**2. Giới thiệu thuật toán K-means và ý tưởng thực hiện**

**a) Tổng quan K-means clustering**

K-Mean là một thuật toán Unsupervised learning (Học không giám sát). Trong thuật toán K-means clustering, chúng ta không biết nhãn (label) của từng điểm dữ liệu.

Clustering (hay phân cụm) là nhiệm vụ phân chia một tập hợp các đối tượng (hay thành viên) thành các nhóm (hay các cụm) khác nhau dựa trên đặc điểm (hay thuộc tính) của đối tượng. Và rằng các thành viên của một nhóm sẽ có nhiều điểm tương đồng hơn so với những thành viên trong nhóm khác.

Ý tưởng đơn giản nhất về cluster (cụm) là tập hợp các điểm ở gần nhau trong một không gian nào đó (không gian này có thể có rất nhiều chiều trong trường hợp thông tin về một điểm dữ liệu là rất lớn). Hình bên dưới là một ví dụ về 3 cụm dữ liệu:

A diagram of different colored squares

Description automatically generated with medium confidence

**b) Ý tưởng và tóm tắt thuật toán**

* **Ý tưởng:**

Với mỗi đối tượng, chúng ta sẽ vector hoá để biểu diễn đặc trưng của đối tượng dưới dạng một data point. Và ta biết rằng, những điểm dữ liệu tương đồng về bản chất thì sẽ ở gần nhau hơn. Khi đó, bài toán phân cụm của ta sẽ là bài toán đi tìm những data point nằm gần nhau.

A diagram of a green circle with dots

Description automatically generated

Mỗi đối tượng đều có một hoặc nhiều thuộc tính riêng. K-means chính là dựa vào một số thuộc tính cụ thể mà đi gom cụm lại những đối tượng có tính chất tương đồng nhau.

Giả sử ta có điểm dữ liệu **xi** cần phân vào cụm K.

Khi đó ta có vector sai số giữa 2 điểm là (**mk – xi**). Ta mong đợi vector sai số càng gần với vector .

Triển khai bài toán, ta có thể dùng 1 trong các phương pháp tính khoảng cách như là Euclid hoặc Manhattan, khoảng cách cosine.

Mục đích cuối là tối ưu hóa hàm mất mát: . Và hàm mất mát tối ưu khi **xi** gần **mk** hơn.

* **Tóm tắt thuật toán:**

Bước 1: Chọn ngẫu nhiên K điểm bất kì trong tập huấn luyện để làm các tâm cụm ban đầu.

Bước 2: Phân nhóm các điểm dữ liệu vào cụm có tâm gần nó nhất:

Bước 3: Cập nhập tâm cụm bằng cách lấy trung bình cộng của các điểm dữ liệu:

Bước 4: Nếu tâm cụm ở bước 3 không thay đổi so với vòng lặp trước đó thì dừng thuật toán.

Bước 5: Quay lại bước 2.

* **Lưu đồ thuật toán:**

A diagram of a basic structure

Description automatically generated with medium confidence

**c) Ưu điểm, nhược điểm**

* **Ưu điểm:**
* K-means là thuật toán đơn giản, dễ dàng sử dụng tốt cho các bài toán phân cụm.
* K-means thực hiện phân cụm tốt mà không cần biết nhãn dữ liệu đầu vào. (Học không giám sát)
* K-means là nền tảng cho nhiều thuật toán phức tạp sau này.
* **Nhược điểm:**
* Số K cần được xác định trước. Ở nhiều bài toán, việc xác định được K không phải là dễ dàng, khi đó K-means sẽ không hiệu quả.
* K-means không đảm bảo tìm được nghiệm tối ưu toàn cục. Và nghiệm cuối cùng phụ thuộc hoàn toàn vào việc khởi tạo các tâm cụm ban đầu.
* K-means sẽ không hiệu quả nếu các cụm chêch lệch về số lượng điểm, phân bố dữ liệu không có dạng cầu, hay bài toán với 1 điểm dữ liệu có thể là con của 2 cụm.