)}$ gần nhất
  - Lặp k = 1 tới K $\min_{\mu_k} J(\ldots)$
    - μ_k = trung bình của các mẫu dữ liệu được phân vào cụm k

Khởi tạo trung tâm

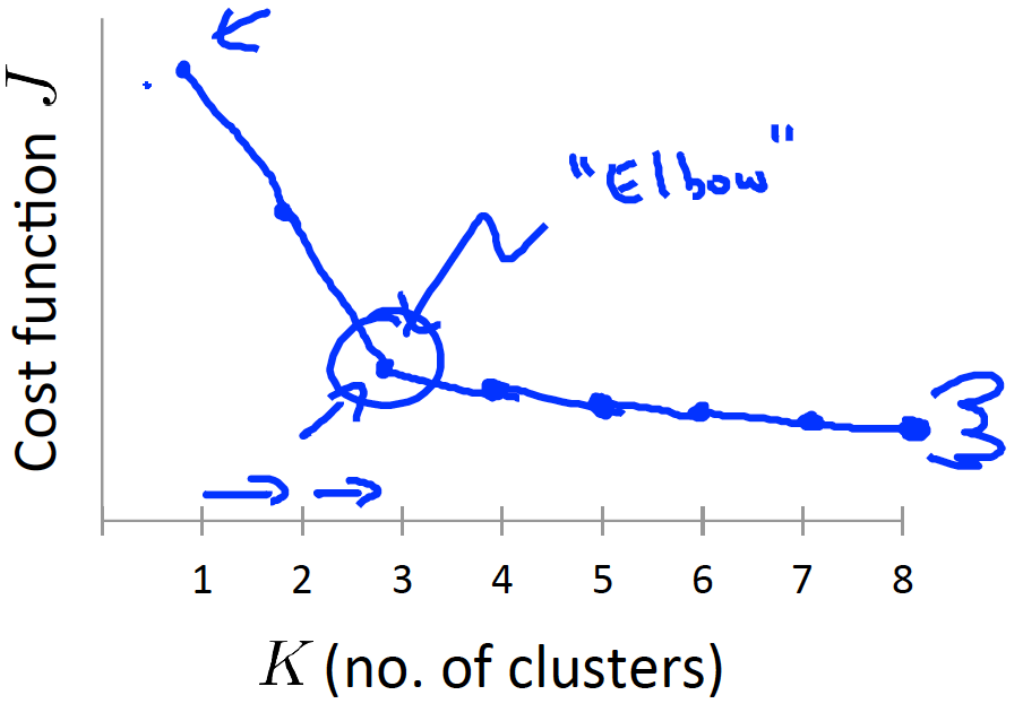
- Lặp i = 1 tới 100
  - Khởi tạo ngẫu nhiên K điểm trung tâm
  - Chạy thuật toán K-mean
  - Tính chi phí

