KHAI PHÁ DỮ LIỆU

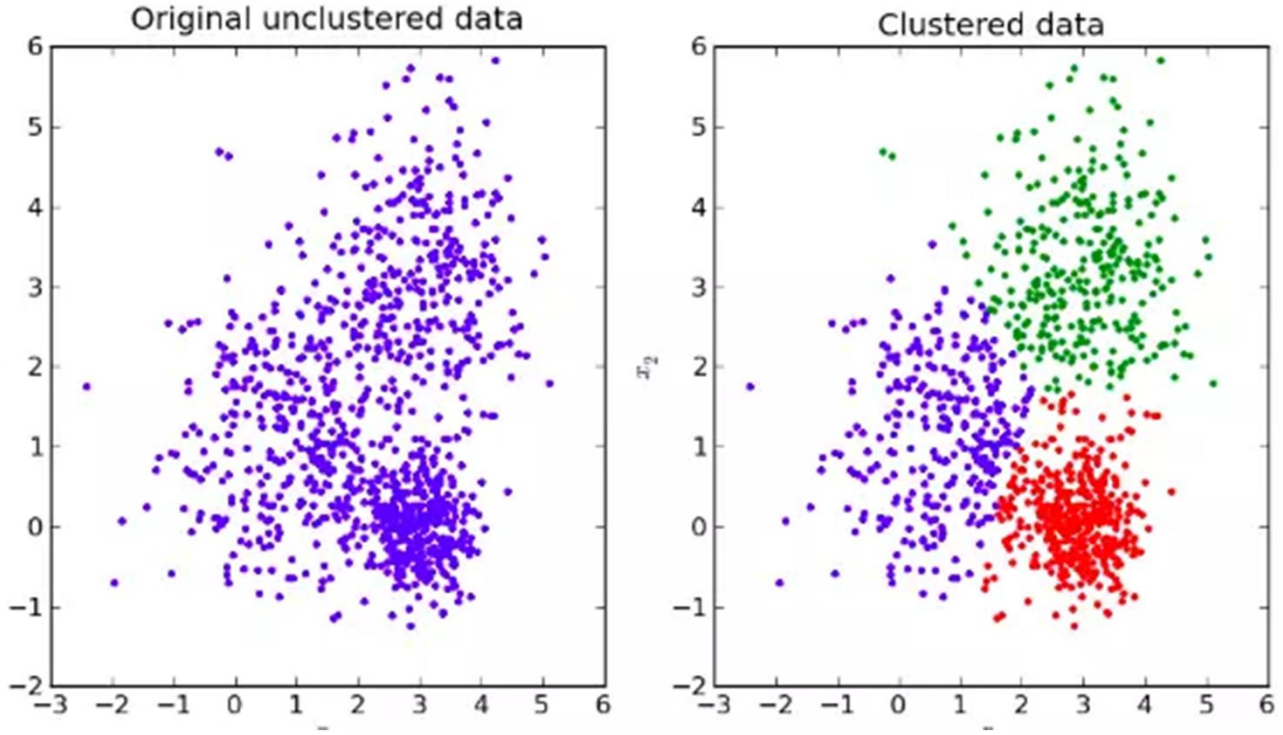
Trường Đại học Nha Trang
Khoa Công nghệ thông tin
Bộ môn Hệ thống thông tin
Giáo viên: TS.Nguyễn Khắc Cường

CHỦ ĐỀ 5

PHÂN CỤM

Phân cụm

- Phân cụm:
  - Các dữ liệu không biết nhãn
  - Tìm cách gom các dữ liệu "CÓ SỰ TƯƠNG ĐỒNG" lại thành một nhóm (cụm) → unsupervised learning

