## Giới thiệu tổng quát.

K-means clustering là một thuật toán phân cụm phổ biến trong học máy và khai thác dữ liệu, được sử dụng rộng rãi để phân nhóm dữ liệu thành các cụm riêng biệt dựa trên các đặc điểm tương đồng. Được giới thiệu lần đầu bởi Stuart Lloyd vào năm 1957, K-means nhanh chóng trở thành công cụ chủ đạo trong nhiều ứng dụng, từ phân tích hình ảnh, nhận diện mẫu đến phân đoạn thị trường và xử lý tín hiệu. Thuật toán này không chỉ dễ hiểu mà còn hiệu quả về mặt tính toán, đặc biệt phù hợp với những dữ liệu lớn và phức tạp trong các lĩnh vực khác nhau.

Nguyên tắc chính của K-means clustering là tìm cách chia dữ liệu thành *k* cụm riêng biệt sao cho khoảng cách giữa các điểm dữ liệu trong cùng một cụm là nhỏ nhất có thể, đồng thời khoảng cách giữa các cụm là lớn nhất. Cụ thể, thuật toán bắt đầu bằng cách chọn ngẫu nhiên kkk điểm làm tâm cụm, sau đó tiến hành hai bước chính: gán mỗi điểm dữ liệu vào cụm có tâm cụm gần nhất và cập nhật lại vị trí của tâm cụm dựa trên các điểm trong cụm đó. Quá trình này lặp lại cho đến khi các tâm cụm không thay đổi hoặc số lần lặp đạt đến một ngưỡng đã định, khi đó thuật toán đạt trạng thái hội tụ.

A diagram of a diagram of a number of dots

Description automatically generated

Hình 2. 1:Minh họa thuật toán phân cụm K-means

Một đặc điểm quan trọng của K-means là tính toán trọng tâm dựa trên khoảng cách Euclidean, vì vậy K-means thường thích hợp nhất khi các cụm có hình cầu và kích thước tương đồng. Đối với dữ liệu có hình dạng cụm phức tạp hoặc không cân đối, thuật toán này có thể không phân cụm tối ưu. Thêm vào đó, việc chọn số lượng cụm kkk là một bước quan trọng trong K-means. Nếu chọn kkk quá nhỏ, các cụm sẽ không phản ánh được hết các đặc điểm của dữ liệu, trong khi chọn kkk quá lớn có thể làm mất tính tổng quát của mô hình. Một trong những phương pháp phổ biến để xác định kkk tối ưu là phương pháp Elbow, dựa vào độ giảm của tổng bình phương sai số (SSE) khi tăng số cụm.

K-means được đánh giá cao nhờ tốc độ nhanh và khả năng triển khai dễ dàng. Vì vậy, nó thường được sử dụng trong các bài toán yêu cầu xử lý nhanh chóng trên lượng lớn dữ liệu, đặc biệt trong các bài toán thực tế như nén ảnh, phân đoạn ảnh, và nhận diện mẫu. Khi ứng dụng trong phân đoạn ảnh, K-means có khả năng phân chia các điểm ảnh thành các cụm tương ứng với các vùng màu sắc tương đồng, giúp làm nổi bật các vùng đặc trưng trong ảnh. Trong xử lý tín hiệu và dữ liệu tài chính, K-means có thể giúp phát hiện các nhóm khách hàng có hành vi tương tự, hoặc xác định các kiểu mẫu trong biến động giá cổ phiếu.

Mặc dù K-means clustering rất hiệu quả, thuật toán này cũng có một số hạn chế, chẳng hạn như nhạy cảm với các điểm dữ liệu ngoại lai (outliers) và phụ thuộc vào giá trị khởi tạo của các tâm cụm. Để cải thiện tính ổn định của thuật toán, nhiều biến thể của K-means đã được phát triển, chẳng hạn như K-means++ giúp chọn giá trị khởi tạo tốt hơn cho các tâm cụm, từ đó giảm thiểu nguy cơ hội tụ tại các cực trị cục bộ.

## Các thuật ngữ liên quan

Trong K-means clustering, có một số thuật ngữ quan trọng thường được sử dụng để mô tả quá trình và các yếu tố liên quan. Bảng dưới đây trình bày tóm tắt các thuật ngữ chính và ý nghĩa của nó trong thuật toán K-means.

