

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

KHAI THÁC DỮ LIỆU

**BÀI TẬP LAP 5**

**NGUYỄN THỊ NGỌC HÀ: 17520421**

Mục lục

[BÀI TẬP 1: 3](#_Toc43059645)

[a. Sử dụng thuật toán K-Means. 3](#_Toc43059646)

[b. Sử dụng mạng Kohonen với các thông số: epochs = 10, R =0, α = 0,8 6](#_Toc43059647)

[c. So sánh kết quả thu được từ thuật toán k-Means và mạng Kohonen với nhau. 14](#_Toc43059648)

[d. Tính toán một vị trí tiềm năng mới xuất hiện: 15](#_Toc43059649)

[BÀI TẬP 2: 16](#_Toc43059650)

[a. Sử dụng thuật toán k-Means: 16](#_Toc43059651)

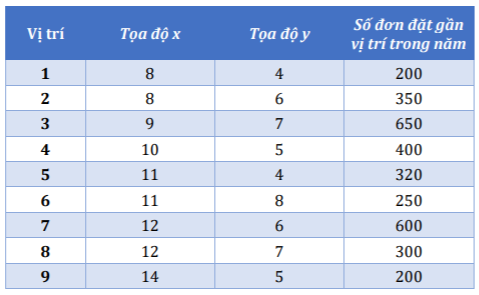
[b. Sử dụng mạng Kohonen với các thông số: epochs = 10, R =0 và α = 0,8: 17](#_Toc43059652)

[c. So sánh kết quả thu được từ thuật toán k-Means và mạng Kohonen với nhau: 22](#_Toc43059653)

[d. Tính toán một đơn hàng mới: 22](#_Toc43059654)

# BÀI TẬP 1:

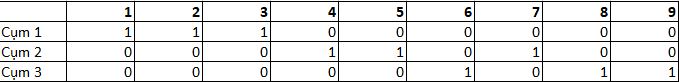
Trong quá trình thống kê doanh thu, một công ty kinh doanh chuỗi cửa hàng Pizza phát hiện những vị trí có nhu cầu cao nhưng chưa có cửa hàng trong khu vực. Tọa độ tương đối của những vị trí tìm năng đó được cho trong bảng sau:



Công ty này muốn xây dựng 3 cửa hàng Pizza mới trong những khu vực trên nhằm mở rộng chuỗi cửa hàng và phục vụ tốt hơn cho những khu vực này. Sinh viên hãy giúp công ty bằng cách tìm ra vị trí đặt 3 cửa hàng sao cho thuận lợi nhất và chỉ ra cụm khách hàng của từng cửa hàng với các yêu cầu cụ thể sau:

## Sử dụng thuật toán K-Means.

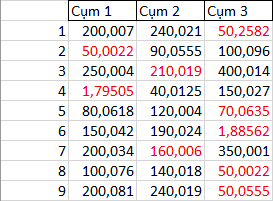
Khởi tạo ma trận phân hoạch



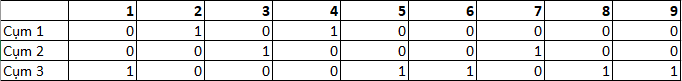
Vector trọng tâm của các cụm lúc này là:

* Cụm 1: V1 = (8.33, 5.67, 400)
* Cụm 2: V2 = (11, 5, 440)
* Cụm 3: V3 = (12.33, 6.67, 250)

Khoảng cách Euclide từ các vị trí đến lần lượt các cụm là:



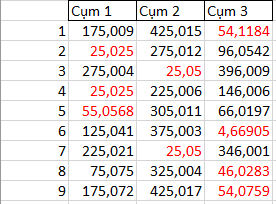
Ma trận phân hoạch các điểm thuộc cụm



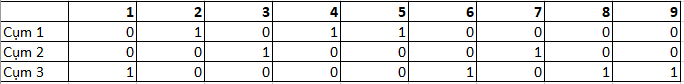
Vector trọng tâm của các cụm lúc này là:

