**BÀI TẬP 2:**

1. **Phương pháp K-means**
2. **Phương pháp PCA**

**Yêu cầu:** Gửi file mã nguồn (.py hoặc .ipnb) và file báo cáo kết quả chương trình .doc/docx đi kèm.

**Bài 1.** Giả sử ta có 4 loại thuốc A,B,C,D. Mỗi loại thuốc được biểu diễn bởi 2 đặc trưng X và Y như sau:

|  |  |  |
| --- | --- | --- |
| Object | X (weight index) | Y (pH) |
| A | 1 | 1 |
| B | 3 | 2 |
| C | 4 | 2 |
| D | 5 | 4 |

Nhóm các thuốc đã cho vào 2 nhóm (K=2) dựa vào các đặc trưng của chúng. Yêu cầu:

1. Thực hiện bằng tay các bước của thuật toán K-means với K=2.
2. Lập trình python kiểm tra lại kết quả trên.

Bài 2. Cho dữ liệu là 6 điểm sau:

|  |  |
| --- | --- |
| X | Y |
| 1 | 1 |
| 2 | 1 |
| 4 | 2 |
| 5 | 3 |
| 5 | 4 |
| 6 | 5 |

1. Thực hiện bằng tay các bước của thuật toán K-means với K=2,
2. Lập trình python kiểm tra lại kết quả trên. Hiển thị các cụm dữ liệu trên đồ thị.

**Bài 3.** Có 20 học sinh có điểm thi ba môn Toán Lý, Hóa như sau:

|  |  |  |  |
| --- | --- | --- | --- |
| maHS | Toan | Ly | Hoa |
| A01 | 6 | 5 | 10 |
| A02 | 9 | 10 | 8 |
| A03 | 8 | 3 | 7 |
| A04 | 6 | 4 | 3 |
| A05 | 8 | 5 | 4 |
| A06 | 6 | 5 | 9 |
| A07 | 7 | 8 | 8 |
| A08 | 5 | 4 | 6 |
| A09 | 7 | 3 | 4 |
| A10 | 9 | 10 | 7 |
| A11 | 2 | 5 | 3 |
| A12 | 8 | 8 | 6 |
| A13 | 3 | 9 | 4 |
| A14 | 3 | 10 | 9 |
| A15 | 7 | 6 | 10 |
| A16 | 10 | 7 | 6 |
| A17 | 7 | 4 | 4 |
| A18 | 7 | 9 | 5 |
| A19 | 4 | 10 | 7 |
| A20 | 1 | 3 | 4 |

Lập trình python:

1. Hãy phân cụm thành 3 cụm theo phương pháp K-means.
2. Thực hiện giảm chiều dữ liệu với PCA với số thành phần chính PC =2. Mô tả kết quả giảm chiều trên hình vẽ. Thực hiện phân cụm dữ liệu với phương pháp K-means với K =3

**Bài 4.** Dữ liệu dùng để phân cụm trong dữ liệu dùng để phân loại khách hàng của ngân hàng (file dữ liệu **bank-data.csv, ví dụ lấy từ** <https://learnersdesk.weebly.com/weka-tutorials.html>). Dữ liệu **bank-data.csv** gồm có 11 thuộc tính và 600 khách hàng (instances). Dưới đây là cấu trúc của **bank-data.csv**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| @attribute | age | numeric |  |  |  |  |
| @attribute | sex | {FEMALE | MALE} |  |  |  |
| @attribute | region | {INNER\_CITY | TOWN | RURAL | SUBURBAN} | |
| @attribute | income | numeric |  |  |  |  |
| @attribute | married | {NO | YES} |  |  |  |
| @attribute | children | {0 | 1 | 2 | 3} |  |
| @attribute | car | {NO | YES} |  |  |  |
| @attribute | save\_act | {NO | YES} |  |  |  |
| @attribute | current\_act | {NO | YES} |  |  |  |
| @attribute | mortgage | {NO | YES} |  |  |  |
| @attribute | pep | {YES | NO} |  |  |  |

1. Thực hiện phân loại khách hàng của ngân hàng bằng phương pháp K-means với K=5.
2. Thực hiện giảm chiều dữ liệu bằng PCA còn 2-3 thành phần chính?

Ví dụ giải bài 1:

Chọn C1(1, 1) là tâm của cụm 1, C2 (3, 2) là tâm cụm 2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Object | X (weight index) | Y (pH) | K/c từ c1 | K/c từ c2 | Cụm |
| A | 1 | 1 | 0 | 2.236 | 1 |
| B | 3 | 2 | 2.236 | 0 | 2 |
| C | 4 | 2 | 3.162 | 1 | 2 |
| D | 5 | 4 | 5 | 2.828 | 2 |

Tính lại tâm:

C1=(1, 1)

C2=(4, 8/3)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Object | X (weight index) | Y (pH) | K/c từ c1 | K/c từ c2 | Cụm |
| A | 1 | 1 | 0 | 3.432 | 1 |
| B | 3 | 2 | 2.236 | 1.202 | 2 |
| C | 4 | 2 | 3.162 | 0.667 | 2 |
| D | 5 | 4 | 5 | 1.667 | 2 |

Ta thấy không còn có sự di chuyển của điểm dữ liệu từ cụm này sang cụm khác, cho nên thuật toán kết thúc. Kết quả là:

|  |  |  |
| --- | --- | --- |
| Tâm cụm 1 | C1=(1, 1) | A |
| Tâm cụm 2 | C2=(4, 8/3) | B, C, D |