Bảng 2. 1: Các thuật ngữ chính trong K-mean và ý nghĩa

|  |  |
| --- | --- |
| **Thuật ngữ** | **Mô tả** |
| Cluster (Cụm) | Nhóm các điểm dữ liệu có đặc điểm tương đồng với nhau. |
| Centroid (Tâm cụm) | Điểm trung tâm của một cụm, đại diện cho cụm, tính bằng giá trị trung bình của các điểm trong cụm. |
| Inertia | Tổng bình phương khoảng cách giữa mỗi điểm dữ liệu và tâm cụm gần nhất, dùng để đánh giá mức độ chặt chẽ của các cụm. |
| Distance Metric | Phương pháp đo khoảng cách giữa các điểm dữ liệu và tâm cụm; thường dùng khoảng cách Euclidean. |
| Initialization | Quá trình chọn vị trí ban đầu cho các tâm cụm trước khi lặp, có thể ngẫu nhiên hoặc dùng K-means++ để tối ưu hóa. |
| Convergence | Trạng thái khi thuật toán đạt hội tụ, các tâm cụm không thay đổi đáng kể nữa. |
| K (Số lượng cụm) | Số cụm mà dữ liệu sẽ được phân thành, xác định trước khi chạy thuật toán. |
| Assignment Step | Bước gán mỗi điểm dữ liệu vào cụm có tâm cụm gần nhất, thực hiện trong mỗi vòng lặp. |
| Update Step | Bước cập nhật lại vị trí tâm cụm bằng cách tính trung bình các điểm trong cụm, thực hiện trong mỗi vòng lặp. |
| Elbow Method | Phương pháp xác định số cụm kkk tối ưu bằng cách chọn điểm "khuỷu tay" trên đồ thị inertia. |
| Silhouette Score | Thước đo hiệu quả phân cụm, giá trị từ -1 đến 1, với giá trị cao cho thấy các cụm được phân tách rõ ràng. |
| Iterations | Số vòng lặp của thuật toán từ khi bắt đầu đến khi hội tụ. |
| Local Minima | Cực tiểu cục bộ của hàm mục tiêu, nơi thuật toán có thể hội tụ nhưng không đạt tối ưu toàn cục. |
| Outliers | Các điểm dữ liệu nằm xa các cụm chính, có thể ảnh hưởng đến độ chính xác của kết quả phân cụm. |
| Dimensionality | Số lượng đặc trưng trong tập dữ liệu, ảnh hưởng đến tính phức tạp của K-means; dữ liệu nhiều chiều có thể cần thêm xử lý hoặc giảm chiều trước khi phân cụm. |

## Cách K-Means hoạt động.

K-means clustering hoạt động dựa trên một quy trình lặp để phân cụm dữ liệu thành các nhóm khác nhau, tối ưu hóa sao cho các điểm dữ liệu trong mỗi cụm có mức độ tương đồng cao nhất. Mỗi cụm được đặc trưng bởi một “tâm cụm” (centroid), được tính toán sao cho khoảng cách giữa tâm cụm và các điểm dữ liệu trong cùng cụm là nhỏ nhất. Thuật toán có thể được chia thành các bước chính như sau:

1. **Khởi tạo tâm cụm ban đầu**: Thuật toán bắt đầu bằng cách xác định trước số cụm kkk và chọn ngẫu nhiên kkk điểm dữ liệu làm tâm cụm ban đầu. Tâm cụm có thể được chọn bằng phương pháp ngẫu nhiên hoặc với các phương pháp tối ưu hơn như K-means++ để đạt được sự phân cụm hiệu quả hơn.
2. **Gán điểm dữ liệu vào các cụm**: Sau khi có các tâm cụm ban đầu, mỗi điểm dữ liệu sẽ được gán vào cụm có tâm cụm gần nhất. Khoảng cách từ mỗi điểm đến tâm cụm được tính toán thường bằng khoảng cách Euclidean, tuy nhiên các khoảng cách khác cũng có thể được sử dụng tùy thuộc vào bài toán.
3. **Cập nhật tâm cụm**: Sau khi các điểm đã được gán vào cụm, thuật toán tính toán lại vị trí của mỗi tâm cụm bằng cách lấy trung bình của tất cả các điểm trong cụm đó. Việc này đảm bảo rằng mỗi tâm cụm được điều chỉnh để nằm chính giữa các điểm trong cụm.
4. **Lặp lại cho đến khi hội tụ**: Bước gán điểm dữ liệu vào cụm và cập nhật tâm cụm sẽ tiếp tục lặp lại cho đến khi các tâm cụm không thay đổi nhiều (hoặc không thay đổi hoàn toàn), tức là đạt trạng thái hội tụ. Đôi khi, quá trình này được dừng lại nếu số vòng lặp đạt đến một ngưỡng nhất định để tránh việc lặp vô hạn.