* Cụm 1: V1 = (9, 5.5, 375)
* Cụm 2: V2 = (10.5, 6.5, 625)
* Cụm 3: V3 = (11.2, 5.6, 254)

Khoảng cách Euclide từ các vị trí đến lần lượt các cụm là:



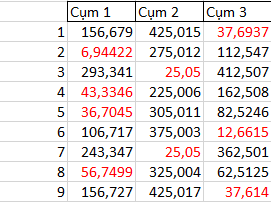
Ma trận phân hoạch các điểm thuộc cụm



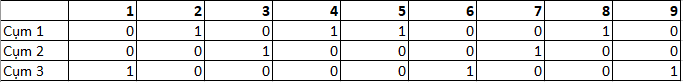
Vector trọng tâm của các cụm lúc này là:

* Cụm 1: V1 = (9.67, 5, 356.67)
* Cụm 2: V2 = (10.5, 6.5, 625)
* Cụm 3: V3 = (11.25, 6, 237.5)

Khoảng cách Euclide từ các vị trí đến lần lượt các cụm là:



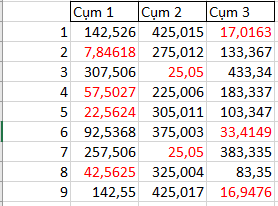
Ma trận phân hoạch các điểm thuộc cụm



Vector trọng tâm của các cụm lúc này là:

* Cụm 1: V1 = (10.25, 5.5, 342.5)
* Cụm 2: V2 = (10.5, 6.5, 625)
* Cụm 3: V3 = (11, 5.67, 216.67)

Khoảng cách Euclide từ các vị trí đến lần lượt các cụm là:



Như vậy các điểm thuộc cụm không thay đổi, thuật toán dừng lại, với kết quả sau:

* Vị trí 2, 4, 5 và 8 thuộc Cụm 1
* Vị trí 3 và 7 thuộc Cụm 2
* Vị trí 1, 6 và 9 thuộc Cụm 3

## Sử dụng mạng Kohonen với các thông số: epochs = 10, R =0, α = 0,8

Mạng Kohonen (Hay còn gọi là SOM) là một mạng nơ ron truyền thẳng sử dụng phương pháp học không gian giám sát, áp dụng trong việc ánh xạ để giảm kích thước dữ liệu đầu vào. Từ tập các đối tượng trong không gian nhiều chiều ở đầu vào, mạng Kohonen có số chiều nhỏ hơn (thường là 2 chiều) được sử dụng dể đặc trưng cho chúng ở đầu ra của thuật toán.

Để áp dụng mạng Kohonen vào bài toán gom cụm thì mạng ở đầu ra có số nơ ron tương ứng với số cụm. Mỗi nơ ron chứa một vecto trọng số có số chiều bằng số chiều của vecto dữ liệu đầu vào. Cụ thể:

* Đầu vào thuật toán: Đối tượng cần gom cụm là tập các vecto trong không gian n chiều, số lần lặp của thuạt toán epochs, bán kính vùng lân cận R và tốc độ học α.
* Đầu ra thuật toán: Bản đồ mạng Kohonen với mỗi nơ ron trên mạng đặc trưng cho một cụm.

Thuật toán trải qua các bước như sau:

|  |  |
| --- | --- |
| **Bước** | **Thao tác** |
| 0 | Khởi tạo giá trị của các vecto trọng số. Gán giá trị cho biến R và α |
| 1 | Nếu chưa thỏa điều kiện dừng thì lặp lại từ bước 2 -> 8 |
| 2 | Với mỗi vecto đầu vào x thực hiện bước 3 -> 5 |
| 3 | Với mỗi nơ ron trên mạng j, tính khoảng cách Euclide đến x theo công thức sau: |
| 4 | Tìm nơ ron J sao cho khoảng cách từ điểm dữ liệu đang xét x đến J là ngắn nhất. |
| 5 | Duyệt qua tất cả những vecto trọng số thuộc nơ ron láng giềng j của J và xét tất cả các chiều I, tiến hành cập nhật  Wij(*new*) = Wij(*old*) + α(xi - Wij(*old*)) |
| 6 | Cập nhật lại tốc độ α. Điều này giúp làm giảm số lần lặp. |
| 7 | Giảm bán kính vùng lân cận vào thười điểm thích hợp |
| 8 | Kiểm tra điều kiện dừng. Thông thường thì đó là lúc giá trị của tốc độ học đã rất nhỏ hoặc giá trị của vecto trọng số hội tụ nên việc cập nhật không còn thay đổi nhiều. |