$$J(c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K)$$

- Chọn các gom cụm có chi phí nhỏ nhất

Chọn số điểm trung tâm K

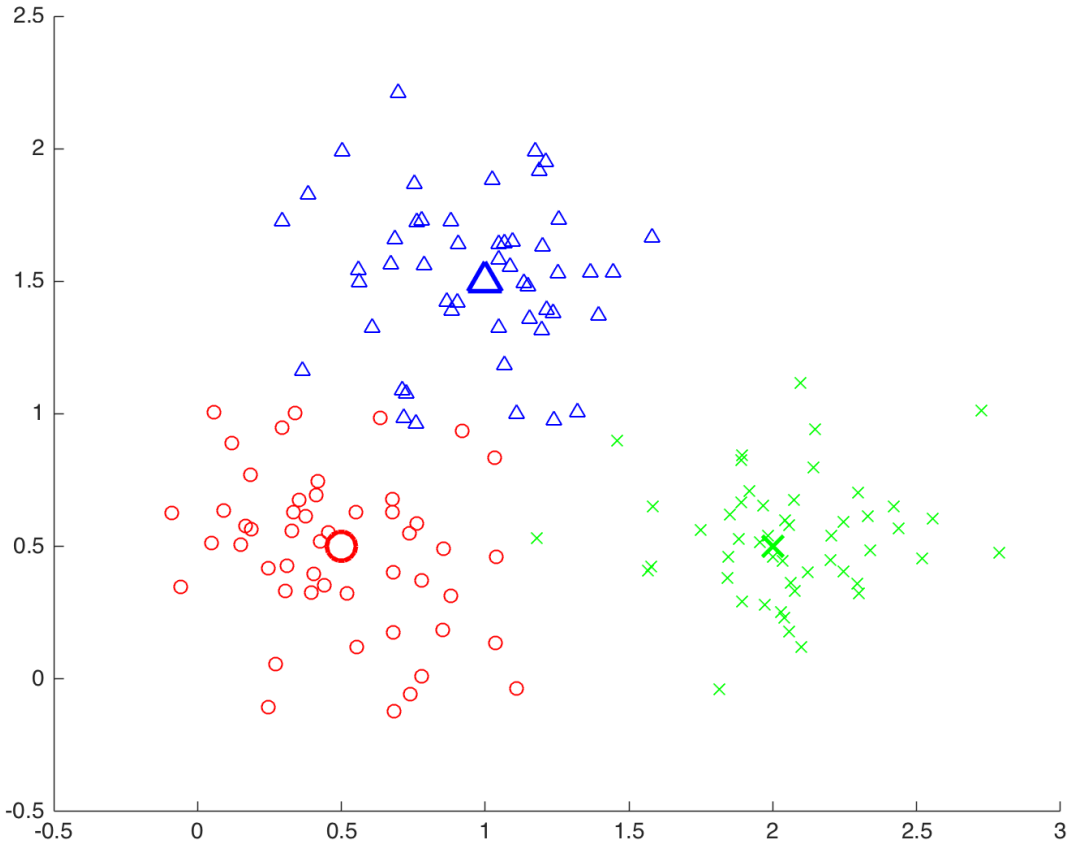
- Phương pháp Elbow: chọn K tại vị trí mà chi phí không đổi sau đó

