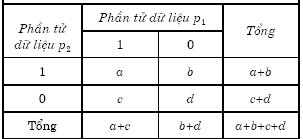
**1. Độ đo tương đồng:**

- Giá trị nhị phân: Lập ma trận kề mô tả số lượng các thuộc tính cùng giá trị và không cùng giá trị.



Sau đó sử dụng độ đo Jaccard:



- Giá trị rời rạc [0,m]: Chuyển m giá trị về nhị phân sau đó áp dụng độ đo Jaccard.

- Giá trị thực sử dụng độ đo cosin 2 vector: cos(p1 , p2) =

**2. Độ đo khác biệt:**

- Giá trị nhị phân: Với thuộc tính đối xứng (các giá trị của nó có sự tương đương), dựa vào ma trận kề, khoảng cách được tính:



Với thuộc tính bất đối xứng (các giá trị của nó có ý nghĩa khác nhau), khoảng cách được tính:



- Giá trị rời rạc [0,m]: Có thể chuyển về m giá trị về nhị phân và sử dụng công thức trên. Hoặc đếm số thuộc tính có giá trị giống nhau và khoảng cách được tính:

(Giả sử có q thuộc tính giống nhau ở p1 và p2)

- Giá trị thực: Sử dụng khoảng cách Mincowski



Khoảng cách Manhatan và Euclide là trường hợp q=1 và q=2 của công thức trên.

**\*Tính chất:**

- Tính xác định dương: d(pi,pj) > 0 nếu i≠j và d(pi,pj)=0 nếu i=j

- Tính đối xứng: d(pi,pj) = d(pj,pi)

- Tính bất đẳng thức tam giác: d(pi,pk) + d(pk,pj) ≠ d(pi,pj)

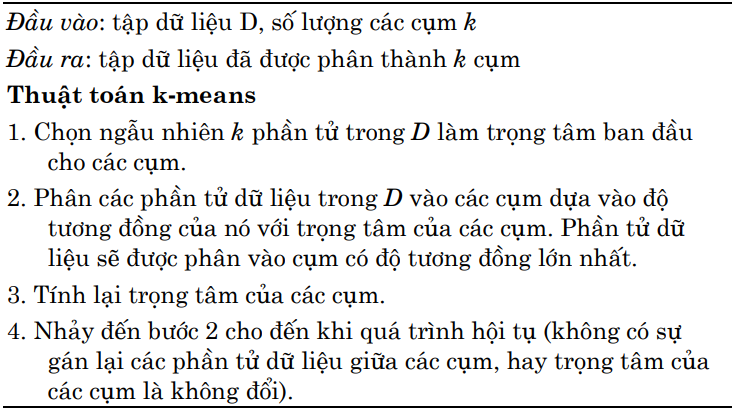
**3. Phân cụm:**

- Là bài toán thuộc lớp các bài toán học không giám sát (unsupervised learning) với nhiệm vụ gán nhãn cho các dữ liệu trong một tập dữ liệu lớn.

- Các phương pháp phân cụm: Phân cụm phẳng và phân cụm phân cấp, Phân cụm theo mật độ, Phân cụm dựa trên mô hình, …

**4. K-means:**

- Là thuật toán thuộc lớp phân cụm phẳng. Được mô tả như sau:



Kết quả cuối cùng của thuật toán phụ thuộc nhiều vào cách lựa chọn k phần tử ban đầu làm tâm cụm. Việc chọn lựa này hoàn toàn ngẫu nhiên nên kết quả thu được có thể khác nhau. Vì vậy cần chạy thuật toán một số lần để thu được kết quả phù hợp. Độ phức tạp của thuật toán: O(nkt) với n là số lượng dữ liệu, k là số lượng cụm và t là số lần thực hiện thuật toán.

- Sau khi huấn luyện, chúng ta cần kiểm tra quy luật phân cụm thông qua biểu đồ phân cụm. Đối với các bộ dữ liệu hai chiều và ba chiều, biểu diễn chúng khá dễ dàng trên mặt phẳng hoặc siêu phẳng. Nhưng với các bộ dữ liệu có nhiều chiều hơn, chúng ta cần áp dụng các phương pháp giảm chiều dữ liệu trước khi đồ thị hóa. Tham khảo link sau để trực quan hóa dữ liệu sau khi thực hiện quá trình trên: <https://blog.exploratory.io/visualizing-k-means-clustering-results-to-understand-the-characteristics-of-clusters-better-b0226fb3dd10>.

**\* Các hạn chế của thuật toán:**

- Cần xác định trước số cụm k cho thuật toán. Chúng ta có thể thực hiện các phương pháp thử để lựa chọn số cụm như Elbow.

- Thuật toán không chính xác với các tập dữ liệu mất cân bằng, có hình dạng phức tạp (thuật toán phù hợp với miền dữ liệu mà các cụm phân bố theo hình cầu).

- Nhạy cảm với ngoại lệ: Việc xuất hiện ngoại lệ thường khiến cho tâm cụm bị sai lệch.

- Áp dụng thuật toán với trường hợp dữ liệu lớn hoặc giải thuật không hội tụ dẫn đến thời gian chạy chương trình có thể rất lớn. Sử dụng điều kiện dừng có thể làm giảm chất lượng của thuật toán.