Với bài tập trên, ta cần xác định các tham số của thuật toán:

* Dữ liệu đầu vào là thông tin của 9 vị trí, mỗi thông tin có thể xem như là một vecto trong không gian n = 3 chiều.
* Bán kính vùng lân cận là R = 0, điều này có nghĩa là khi thay đổi trọng số của một nơ ron thì những nơ ron lân cận không bị ảnh hưởng.
* Số lần lặp của thuật toán là epochs = 10
* Tốc độ học là α = 0,8
* Vì mục đích của bài toán là gom thành 3 cụm nên mạng Kohonen ở đầu ra gồm 3 nơ ron. Mỗi nơ ron chứa 1 trọng số là vecto wi có số chiều = n.

Khởi tạo giá trị của các vecto trọng số:

W1 = (10, 6, 343)

W2 = (11, 7, 625)

W3 = (11, 6, 217)

Ở lần lặp thứ nhất:

* Xét vecto đầu tiên (Vị trí 1) x1

Khoảng cách từ x1 đến W1 là:

D11 = = 143

Tương tự, ta tính được:

D12 = 425 và D13 = 17,4

Như vậy, nơ ron có trọng số W3 là nơ ron có khoảng cách đến x1 ngắn nhất. Tiến hành cập nhật lại trọng số của W3

W31(new) = 11 + 0,8 x (8 – 11) = 8,6

W32(new) = 6 + 0,8 x (4 – 6) = 4,4

W33(new) = 217 + 0,8 x (200 – 217) = 203,4

Lúc này, do R = 0 nên không cần cập nhật lại vùng lân cận, W2 và W1 sẽ giữ nguyên giá trị, ta có:

W1 = (10, 6, 343)

W2 = (11, 7, 625)

W3 = (8.6, 4.4, 203.4)

* Xét vector đầu tiên (Vị trí 1)

=15.21 ; = 150.008; =150.06

Cập nhật lại vector

* = (8.5, 6, 353)
  + = (9.4, 5.2, 200)
  + = (5, 3, 500)
* Xét vector đầu tiên (Vị trí 1)

=297 ; = 450; =150.11

Cập nhật lại vector

* = (8.5, 6, 353)
  + = (9.4, 5.2, 200)
  + = (8.2, 6.2, 620)
* Xét vector đầu tiên (Vị trí 1)

=47.03 ; = 200.001; =220.01

Cập nhật lại vector

* = (9.7, 5.2, 390.6)
  + = (9.4, 5.2, 200)
  + = (8.2, 6.2, 620)
* Xét vector đầu tiên (Vị trí 1)

=70.62 ; = 120.02; =300.02

Cập nhật lại vector

* = (10.74, 4.24, 334.12)
  + = (9.4, 5.2, 200)
  + = (8.2, 6.2, 620)
* Xét vector đầu tiên (Vị trí 1)

=84.2 ; = 50.1; =370.01

Cập nhật lại vector

* = (10.74, 4.24, 334.12)
  + = (10.68, 7.44, 240)
  + = (8.2, 6.2, 620)
* Xét vector đầu tiên (Vị trí 1)

=265.89 ; = 360; =20.36

Cập nhật lại vector

* = (10.74, 4.24, 334.12)
  + = (10.68, 7.44, 240)
  + = (11.24, 6.04, 604)
* Xét vector đầu tiên (Vị trí 1)

=34.25 ; = 60.02; =304

Cập nhật lại vector

* = (11.75, 6.45, 546.82)
  + = (10.68, 7.44, 240)
  + = (11.24, 6.04, 604)
* Xét vector đầu tiên (Vị trí 1)

