K-means

K-means là một thuật toán **phân cụm không giám sát (unsupervised clustering)**. Mục tiêu là **chia tập dữ liệu thành K cụm**, sao cho các điểm trong cùng một cụm thì **gần nhau** (theo khoảng cách Euclidean hoặc tương tự).

Ý tưởng chính

Thuật toán lặp lại **2 bước chính** cho đến khi hội tụ:

1. **Gán điểm dữ liệu vào cụm gần nhất (Assignment step)**

   Mỗi điểm dữ liệu được gán vào cụm có **tâm cụm (centroid)** gần nhất.

   Với mỗi điểm $x^{(i)}$, Gán chỉ số cụm $c^{(i)}$ là cụm có tâm gần nhất:

$$c^{(i)} := \arg\min_{k \in \{1,\dots,K\}} ||x^{(i)} - \mu_k||^2$$

1. **Cập nhật tâm cụm (Update step)**

   Tính **trung bình tất cả các điểm** trong mỗi cụm và **di chuyển centroid đến vị trí trung bình đó**.

   Với mỗi cụm k:

   - Tính tâm mới μ_k là trung bình các điểm thuộc cụm đó:

$$\mu_k := \frac{1}{|C_k|} \sum_{i \in C_k} x^{(i)}$$

     Trong đó C_k là tập hợp các điểm được gán vào cụm k.

Khi hội tụ

- Các cụm **ổn định**: không thay đổi gán cụm và vị trí centroid.

Hàm chi phí (cost function)

Dù K-means không dùng gradient descent, nhưng nó **vẫn tối ưu một hàm chi phí cụ thể**, gọi là **hàm biến dạng (distortion function)**.

Ký hiệu:

- $x^{(i)}$: điểm dữ liệu thứ i
- $c^{(i)}$: chỉ số cụm mà $x^{(i)}$ được gán vào (1 đến K)
- μ_k: tâm cụm thứ k
- $\mu_{c^{(i)}}$: tâm cụm mà $x^{(i)}$ được gán vào

Hàm chi phí (distortion cost function) J:

$$J(c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K) = \frac{1}{m} \sum_{i=1}^{m} \|x^{(i)} - \mu_{c^{(i)}}\|^2$$

⟹ Trung bình bình phương khoảng cách từ mỗi điểm đến tâm cụm mà nó được gán.

- Khi J không còn giảm nữa ⇒ **thuật toán hội tụ**.
- Dừng lặp khi:
  - Giá trị J không thay đổi giữa hai vòng lặp.
  - Hoặc thay đổi quá nhỏ (giảm rất chậm).

Khởi tạo Centroids

Cách phổ biến nhất là:

- **Chọn ngẫu nhiên K điểm từ tập huấn luyện** làm vị trí ban đầu cho các tâm cụm
- Điều kiện: $K < m$ (số cụm nhỏ hơn số điểm dữ liệu).

Cách chọn cụm tốt nhất:

- Chạy K-means **nhiều lần với khởi tạo ngẫu nhiên khác nhau**
- Với mỗi lần chạy:
  1. Khởi tạo K điểm cụm từ dữ liệu
  2. Chạy K-means đến hội tụ
  3. Tính **hàm chi phí J (distortion function)** cho kết quả đó
- Sau tất cả các lần chạy, **chọn kết quả có J nhỏ nhất**

Tại sao khởi tạo ngẫu nhiên lại quan trọng?

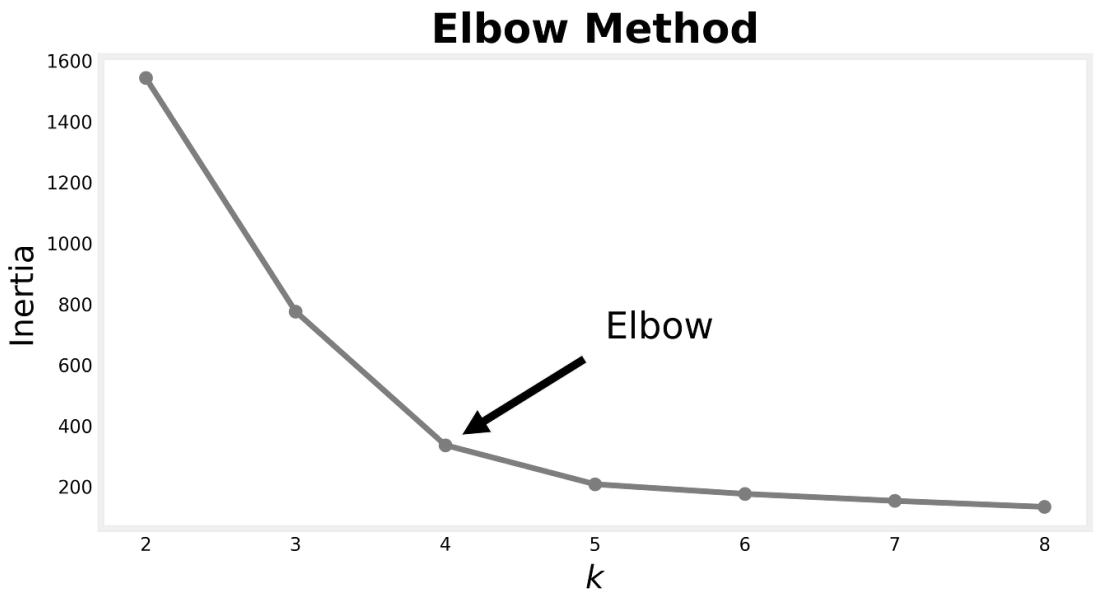
- Kết quả của K-means **phụ thuộc mạnh** vào cách khởi tạo ban đầu.
- Một khởi tạo tốt có thể dẫn đến **các cụm rõ ràng và trực quan**.
- Một khởi tạo kém có thể khiến K-means **mắc kẹt tại cực tiểu cục bộ**, dẫn đến kết quả tệ hơn.

Chọn số cụm

Số cụm "đúng" thường **mơ hồ** và phụ thuộc vào:

- Mục tiêu bài toán
- Cách nhìn nhận dữ liệu
- Ứng dụng thực tế

Elbow Method (Phương pháp Khuỷu tay) – phương pháp phổ biến



- **Ý tưởng:** Vẽ đồ thị hàm chi phí J (distortion) theo số cụm K
- Khi K tăng ⇒ J giảm
- Nếu đồ thị có "khuỷu" (chỗ gấp khúc rõ rệt), đó có thể là điểm dừng hợp lý

Cách thực tế & hiệu quả nhất: **Dựa vào mục tiêu sử dụng**

- Không chọn K chỉ vì tối ưu hoá hàm chi phí
- Thay vào đó, chọn K sao cho **cụm tạo ra phục vụ tốt mục đích sau đó**