KMEANS- CLUSTERING

I. CƠ SỞ LÍ THUYẾT

1. KHÁI NIỆM

- **Thuật toán K-Means** là một thuật toán phân cụm (clustering) trong học máy không giám sát (unsupervised learning). Mục tiêu của K-Means là chia tập dữ liệu thành K nhóm (cụm) sao cho các điểm dữ liệu trong cùng một cụm có độ tương đồng cao, còn các cụm khác nhau thì khác biệt rõ rệt.

2. TỔNG QUAN VỀ VỊ TRÍ TRONG THUẬT TOÁN

1. Khái niệm về Phân cụm (Clustering)

- Phân cụm là một **kỹ thuật trong học máy không giám sát (unsupervised learning)**, nhằm **chia tập dữ liệu thành các nhóm (cụm)** sao cho:

- **Các điểm trong cùng cụm** có đặc điểm **giống nhau hoặc gần nhau** nhất có thể.
- **Các điểm ở cụm khác nhau** thì **khác biệt rõ ràng**.

2. Mục tiêu của Phân cụm

- **Khám phá mẫu (pattern)** trong dữ liệu.
- **Tóm tắt, giảm kích thước dữ liệu** để dễ phân tích.
- **Hỗ trợ ra quyết định** trong marketing, tài chính, hoặc nhận dạng hành vi.

3. Vị trí của phân cụm trong học máy

- Trong học máy (Machine Learning), phân cụm thuộc **nhánh học không giám sát (unsupervised learning)**:

Loại học máy	Đặc điểm chính	Ví dụ thuật toán
Học có giám sát (Supervised)	Dữ liệu có nhãn sẵn (label)	Hồi quy tuyến tính, KNN, Cây quyết định
Học không giám sát (Unsupervised)	Dữ liệu không có nhãn, cần tự tìm mẫu	**Phân cụm (Clustering)**, Giảm chiều PCA
Học tăng cường (Reinforcement)	Học thông qua phần thưởng/phạt	Q-learning, DQN

*→ **K-Means** là **thuật toán tiêu biểu nhất** trong nhóm **phân cụm** của **học không giám sát**.*

4. Các phương pháp phân cụm phổ biến

Loại phương pháp	Mô tả	Ví dụ thuật toán
Phân cụm phân hoạch (Partitioning)	Chia dữ liệu thành K cụm rõ ràng	**K-Means**, K-Medoids
Phân cụm phân cấp (Hierarchical)	Tạo cây phân cấp các cụm	Agglomerative, Divisive
Phân cụm dựa trên mật độ (Density-based)	Xác định cụm theo vùng có mật độ cao	DBSCAN, OPTICS
Phân cụm dựa trên mô hình (Model-based)	Giả định dữ liệu theo mô hình xác suất	Gaussian Mixture Model (GMM)

5. Vị trí của thuật toán K-Means

- Nằm trong **nhóm phân cụm phân hoạch (Partitioning Clustering)**.
- Là **thuật toán cơ bản, đơn giản và phổ biến nhất** trong phân cụm.
- Được xem là **bước khởi đầu** cho nhiều phương pháp phân cụm nâng cao khác.

6. Vai trò của K-Means trong phân cụm

- Là **chuẩn so sánh (baseline)** cho các thuật toán phân cụm khác.
- Dễ triển khai, nhanh, phù hợp cho dữ liệu lớn.
- Là nền tảng để hiểu các thuật toán phức tạp hơn như **K-Medoids**, **GMM**, **DBSCAN**.
- Phân cụm là **kỹ thuật cốt lõi trong học không giám sát**, giúp **khám phá cấu trúc ẩn trong dữ liệu**. Trong đó, **K-Means** giữ **vị trí trung tâm và nền tảng nhất**, được ứng dụng rộng rãi nhờ **đơn giản, hiệu quả và trực quan**.

3. NGUYÊN LÍ HOẠT ĐỘNG

1. Mục tiêu chính

- Thuật toán K-Means hoạt động dựa trên nguyên lý phân nhóm dữ liệu thành K cụm sao cho:

+ Các điểm trong cùng một cụm thì giống nhau nhất có thể.

+ Các điểm ở cụm khác nhau thì khác nhau nhiều nhất có thể.

Nói cách khác, K-Means tối thiểu hóa tổng bình phương khoảng cách từ mỗi điểm dữ liệu đến tâm cụm (centroid) gần nhất.

2. Ý tưởng cốt lõi

K-Means tìm K điểm trung tâm (centroid) sao cho

$$J = \sum_{i=1}^{K} \sum_{x \in C_i} ||x - \mu_i||^2$$

được nhỏ nhất có thể, trong đó:

- C_i: cụm thứ i
- μ_i: tâm cụm thứ i
- $|| x - \mu_i ||^2$: bình phương khoảng cách từ điểm dữ liệu xđến tâm cụm μ_i

3. Các bước hoạt động của K-Means

Bước 1: Chọn số cụm K

- Người dùng hoặc chuyên gia chọn trước số cụm cần chia (ví dụ: 3 cụm, 4 cụm,…).