=346.83 ; = 40.21; =404.01

Cập nhật lại vector

* = (11.75, 6.45, 546.82)
  + = (13.34, 5.49, 208)
  + = (11.24, 6.04, 604)
* Giảm α một nửa, = 0.4

**Lặp lần 2:**

* Xét vector đầu tiên (Vị trí 1)

=346.85 ; = 9.73; =404.01

Cập nhật lại vector

* = (11.75, 6.45, 546.82)
  + = (11.204, 4.89, 204.8)
  + = (11.24, 6.04, 604)
* Xét vector đầu tiên (Vị trí 1)

=196.86 ; = 145.24; =254.02

Cập nhật lại vector

* = (11.75, 6.45, 546.82)
  + = (9.92, 5.334, 262.88)
  + = (11.24, 6.04, 604)
* Xét vector đầu tiên (Vị trí 1)

=103.22 ; = 387.12; =46.06

Cập nhật lại vector

* = (11.75, 6.45, 546.82)
  + = (9.92, 5.334, 262.88)
  + = (10.344, 6.42, 622.4)
* Xét vector đầu tiên (Vị trí 1)

=146.84 ; = 137.12; = 222.405

Cập nhật lại vector

* = (11.75, 6.45, 546.82)
  + = (9.95, 5.2004, 317.73)
  + = (10.344, 6.42, 622.4)
* Xét vector đầu tiên (Vị trí 1)

=226.83 ; = 2.77; = 302.41

Cập nhật lại vector

* = (11.75, 6.45, 546.82)
  + = (10.37, 4.72, 318.64)
  + = (10.344, 6.42, 622.4)
* Xét vector đầu tiên (Vị trí 1)

=296.83 ; = 68.72; = 372.4

Cập nhật lại vector

* = (11.75, 6.45, 546.82)
  + = (10.62, 6.03, 291.18)
  + = (10.344, 6.42, 622.4)
* Xét vector đầu tiên (Vị trí 1)

=53.18 ; = 308.82; = 22.47

Cập nhật lại vector

* = (11.75, 6.45, 546.82)
  + = (10.62, 6.03, 291.18)
  + = (11.006, 6.25, 613.44)
* Xét vector đầu tiên (Vị trí 1)

=246.82 ; = 8.98; = 313.44

Cập nhật lại vector

* = (11.75, 6.45, 546.82)
  + = (11.17, 6.42, 294.71)
  + = (11.006, 6.25, 613.44)
* Xét vector đầu tiên (Vị trí 1)

=346.83 ; = 94.76; = 413.45

Cập nhật lại vector

* = (11.75, 6.45, 546.82)
  + = (12.302, 5.85, 256.83)
  + = (11.006, 6.25, 613.44)
* Giảm α một nửa, = 0.2

**Lặp lại tươg tự 8 lần, ta được kết quả:**

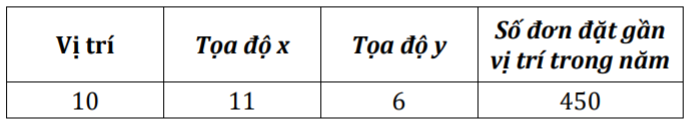
* Vị trí 1, 6, 9 thuộc cụm 1
* Vị trí 2, 4, 5, 8 thuộc cụm 2
* Vị trí 3, 7 thuộc cụm 3

## So sánh kết quả thu được từ thuật toán k-Means và mạng Kohonen với nhau.

Từ kết quả thu được ở câu 1 và 2, có thể thấy ở cả hai thuật toán các đơn hàng trong một cụm đều giống nhau. Các cụm tương ứng ở hai thuật toán là:

|  |  |
| --- | --- |
| k-Means | Mạng Kohonen |
| Cụm 1 | Cụm 1 |
| Cụm 2 | Cụm 2 |
| Cụm 3 | Cụm 3 |

## Tính toán một vị trí tiềm năng mới xuất hiện:



Sinh viên hãy giúp công ty xác định cửa hàng nào trong 3 cửa hàng trên sẽ phục vụ cho vị trí này tốt nhất.