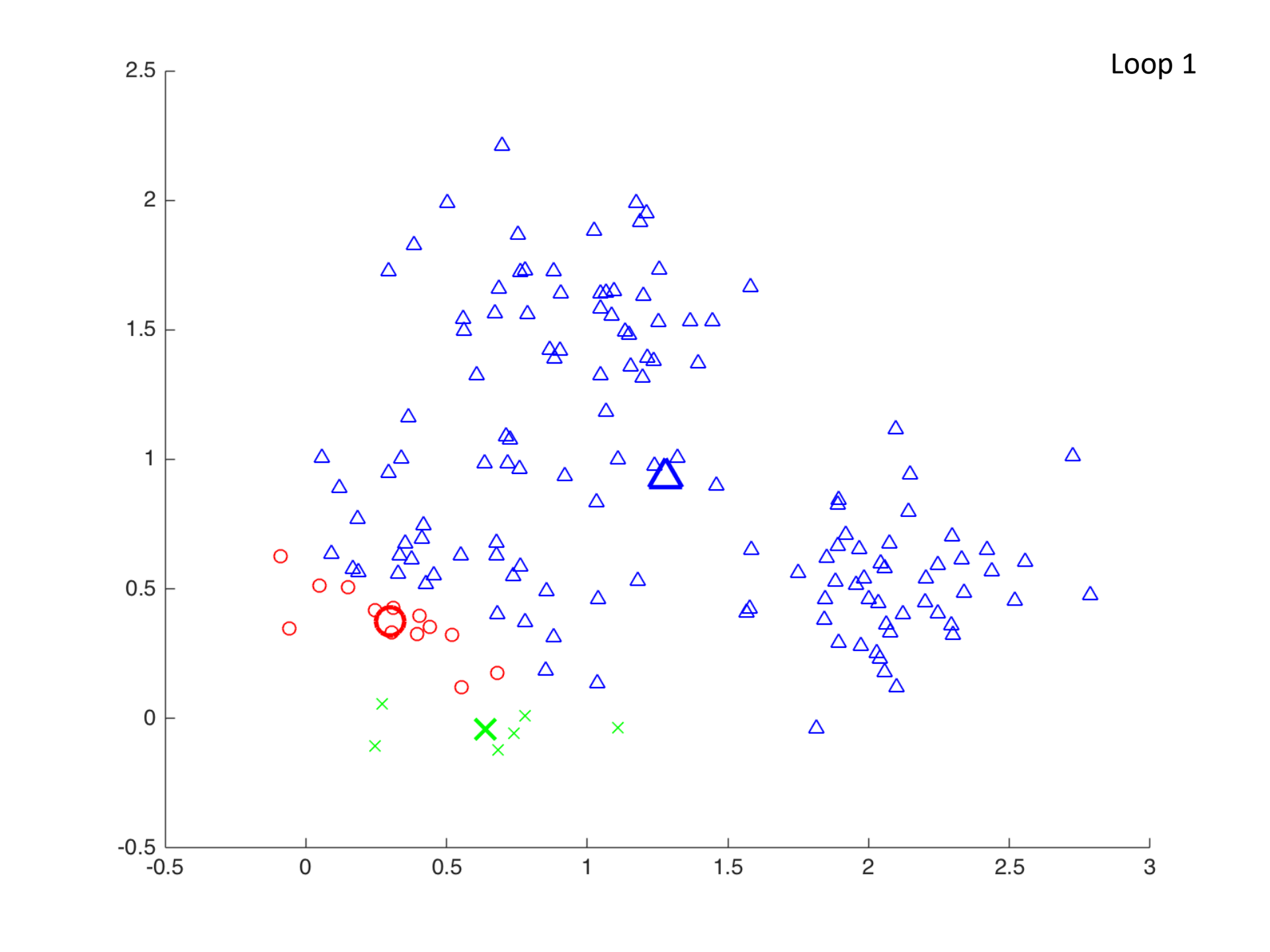


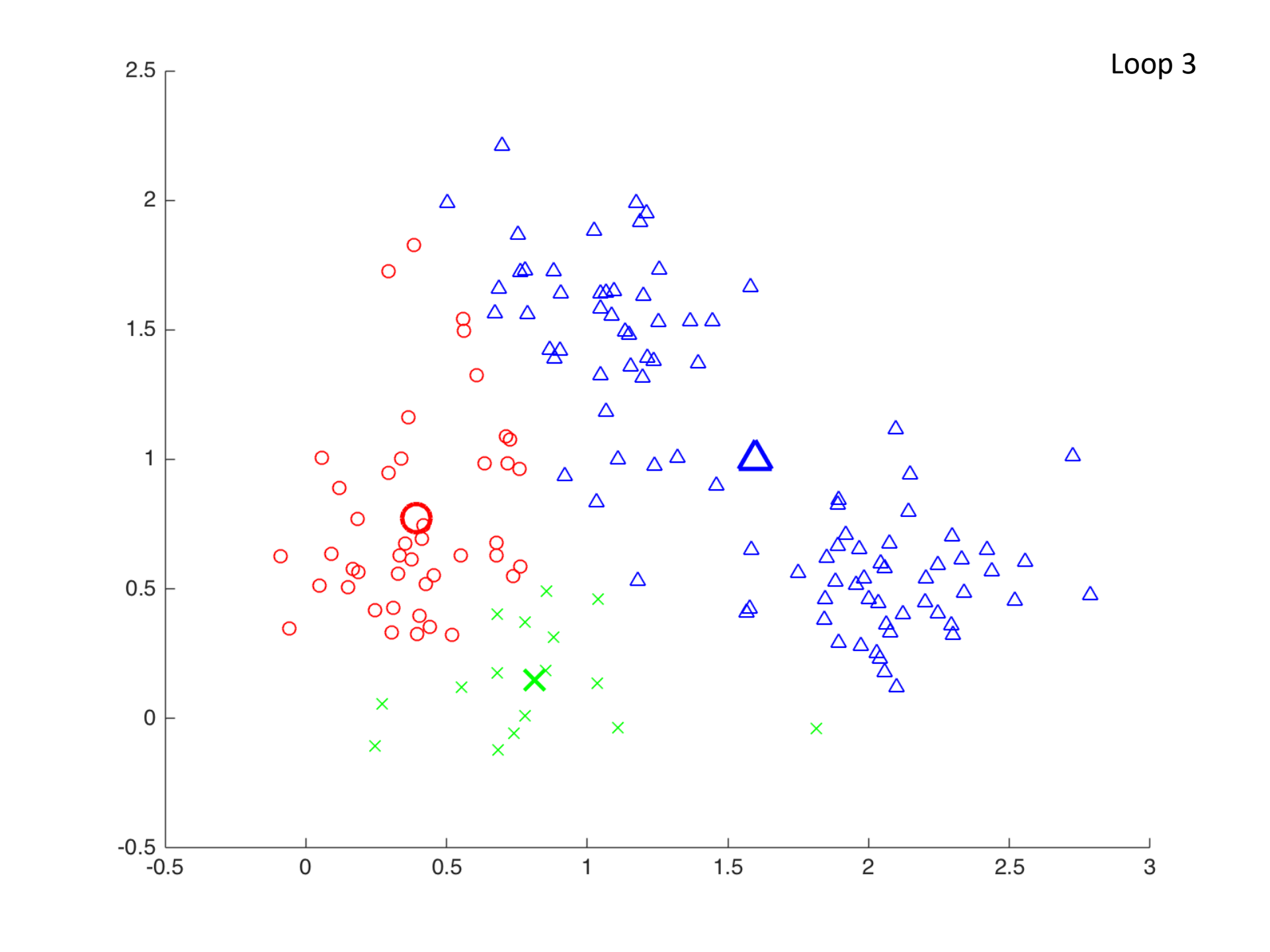
Các khoảng cách khác

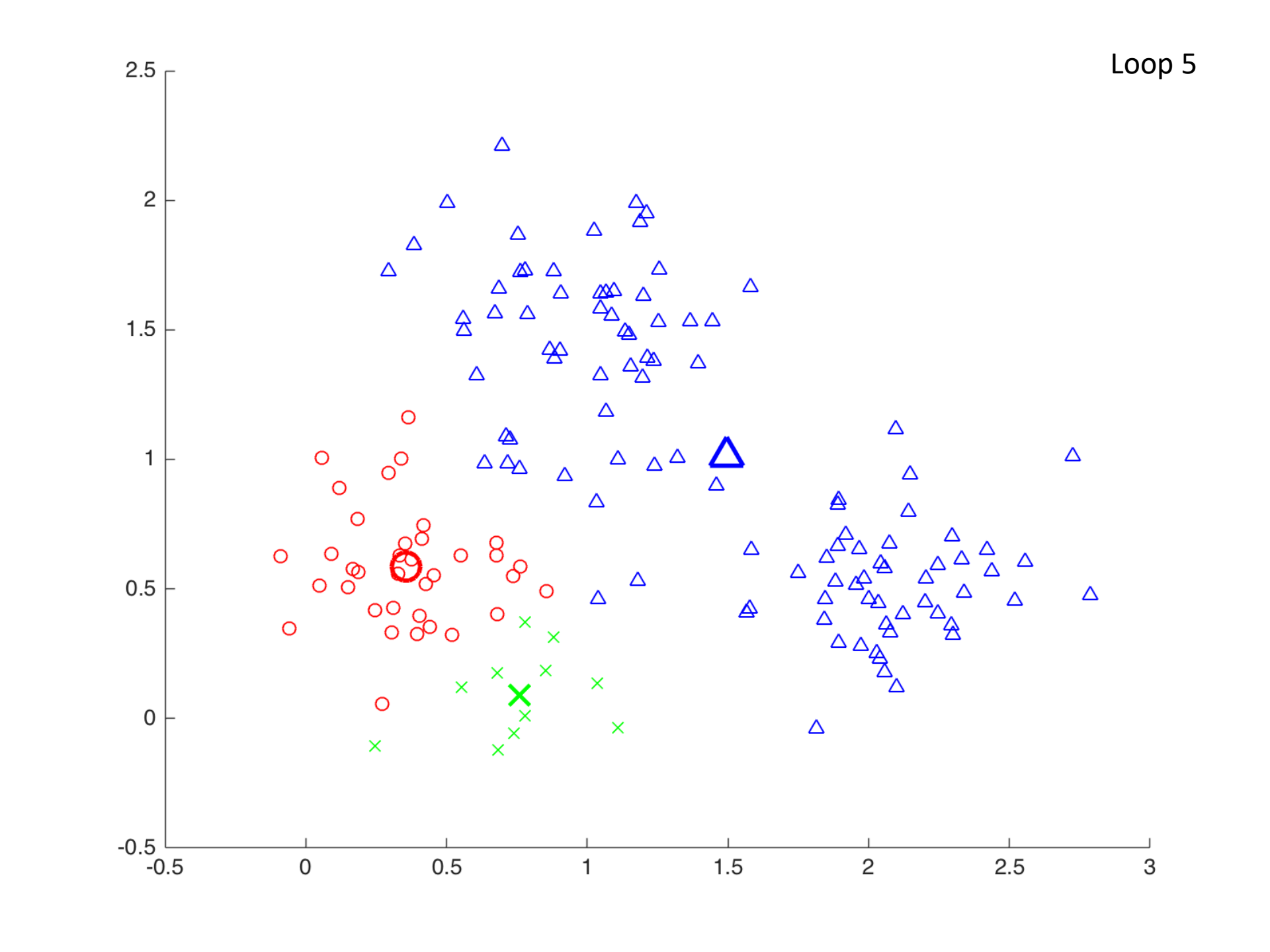
- Khoảng cách Euclide
  - $d(x,y) = \sqrt{\sum_{i=i}^{n} |x_i - y_i|^2}$
- Khoảng cách Manhattan
  - $d(x,y) = \sum_{i=1}^{n} |x_i - y_i|$, n: số đặc trưng
- Maximum norm
  - $d(x,y) = \max_{1 \le i \le n} |x_i - y_i|$, n: số đặc trưng
- Khoảng cách cosine
  - $d(x,y) = 1 - \frac{x^T y}{\|x\|\|y\|}$, d có giá trị từ 0 đến 2
- Khoảng cách Hamming
  - Số thành phần khác nhau giữa 2 véc tơ x và y
  - Ví dụ: 2 véc tơ (0, 1, **1**) và (0, 1, **0**) có khoảng cách Hamming là 1