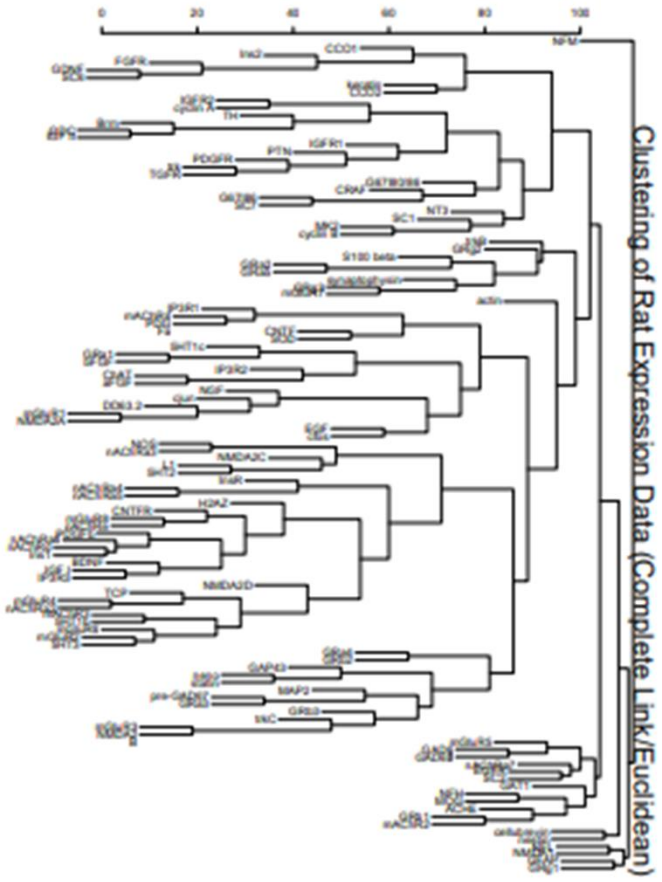


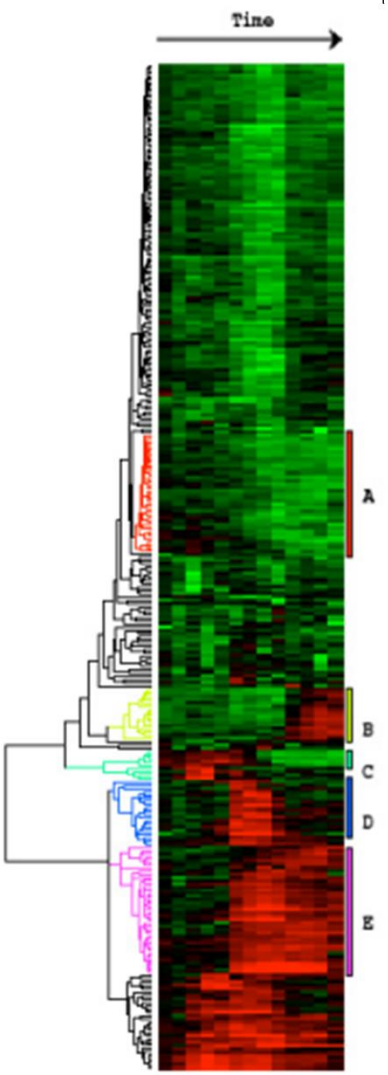
Phân cụm

- Đặc điểm:
  - Không biết trước được số cụm
  - Cùng một dữ liệu, nhưng phương pháp phân cụm khác nhau sẽ tạo thành các kết quả khác nhau
- Một số ứng dụng của phân cụm
  - Image segmentation

Phân cụm

- Một số ứng dụng của phân cụm
  - Clustering gene / expression data
  - market research
  - pattern recognition
  - data analysis
  - image processing
  - . . .





Phân cụm

- Idea để đánh giá các data "tương đồng"
  - Một trong các idea đó là dùng distance
  - Một số các công thức tính distance:

$$d_{euclidean} = \sqrt{\sum_{i=1}^{n} \left(\vec{x}_i - \vec{y}_i\right)^2}$$

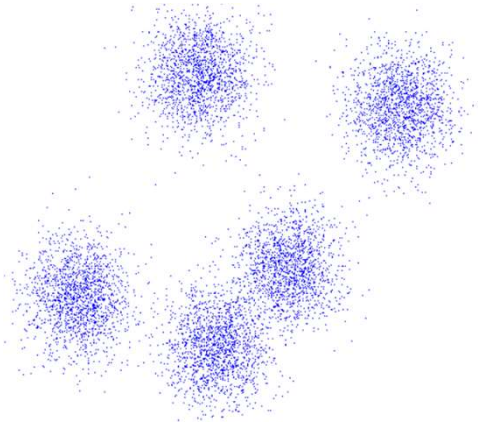
$$d_{manhattan} = \sum_{i=1}^{n} \left|\vec{x}_i - \vec{y}_i\right|$$