Bước 2: Khởi tạo tâm cụm ban đầu

- Chọn ngẫu nhiên Kđiểm dữ liệu làm tâm cụm ban đầu (centroid).

Bước 3: Gán cụm cho từng điểm dữ liệu

- Tính khoảng cách (thường dùng Euclidean) từ mỗi điểm đến tất cả các tâm cụm,
→ Gán điểm đó vào cụm có tâm gần nhất.

Bước 4: Cập nhật lại tâm cụm

- Sau khi gán xong, tính lại tâm cụm mới bằng trung bình cộng của tất cả các điểm trong cụm đó:

$$\mu_i = \frac{1}{| C_i |} \sum_{x \in C_i} x$$

Bước 5: Lặp lại

- Lặp lại Bước 3 và 4 cho đến khi:
  - Tâm cụm không thay đổi đáng kể, hoặc
  - Số vòng lặp đạt giới hạn cho phép.

Khi đó, mô hình hội tụ (đã ổn định).

4. Minh họa trực quan (mô tả bằng lời)

Ví dụ:
Giả sử ta có 2 cụm (K=2):

1. Ban đầu chọn ngẫu nhiên 2 tâm cụm.
2. Mỗi điểm dữ liệu “chạy” về phía tâm gần nó nhất.

3. Tính lại trung bình của từng cụm để cập nhật tâm mới.
4. Lặp đi lặp lại → các điểm "tụ" lại quanh 2 tâm ổn định → đó là kết quả phân cụm cuối cùng.

5. Đặc điểm chính

- Khoảng cách dùng phổ biến: Euclidean (khoảng cách Euclid).
- Dừng khi hội tụ: Không còn thay đổi lớn ở tâm cụm.
- Tính chất: Dữ liệu dạng số, cụm hình cầu cho kết quả tốt nhất.

4. QUY TRÌNH THUẬT TOÁN

1. Quy trình của thuật toán K-Means

- Thuật toán **K-Means** có quy trình gồm **5 bước chính**, được lặp lại cho đến khi kết quả ổn định (hội tụ). Dưới đây là mô tả chi tiết từng bước.

Bước 1: Xác định số cụm K

- Chọn **số cụm (K)** cần phân chia trong tập dữ liệu.
- Số K thường được **chọn trước** dựa trên kinh nghiệm hoặc các phương pháp đánh giá như **Elbow Method** hoặc **Silhouette Score**.

Ví dụ: Muốn chia khách hàng thành 3 nhóm → chọn K = 3.

Bước 2: Khởi tạo tâm cụm ban đầu (Centroid)

- Chọn ngẫu nhiên **K điểm dữ liệu** trong tập dữ liệu làm **tâm cụm ban đầu**.
- Các tâm này đại diện cho vị trí trung bình của mỗi cụm.

Ví dụ: Nếu K = 3 → chọn 3 điểm dữ liệu làm tâm cụm tạm thời.

Bước 3: Gán điểm dữ liệu vào cụm gần nhất

- Với mỗi điểm dữ liệu, tính **khoảng cách** đến các tâm cụm (thường là khoảng cách Euclidean).
- Gán điểm đó vào cụm có **tâm gần nhất**.

Công thức khoảng cách Euclidean:

$$d(x, \mu_i) = \sqrt{\sum_{j=1}^{n} (x_j - \mu_{ij})^2}$$

Kết quả của bước này: mỗi điểm dữ liệu được gắn với một cụm.

Bước 4: Cập nhật lại tâm cụm

- Sau khi gán xong, tính lại **tâm cụm mới** bằng trung bình của tất cả các điểm trong cùng cụm:

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$$

- Các tâm cụm sẽ **di chuyển** dần đến vị trí tối ưu, nơi dữ liệu trong cụm đồng nhất nhất.

Bước 5: Kiểm tra điều kiện dừng

- So sánh tâm cụm mới với tâm cụm cũ:
  - Nếu **không thay đổi đáng kể**, thuật toán **hội tụ → dừng lại**.
  - Nếu **vẫn thay đổi**, quay lại **Bước 3** để tiếp tục lặp.

