*Thuật toán K-Means*

K-Means là một thuật toán phân cụm phổ biến trong học máy, thuộc nhóm học không giám sát (unsupervised learning). Thuật toán này chia một tập dữ liệu thành K cụm, với mỗi cụm được đại diện bởi một tâm cụm (centroid). Mục tiêu của K-Means là phân chia dữ liệu sao cho các điểm dữ liệu trong cùng một cụm có độ tương đồng cao, trong khi các điểm dữ liệu thuộc các cụm khác nhau có độ tương đồng thấp.

**Nguyên lý hoạt động của K-Means**

1. **Khởi tạo các tâm cụm:** Ngẫu nhiên chọn K điểm trong không gian dữ liệu làm tâm cụm ban đầu.
2. **Gán nhãn cho các điểm dữ liệu:** Mỗi điểm dữ liệu được gán vào cụm có tâm cụm gần nhất.
3. **Cập nhật tâm cụm:** Tính toán lại vị trí của các tâm cụm bằng cách lấy trung bình của tất cả các điểm dữ liệu trong mỗi cụm.
4. **Lặp lại bước 2 và 3:** Tiếp tục lặp lại các bước 2 và 3 cho đến khi không có điểm dữ liệu nào thay đổi cụm hoặc đạt đến một số tiêu chí dừng khác.

**Ưu điểm của K-Means**

* **Đơn giản, dễ hiểu:** Thuật toán có ý tưởng trực quan và dễ dàng triển khai.
* **Hiệu quả:** Thời gian chạy của K-Means tương đối nhanh, đặc biệt với các tập dữ liệu lớn.
* **Linh hoạt:** Có thể áp dụng cho nhiều loại dữ liệu khác nhau.

**Nhược điểm của K-Means**

* **Phải xác định trước số cụm K:** Việc chọn số cụm K phù hợp là một vấn đề khó và ảnh hưởng lớn đến kết quả phân cụm.
* **Nhạy cảm với giá trị ban đầu của các tâm cụm:** Kết quả phân cụm có thể khác nhau tùy thuộc vào cách khởi tạo ban đầu.
* **Giả định dữ liệu phân bố tròn:** K-Means thường hoạt động tốt với các cụm có hình dạng gần tròn.
* **Không thể phát hiện các cụm có hình dạng phức tạp:** K-Means khó phát hiện các cụm có hình dạng bất thường hoặc có độ đậm đặc khác nhau.

**Các cải tiến của K-Means**

* **K-Means++:** Một phương pháp khởi tạo các tâm cụm tốt hơn, giúp giảm khả năng mắc kẹt vào các nghiệm cục bộ.
* **Bisecting K-Means:** Một biến thể của K-Means, chia dữ liệu thành hai cụm tại mỗi bước lặp, cho phép tìm kiếm các cấu trúc phân cấp trong dữ liệu.
* **Mini Batch K-Means:** Một phiên bản nhanh hơn của K-Means, sử dụng các mini-batch dữ liệu để cập nhật các tâm cụm.

**Ứng dụng của K-Means**

* **Phân nhóm khách hàng:** Chia khách hàng thành các nhóm có hành vi mua sắm tương tự.
* **Phân tích hình ảnh:** Phân cụm các pixel trong ảnh để nén hình ảnh hoặc phát hiện các đối tượng.
* **Phân tích văn bản:** Phân cụm các tài liệu thành các chủ đề.
* **Sinh học:** Phân loại các gen, phân tích dữ liệu microarray.

A screenshot of a computer screen

Description automatically generated

Trong hình trên, các điểm dữ liệu được phân thành 3 cụm khác nhau. Mỗi cụm được đại diện bởi một tâm cụm (điểm màu đen).

*Fuzzy C-means (FCM)*

**Fuzzy C-means (FCM)** là một thuật toán phân cụm (clustering) trong học máy, dựa trên nguyên lý "fuzzy" (mờ) thay vì phân cụm rắn (hard clustering) như trong các thuật toán phân cụm truyền thống (ví dụ: K-means). Trong K-means, mỗi điểm dữ liệu chỉ thuộc về một cụm duy nhất. Tuy nhiên, trong **Fuzzy C-means**, mỗi điểm dữ liệu có thể thuộc vào nhiều cụm với một mức độ "membership" (sự tham gia) khác nhau.

1. **Mục đích và Cách thức hoạt động**

Fuzzy C-means được dùng để phân chia một tập hợp dữ liệu thành **c** cụm sao cho các điểm dữ liệu trong mỗi cụm có sự tương đồng lớn hơn với các điểm dữ liệu trong các cụm khác. Tuy nhiên, không giống như K-means, FCM cho phép một điểm dữ liệu có thể "thuộc" vào nhiều cụm cùng một lúc, với mức độ khác nhau.

2. **Công thức và Các Bước trong Fuzzy C-means**

Fuzzy C-means tìm kiếm các cụm dựa trên mức độ thuộc về của từng điểm dữ liệu đối với mỗi cụm. Công thức của FCM dựa trên hàm **objective function** sau:

A black text with white text

Description automatically generated with medium confidence

Các bước trong thuật toán Fuzzy C-means:

*Khởi tạo:*

* Chọn số cụm C mà bạn muốn phân chia dữ liệu.
* Chọn một giá trị cho tham số m (thường lấy giá trị 2).
* Khởi tạo **membership matrix** U với các giá trị ngẫu nhiên, sao cho và tổng các giá trị theo mỗi dòng phải bằng 1.

*Tính toán trung tâm cụm:*

A white background with black text

Description automatically generated

*  Trung tâm cụm là giá trị trọng số của các điểm dữ liệu, với các trọng số là các mức độ "membership".

*Cập nhật membership matrix U:*

A math equations and numbers

Description automatically generated with medium confidence

*Kiểm tra điều kiện dừng:*

Thuật toán tiếp tục cập nhật trung tâm cụm và U cho đến khi một trong các điều kiện dừng sau đây được thỏa mãn:

* Sự thay đổi trong U hoặc các trung tâm cụm giữa các vòng lặp là rất nhỏ (chỉ số hội tụ).
* Đạt số vòng lặp tối đa.

**Ứng dụng của Fuzzy C-means:**

* Phân tích dữ liệu hình ảnh: Fuzzy C-means có thể được dùng để phân cụm các pixel trong hình ảnh, đặc biệt là khi các khu vực có sự chuyển giao mờ giữa các đối tượng.
* Phân cụm trong học máy: Dùng để phân cụm dữ liệu trong các bài toán như phân tích thị trường, nhận diện đối tượng trong dữ liệu đa chiều, v.v.
* Y tế: FCM có thể phân cụm dữ liệu hình ảnh y tế, ví dụ như phân tích ảnh chụp CT, MRI hoặc siêu âm để phân loại các mô hoặc tổ chức trong cơ thể.