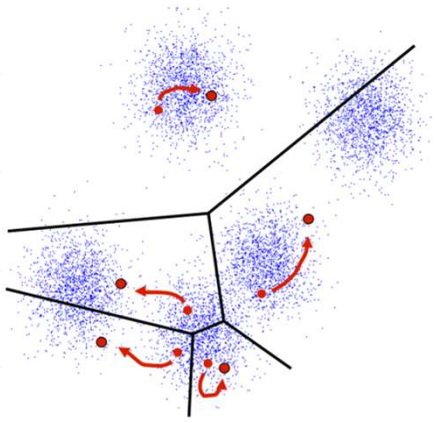
$$d_{minkowski} = \left(\sum_{i=1}^{n} \left|\vec{x}_i - \vec{y}_i\right|^p \right)^{\frac{1}{p}}$$

- K-mean
  - Quá trình "học"
    - Khởi tạo (tự chọn) số nhóm K
    - Chọn ngẫu nhiên k training data → làm k "điểm trung tâm"
    - Xét từng training data còn lại

Phân cụm

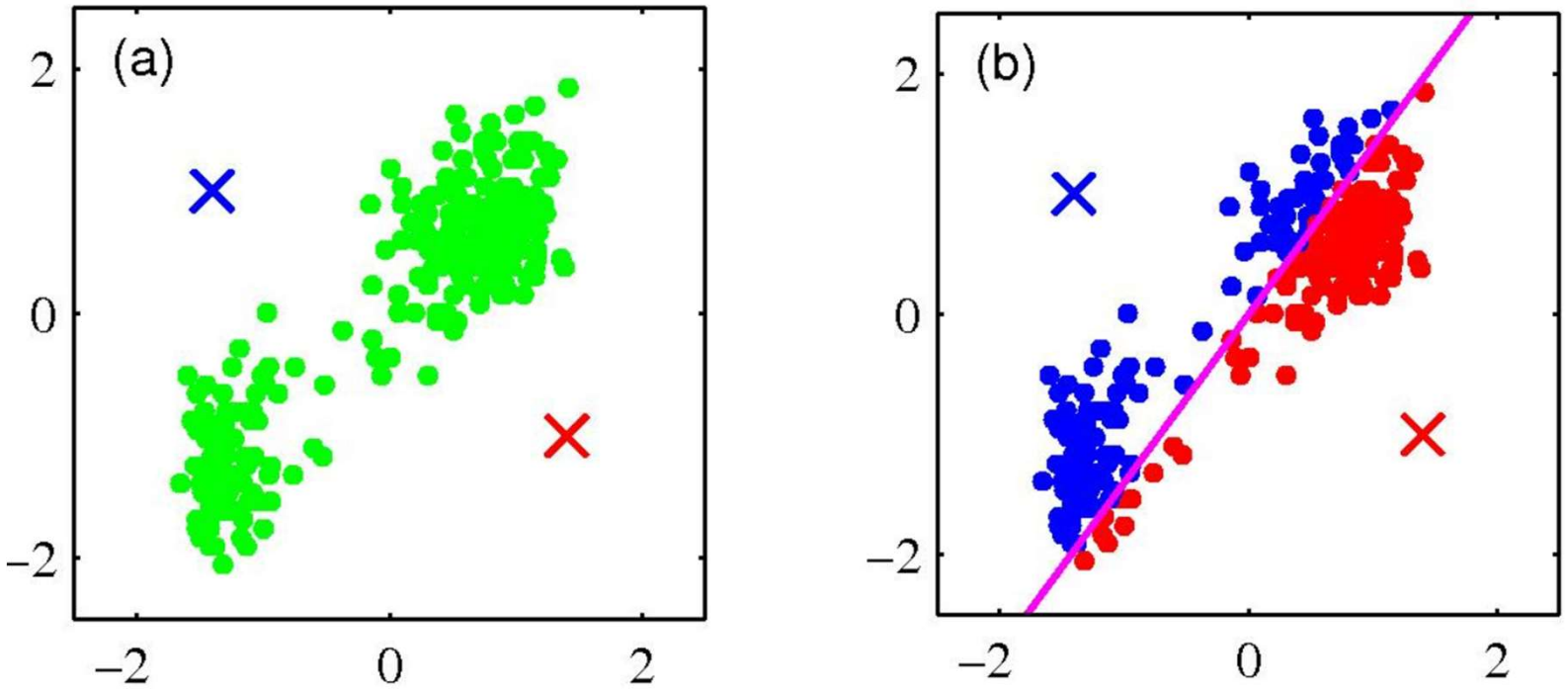
- K-mean
  - Quá trình "học"
    - Gán từng training data vào cùng nhóm với "điểm trung tâm" gần nhất
    - Sau khi xét hết các training data → tính lại "điểm trung tâm" mới của từng cụm (tìm vector trung bình)
    - Xét lại từng training data, gán từng data cho k "điểm trung tâm" mới đó
    - Lặp lại cho đến khi không có sự thay đổi thì dừng lại