* Sử dụng thuật toán k-Means:

Khoảng cách Euclide từ các vị trí đến lần lượt các cụm là

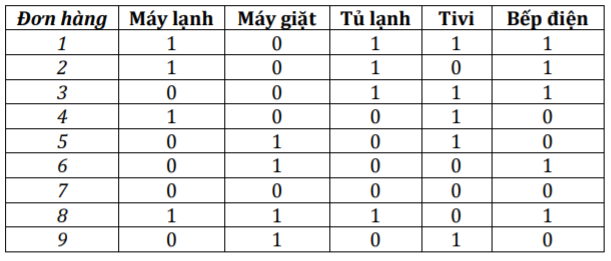


Vì khoảng cách từ Vị trí 10 đến tâm Cụm 1 là ngắn nhất nên có thể kết luận được rằng Vị trí 10 có sự tương đồng với các Vị trí 2, 4, 5 và 8.

* Sử dụng thuật toán mạng Kohonen:

# BÀI TẬP 2:

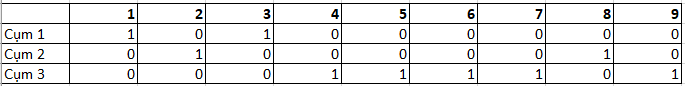
Một Website Thương mại điện tử chuyên kinh doanh thiết bị gia dụng thu thập được những đơn hàng sau:



Công ty dự định xây dựng một hệ thống khuyến nghị dựa trên việc gom cụm những đơn hàng trên thành 3 nhóm có độ tương đồng với nhau. Từ đó gợi ý cho việc mua hàng của khách hàng dựa trên sự tương đồng này. Sinh viên hãy tham gia xây dựng hệ thống khuyến nghị này bằng cách thực hiện những yêu cầu sau:

## Sử dụng thuật toán k-Means:

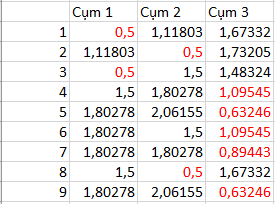
Khởi tạo ma trận phân hoạch



Vector trọng tâm của các cụm lúc này là:

* Cụm 1: V1 = (0.5, 0, 1, 1, 1)
* Cụm 2: V2 = (1, 0.5, 1, 0, 1)
* Cụm 3: V3 = (0.2, 0.6, 0, 0.6, 0.2)

Khoảng cách Euclide từ các vị trí đến lần lượt các cụm là:



Như vậy các điểm thuộc cụm không thay đổi, thuật toán dừng lại, với kết quả sau:

* Đơn hàng 1 và 3 thuộc Cụm 1
* Đơn hàng 2 và 8 thuộc Cụm 2
* Đơn hàng 4, 5, 6, 7 và 9 thuộc Cụm 3

## Sử dụng mạng Kohonen với các thông số: epochs = 10, R =0 và α = 0,8:

Với bài tập trên, ta cần xác định các tham số của thuật toán:

* Dữ liệu đầu vào là thông tin của 9 đơn hàng, mỗi thông tin có thể xem như là một vecto trong không gian n = 5 chiều.
* Bán kính vùng lân cận là R = 0, điều này có nghĩa là khi thay đổi trọng số của một nơ ron thì những nơ ron lân cận không bị ảnh hưởng.
* Số lần lặp của thuật toán là epochs = 10
* Tốc độ học là α = 0,8
* Vì mục đích của bài toán là gom thành 3 cụm nên mạng Kohonen ở đầu ra gồm 3 nơ ron. Mỗi nơ ron chứa 1 trọng số là vecto wi có số chiều = n.