Giả sử chúng ta có dữ liệu của một chuỗi cửa hàng cà phê muốn phân chia khách hàng thành hai nhóm dựa trên số lần đến cửa hàng (biểu thị thói quen trung thành) và mức chi tiêu trung bình mỗi lần đến. Mỗi khách hàng sẽ là một điểm dữ liệu với hai thuộc tính: “số lần đến” và “mức chi tiêu trung bình”. Các bước thực hiện K-means với ví dụ này như sau:

1. Chọn số lượng cụm k=2: Do chuỗi cà phê muốn phân chia thành hai nhóm khách hàng, chúng ta chọn số lượng cụm k=.
2. Khởi tạo tâm cụm ban đầu: Chọn ngẫu nhiên hai điểm khách hàng từ dữ liệu, hoặc tạo hai điểm đại diện, làm tâm cụm ban đầu.
3. Gán khách hàng vào các cụm: Với mỗi khách hàng (điểm dữ liệu), tính khoảng cách đến hai tâm cụm. Nếu khách hàng có khoảng cách gần tâm cụm 1 hơn tâm cụm 2, thì khách hàng đó sẽ được gán vào cụm 1, và ngược lại.
4. Cập nhật tâm cụm: Sau khi các khách hàng đã được gán vào các cụm, tính lại tâm cụm bằng cách lấy giá trị trung bình của số lần đến và mức chi tiêu trung bình của tất cả các khách hàng trong mỗi cụm. Tâm cụm mới sẽ phản ánh chính xác vị trí trung tâm của các khách hàng trong cụm đó.
5. Lặp lại quá trình: Các bước gán và cập nhật tâm cụm tiếp tục lặp lại cho đến khi các tâm cụm không thay đổi hoặc thay đổi rất ít. Khi đó, thuật toán đã hội tụ và chúng ta có được hai cụm khách hàng với đặc điểm khác biệt rõ ràng: một nhóm có tần suất đến cửa hàng cao và chi tiêu nhiều, và nhóm còn lại đến ít hơn hoặc chi tiêu ít hơn.

Kết quả cuối cùng sẽ là hai cụm khách hàng, mỗi cụm có một đặc trưng riêng. Ví dụ, cụm đầu tiên có thể đại diện cho nhóm khách hàng trung thành, ghé thăm thường xuyên và chi tiêu nhiều, trong khi cụm thứ hai có thể đại diện cho những khách hàng đến ít hơn và chi tiêu ít hơn. Thông tin này giúp chuỗi cà phê tối ưu hóa chiến lược marketing, thiết kế chương trình khuyến mãi cho nhóm khách hàng trung thành, đồng thời tìm cách thu hút và khuyến khích nhóm khách hàng ít đến ghé thăm thường xuyên hơn. Phương pháp K-means, qua ví dụ này, cho thấy khả năng phân nhóm dữ liệu rõ ràng và hiệu quả, đồng thời cung cấp cái nhìn trực quan về đặc điểm của các nhóm dữ liệu trong các ứng dụng thực tế.

## 2.4. Các ứng dụng của Kmeans

Kmeans là thuật toán phổ biến và nổi bật trong việc phân cụm, vì thế nó có nhiều ứng dụng trong nhiều lĩnh vực khác nhau, có thể kể đến như:

* Phân cụm khách hàng:Phân nhóm khách hàng theo hành vi mua sắm hoặc sở thích để tối ưu hóa chiến dịch marketing và cá nhân hóa đề xuất sản phẩm.
* Nén ảnh: Giảm số lượng màu sắc trong hình ảnh bằng cách nhóm màu tương tự lại với nhau, giúp giảm kích thước ảnh mà vẫn giữ được chất lượng.
* Phân tích thị trường chứng khoán: Phân nhóm cổ phiếu dựa trên các đặc điểm tài chính để đánh giá rủi ro và xây dựng danh mục đầu tư.
* Phân loại văn bản: Phân nhóm các văn bản theo chủ đề, hỗ trợ trong xử lý ngôn ngữ tự nhiên như phân loại email, tài liệu hoặc bài báo.
* Phân cụm dữ liệu địa lý: Sử dụng trong GIS để phân nhóm các điểm dữ liệu địa lý, hỗ trợ quản lý tài nguyên hoặc lập kế hoạch đô thị.
* Phân tích gen trong sinh học: Phân nhóm các gen tương tự giúp nghiên cứu sự phát triển và mối liên hệ giữa các gen.
* Phân tích hành vi người dùng: Phân loại hành vi của người dùng trên trang web, giúp cải thiện trải nghiệm người dùng và tối ưu hóa quảng cáo.
* Chẩn đoán lỗi trong công nghiệp: Sử dụng trong IoT để phân tích dữ liệu từ các cảm biến, phát hiện các sự cố và bảo trì phòng ngừa trong hệ thống công nghiệp.

K-means được áp dụng rộng rãi nhờ tính linh hoạt và hiệu quả trong việc phân cụm dữ liệu lớn, từ thương mại đến công nghệ sinh học và kỹ thuật công nghiệp.