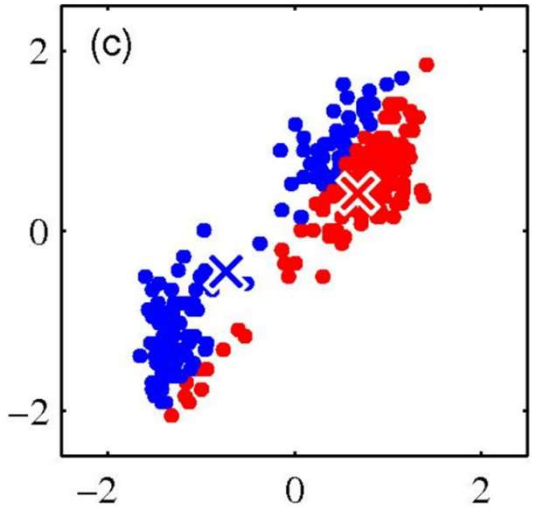
Phân cụm

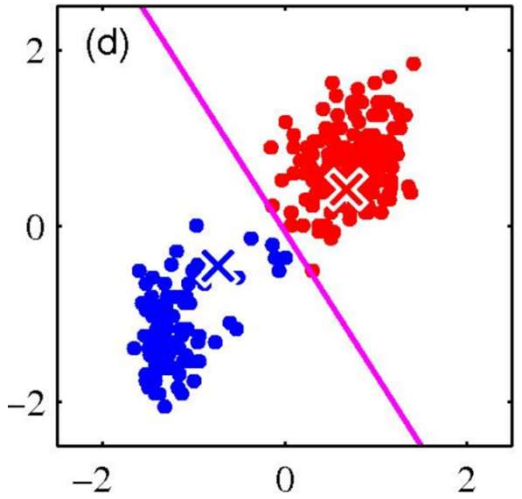
- Ví dụ K-means với K = 2



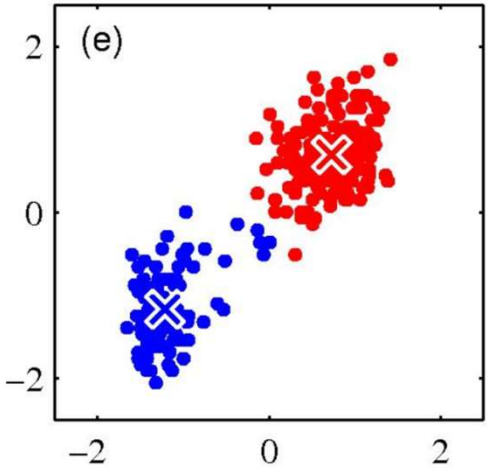
Phân cụm

- Ví dụ K-means với K = 2





→ Kết quả:



Q / A