Khởi tạo giá trị của các vecto trọng số:

W1 = (1,0,0,1,1)

W2 = (0,0,1,1,0)

W3 = (1,1,0,0,0)

Ở lần lặp thứ nhất:

* Xét vecto đầu tiên (Vị trí 1) x1

Khoảng cách từ x1 đến W1 là:

D11 = = 1

Tương tự, ta tính được:

D12 = 1,41 và D13 = 2

Như vậy, nơ ron có trọng số W1 là nơ ron có khoảng cách đến x1 ngắn nhất. Tiến hành cập nhật lại trọng số của W1

W11(new) = 1 + 0,8 x (1 – 1) = 1

W12(new) = 0 + 0,8 x (0 – 0) = 0

W13(new) = 0 + 0,8 x (1 – 0) = 1

W14(new) = 1 + 0,8 x (1 – 1) = 1

W15(new) = 1 + 0,8 x (1 – 1) = 1

Lúc này, do R = 0 nên không cần cập nhật lại vùng lân cận, W2 và W1 sẽ giữ nguyên giá trị, ta có:

W1 = (1,0,1,1,1)

W2 = (0,0,1,1,0)

W3 = (1,1,0,0,0)

* Xét các vecto còn lại:

Tương tự như trên, sau khi xét lần lượt các vecto còn lại, ta có trọng số lúc này là:

W1 = (1, 0.8, 0.99, 0.04, 1)

W2 = (0, 0, 1, 1, 0.8)

W3 = (0, 0.84, 0, 0.81, 0.03)

* Trước khi kết thúc lần lặp đầu tiên, ta giảm tốc độ học đi một nửa:

α = = 0.4

Tiếp tục lặp lần 2, ta được kết quả:

W1 = (1, 0.69, 1, 0.01, 1)

W2 = (0.24, 0, 1, 1, 0.93)

W3 = (0.05, 0.7, 0, 0.6, 0.15)

* Trước khi kết thúc lần lặp thứ 2, ta giảm tốc độ học đi một nửa:

α = = 0.2

Tiếp tục lặp lần 3, ta được kết quả:

W1 = (1, 0.53, 1, 0.01, 1)

W2 = (0.24, 0, 1, 1, 0.97)

W3 = (0.07, 0.58, 0, 0.45, 0.21)

* Trước khi kết thúc lần lặp thứ 3, ta giảm tốc độ học đi một nửa:

α = = 0.1

Tiếp tục lặp lần 4, ta được kết quả:

W1 = (1, 0.53, 1, 0.01, 1)

W2 = (0.28, 0, 1, 1, 0.97)

W3 = (0.11, 0.6, 0, 0.50, 0.20)

* Trước khi kết thúc lần lặp thứ 4, ta giảm tốc độ học đi một nửa:

α = = 0.05

Tiếp tục lặp lần 5, ta được kết quả:

W1 = (1, 0.60, 1, 0, 1)

W2 = (0.34, 0, 1, 1, 0.98)

W3 = (0.13, 0.62, 0, 0.56, 0.20)

* Trước khi kết thúc lần lặp thứ 5, ta giảm tốc độ học đi một nửa:

α = = 0.025

Tiếp tục lặp lần 6, ta được kết quả:

W1 = (1, 0.59, 1, 0, 1)

W2 = (0.35, 0, 1, 1, 0.98)

W3 = (0.14, 0.62, 0, 0.56, 0.20)

* Trước khi kết thúc lần lặp thứ 6, ta giảm tốc độ học đi một nửa:

α = = 0.0125

Tiếp tục lặp lần 7, ta được kết quả:

W1 = (1, 0.59, 1, 0, 1)

W2 = (0.35, 0, 1, 1, 0.98)

W3 = (0.14, 0.62, 0, 0.56, 0.20)

* Trước khi kết thúc lần lặp thứ 7, ta giảm tốc độ học đi một nửa:

α = = 0.00625

Tiếp tục lặp lần 8, ta được kết quả:

W1 = (1, 0.59, 1, 0, 1)

W2 = (0.35, 0, 1, 1, 0.98)

W3 = (0.14, 0.62, 0, 0.56, 0.19)

* Trước khi kết thúc lần lặp thứ 8, ta giảm tốc độ học đi một nửa:

α = = 0.003125

Tiếp tục lặp lần 9, ta được kết quả:

W1 = (1, 0.59, 1, 0, 1)

W2 = (0.36, 0, 1, 1, 0.98)

W3 = (0.14, 0.62, 0, 0.56, 0.20)

* Trước khi kết thúc lần lặp thứ 9, ta giảm tốc độ học đi một nửa:

α = = 0.001563

Tiếp tục lặp lần 10, ta được kết quả:

W1 = (1, 0.59, 1, 0, 1)

W2 = (0.36, 0, 1, 1, 0.98)

W3 = (0.15, 0.62, 0, 0.56, 0.20)

Tính khoảng cách từ mỗi điểm đến các vecto, ta kết luận được:

* + Đơn hàng 2 và 8 thuộc cụm đại diện bởi vecto cụm 1
  + Đơn hàng 1 và 3 thuộc cụm đại diện bởi vecto cụm 2
  + Đơn hàng 4, 5, 6, 7, 9 thuộc cụm đại diện bởi vecto cụm 3

## So sánh kết quả thu được từ thuật toán k-Means và mạng Kohonen với nhau:

Từ kết quả thu được ở câu 1 và 2, có thể thấy ở cả hai thuật toán các đơn hàng trong một cụm đều giống nhau. Các cụm tương ứng ở hai thuật toán là:

|  |  |
| --- | --- |
| k-Means | Mạng Kohonen |
| Cụm 1 | Cụm 2 |
| Cụm 2 | Cụm 1 |
| Cụm 3 | Cụm 3 |

## Tính toán một đơn hàng mới:

Một khách đang thực hiện đặt hàng trên hệ thống đã mua Máy lạnh, tủ lạnh. Theo em, những mặt hàng nào nên được khuyến nghị dựa theo kết quả đã tính được từ các câu trên.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Đơn hàng | Máy lạnh | Máy giặt | Tủ lạnh | Tivi | Bếp điện |
| 10 | 1 | 0 | 1 | 0 | 0 |

* Sử dụng thuật toán k-Means:

Áp dụng kết quả gom cụm từ thuật toán k-Means, chúng ta phải tính khoảng cách Euclide từ điểm đại diện cho Đon hàng 10 đến tâm các cụm. Sau đó chọn ra cụm có khoảng cách gần nhất, đó chính là cụm chứa các Đơn hàng có sự tương đồng nhiều nhất đối với Đơn hàng 10.

Khoảng cách đến cụm 1:

D101 = = 1.5

Tương tự, khoảng cách đến cụm 2 và cụm 3 lần lượt là D102 = 1.118; D103 = 1.549.

Vì khoảng cách từ Đơn hàng 10 đến tâm cụm 2 là ngắn nhất nên có thể kết luận được rằng Đơn hàng 10 có sự tương đồng với các Đơn hàng 2,8.

Đơn hàng 2,8 gồm các món hàng sau:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Đơn hàng | Máy lạnh | Máy giặt | Tủ lạnh | Ti vi | Bếp điện |
| 2 | 1 | 0 | 1 | 0 | 1 |
| 8 | 1 | 1 | 1 | 0 | 1 |

* Sử dụng thuật toán mạng Kohonen:

Khoảng cách từ x1 đến W1 là:

D11 = = 1,16

Tương tự, ta tính được:

D12 = 1,54 và D13 = 1,57

Vì khoảng cách từ Đơn hàng 10 tới nơ ron 1 là ngắn nhất nên có thể kết luận được rằng Đơn hàng 10 có sự tương đồng với các Đơn hàng 2 và 8

Như vậy dù với phương pháp nào thì chúng ta cũng có cùng một kết luận về sự tương đồng của Đơn hàng 2 và 8.

* Dựa vào kết quả thu được, 2 đơn hàng trên đều mua các mặt hàng Máy lạnh, Tủ lạnh và Bếp điện. Vì vậy ta có thể gợi ý cho khách hàng của Đơn hàng 10 mua thêm Bếp điện.