**K-mean Clustering**

1. **Giới thiệu**

K-Means Clustering là thuật toán máy học không có giám sát (unsupervised learning) được sử dụng để phân chia một tập dữ liệu thành các nhóm (clusters) sao cho các điểm dữ liệu trong cùng một nhóm có đặc điểm giống nhau và các nhóm khác nhau có đặc điểm khác nhau.

***Unsupervised machine learning****: là một quá trình dạy máy học các dữ liệu không dán nhãn, chưa được phân loại và khởi tạo trên dữ liệu đó mà không cần có sự giám sát. Không cần dữ liệu đã được đào tạo trước đó, máy học trong trường hợp này chỉ cần sắp xếp dữ liệu thành các cụm dựa theo các điểm tương đồng, mẫu, biến thể.*

1. **Mục đích của K-means**

Mục đích của K-means là phân dữ liệu thành các cụm (cluster) khác nhau sao cho dữ liệu trong cùng một cụm có tính chất giống nhau. Bằng cách từ dữ liệu đầu vào và số lượng nhóm muốn tìm, chỉ ra center của mỗi nhóm và phân các điểm dữ liệu vào các nhóm tương ứng.[1]

1. **K-means clustering hoạt động như thế nào?** [2][3]

Trong thuật toán K-means mỗi cụm dữ liệu được đặc trưng bởi một tâm (centroid). Tâm là điểm đại diện nhất cho một cụm và có giá trị bằng trung bình của toàn bộ các quan sát nằm trong cụm.

Chúng ta sẽ dựa vào khoảng cách từ mỗi quan sát đến tâm để xác định nhãn cho chúng thuộc về tâm gần nhất. Ban đầu thuật toán sẽ khởi tạo ngẫu nhiên một số lượng xác định trước tâm cụm. Sau đó, tiến hành xác định nhãn cho từng điểm dữ liệu và tiếp tục cập nhật lại tâm cụm. Thuật toán sẽ dừng cho tới khi toàn bộ các điểm dữ liệu được phân về đúng cụm hoặc số lượng cập nhật tâm chạm ngưỡng.

Các bước của thuật toán k-Means được tóm tắt như sau:

1. Khởi tạo ngẫu nhiên k tâm cụm
2. Lặp lại quá trình cập nhật tâm cụm cho tới khi dừng:
3. Xác định nhãn cho từng điểm dữ liệu i dựa vào khoảng cách tới từng tâm cụm

A black text with a white background

Description automatically generated

**Trong công thức 2.a:**

* i: điểm dữ liệu thứ
*  : bình phương của khoảng cách Euclid giữa điểm dữ liệu i và tâm cụm j
* i: nhãn của điểm dữ liệu i , là chỉ số của cụm mà i thuộc về, được các định bằng cách tìm tâm cụm gần nhất.

1. Tính toán lại tâm cho từng cụm theo trung bình của toàn bộ các điểm dữ liệu trong một cụm:

A mathematical equation with numbers and symbols

Description automatically generated

**Trong công thức 2.b:**

* j: tâm cụm thứ
* **:** hàm chỉ thị, trả về giá trị 1 nếu điểm dữ liệu i thuộc về cụm , ngược lại trả về 0.
* i: điểm dữ liệu thứ
* Biểu thức tính toán trung bình có nghĩa là lấy tổng các điểm dữ liệu thuộc về cụm và chia cho số lượng các điểm dữ liệu đó để cập nhật lại tâm cụm j

Trong thuật toán trên thì tham số mà chúng ta cần chọn chính là số lượng cụm . Thời điểm ban đầu ta sẽ khởi tạo điểm dữ liệu một cách ngẫu nhiên và sau đó gán các tâm bằng giá trị của điểm dữ liệu này. Các bước trong vòng lặp ở bước thứ 2 thực chất là:

1. Gán nhãn cho mỗi điểm dữ liệu bằng với nhãn của tâm cụm gần nhất.
2. Dịch chuyển dần dần tâm cụm **j** tớitrung bình của những điểm dữ liệu mà được phân về

[1] Cống, H. (2017, January 1). **K-means clustering**. Retrieved from <https://machinelearningcoban.com/2017/01/01/kmeans/>

[2] Harris, N. (n.d.). **Visualizing k-means clustering**. Truy xuất từ <https://www.naftaliharris.com/blog/visualizing-k-means-clustering/>

[3] Ng, A., Jordan, M., & Weiss, Y. (2020). **CS229 Lecture Notes - K-means clustering**. Retrieved from <http://cs229.stanford.edu/notes2020spring/cs229-notes7a.pdf>