Chương 3:
K-means (phần 2)

TS Phạm Tuấn

no_reply@example.com

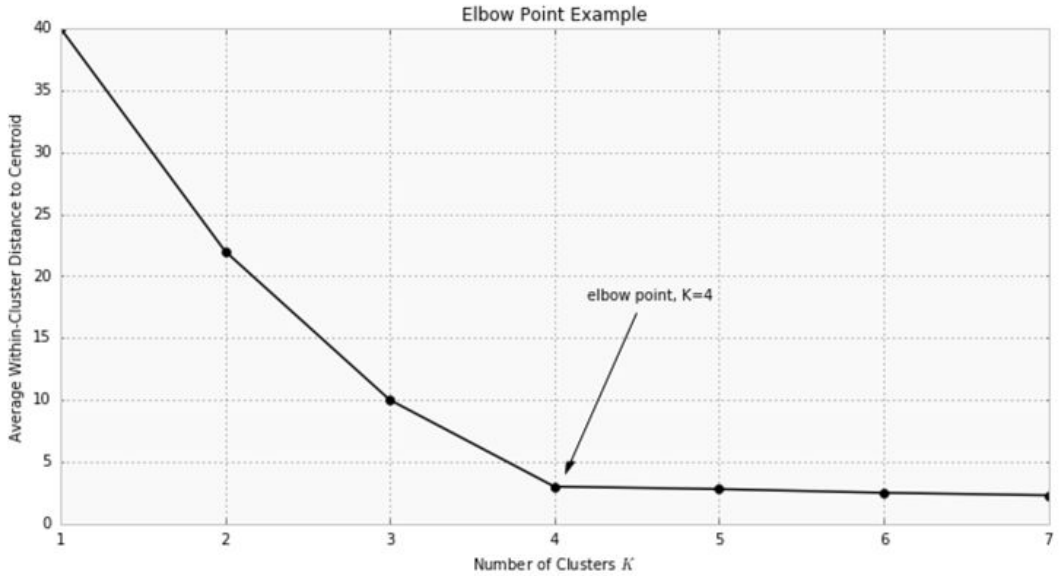
Xác định số K

Chúng ta sẽ dùng phương pháp Elbow để xác định giá trị K của một bài toán. Elbow sẽ dùng một thông số để xác định giá trị K đó là Within Cluster Sum of Squares (WCSS), nó sẽ tính tổng khoảng cách tất cả điểm trong từng cụm, rồi sau đó tính tổng khoảng cách ở tất cả các cụm. Công thức WCSS được tính như sau

- **Within Cluster Sums of Squares :** $$\mathrm{WSS} = \sum_{i=1}^{N_C} \sum_{x \in C_i} d(\mathbf{x}, \bar{\mathbf{x}}_{C_i})^2$$

Xác định số K

Chúng ta sẽ cho K chạy từ 1 đến 1 số hữu hạn và chúng ta sẽ có hình như sau:



Bài tập số 4

1, Download dữ liệu tại đây và in ra head. (1 điểm)

2, Xây dựng hàm tính khoảng cách từ tất cả các điểm đến centroid. (1 điểm)

3. Dùng K-means trong thư viện Sklearn để xác định centroid. Với K từ 1 đến 10, tương ứng với mỗi K thì tính WCSS bằng dùng hàm ở trên. (3 điểm)

4. Vẽ đường Elbow để xác định 2 số K tối ưu. (1 điểm) (K=4 và 6)

5, Đưa ra nhận xét về từng cụm trong mỗi số K bằng cách vẽ lên trục toạ độ tất cả các điểm và Centroid. Dựa vào kết quả nhận xét số K nào tối ưu hơn. (2 điểm)

6. Tìm các điểm kỳ dị trong từng cụm với số K tối ưu. (2 điểm)