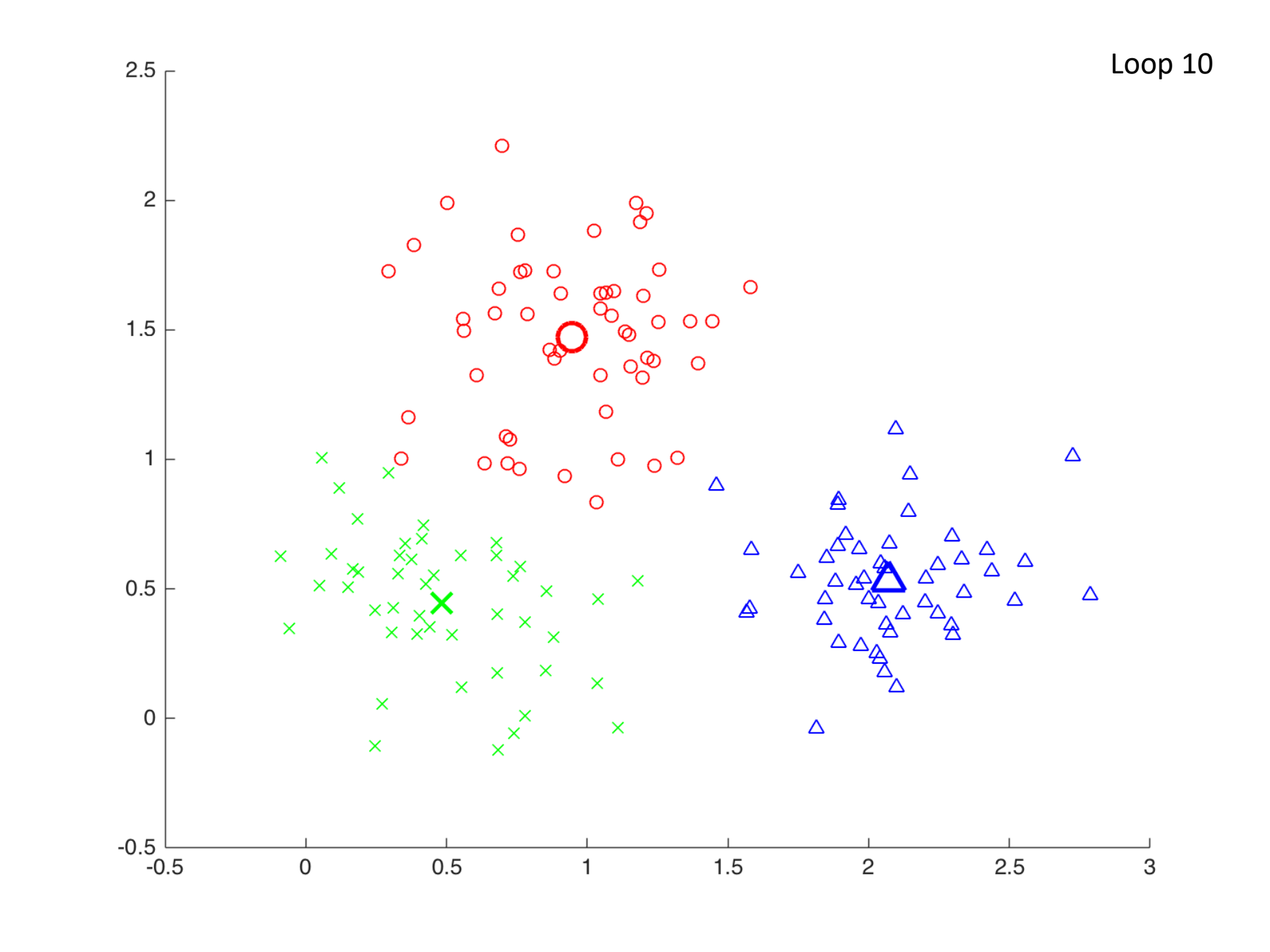
Ví dụ




Loop 1
2.5
2
1.5
1
0.5
0
-0.5
-0.5
0
0.5
1
1.5
2
2.5
3


Loop 3
2.5
2
1.5
1
0.5
0
-0.5
-0.5
0
0.5
1
1.5
2
2.5
3


Loop 5
2.5
2
1.5
1
0.5
0
-0.5
-0.5
0
0.5
1
1.5
2
2.5
3


Loop 10
2.5
2
1.5
1
0.5
0
-0.5
-0.5
0
0.5
1
1.5
2
2.5
3

Ưu điểm của k-mean

- Tìm ra các cụm có variance nhỏ
- Đơn giản và nhanh
- Dễ cài đặt

Khuyết điểm của k-mean

- Cần phải chọn tham số K trước
- Bị ảnh hưởng bởi outliers
- Dễ rơi vào cực tiểu địa phương
- Phụ thuộc lớn vào việc khởi tạo các cụm ban đầu
- Có thể chậm. Độ phức tạp của mỗi lần lặp: O(Kmn), m là số mẫu, n là số chiều