2. Sơ đồ tóm tắt quy trình K-Means

Bắt đầu

↓

Chọn số cụm K

↓

Khởi tạo K tâm cụm ban đầu

↓

Gán từng điểm dữ liệu vào cụm gần nhất

↓

Tính lại tâm cụm (trung bình các điểm trong cụm)

↓

Kiểm tra hội tụ

↓

┌──────────────────────────┐

│ Còn thay đổi? │—— Có —> Quay lại bước gán cụm

└──────────────────────────┘

↓

Không

↓

Kết thúc → Xuất kết quả phân cụm

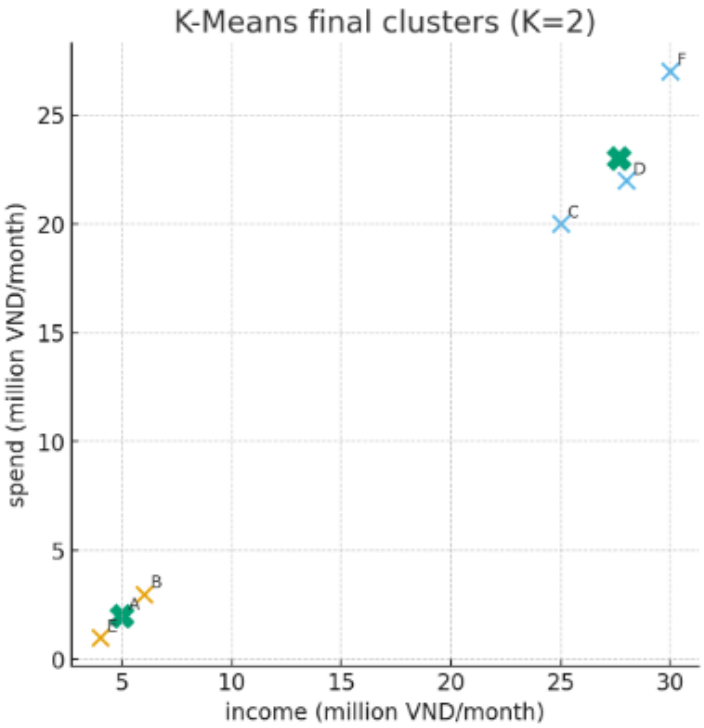
3. Kết quả đầu ra

Sau khi thuật toán dừng lại, ta thu được:

- **K tâm cụm cuối cùng (centroids)**.
- **Nhãn cụm của mỗi điểm dữ liệu** (điểm đó thuộc cụm nào).

5. VÍ DỤ MINH HỌA VÀ THỰC HÀNH

Khách hàng	Thu nhập (x)	Chi tiêu (y)
A	5	2
B	6	3
C	25	20
D	28	22
E	4	1
F	30	27



Tóm tắt nhanh (kết quả cuối cùng):

- Bộ điểm: A(5,2), B(6,3), C(25,20), D(28,22), E(4,1), F(30,27).
- Khởi tạo tâm: C1 = A(5,2), C2 = D(28,22).
- Kết quả phân cụm cuối cùng (cluster 1/2):

Cụm 1: A, B, E (thu nhập thấp, chi tiêu thấp)

- Cụm 2: C, D, F (thu nhập cao, chi tiêu cao)
- Tâm cụm cuối cùng (x, y):
- Centroid 1 ≈ (5.0, 2.0) — gần nhóm A,B,E trung bình.
- Centroid 2 ≈ (27.67, 23.0) — gần nhóm C,D,F trung bình.

6. KẾT LUẬN VÀ TÀI LIỆU THAM KHẢO

1. Kết Luận:

- Thuật toán **K-Means** là một trong những phương pháp **phân cụm dữ liệu (clustering)** phổ biến và hiệu quả nhất trong **học máy không giám sát**.
Nó giúp **chia dữ liệu thành các nhóm (cụm)** dựa trên **độ tương đồng về đặc trưng**, qua đó hỗ trợ việc **phân tích, phát hiện mẫu, hoặc dự đoán xu hướng** trong nhiều lĩnh vực.

- Qua ví dụ thực tế với dữ liệu "thu nhập – chi tiêu", ta thấy K-Means có thể:

- Nhóm **các đối tượng có hành vi tương tự nhau** lại với nhau (ví dụ nhóm khách hàng).
    - Giúp **hiểu rõ hơn về cấu trúc ẩn** trong dữ liệu mà không cần nhãn có sẵn.
    - Tuy nhiên, kết quả phụ thuộc vào **số cụm K lựa chọn**, **vị trí khởi tạo ban đầu**, và chỉ phù hợp khi **các cụm có hình dạng tương đối tròn, cách biệt rõ**.

- Nhìn chung, **K-Means** là một công cụ mạnh mẽ, dễ triển khai và có giá trị ứng dụng cao trong thực tế như:

- Phân khúc khách hàng (Customer Segmentation)
- Phân loại vùng địa lý hoặc hành vi tiêu dùng
- Nhận dạng mẫu trong dữ liệu lớn

2. Tài liệu tham khảo

1. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning* (2nd ed.). Springer.
2. Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
3. scikit-learn.org. "Clustering — K-Means" [Online]. Available: https://scikit-learn.org/stable/modules/clustering.html#k-means
4. Wikipedia. "K-means clustering." [Online]. Available: https://en.wikipedia.org/wiki/K-means_clustering