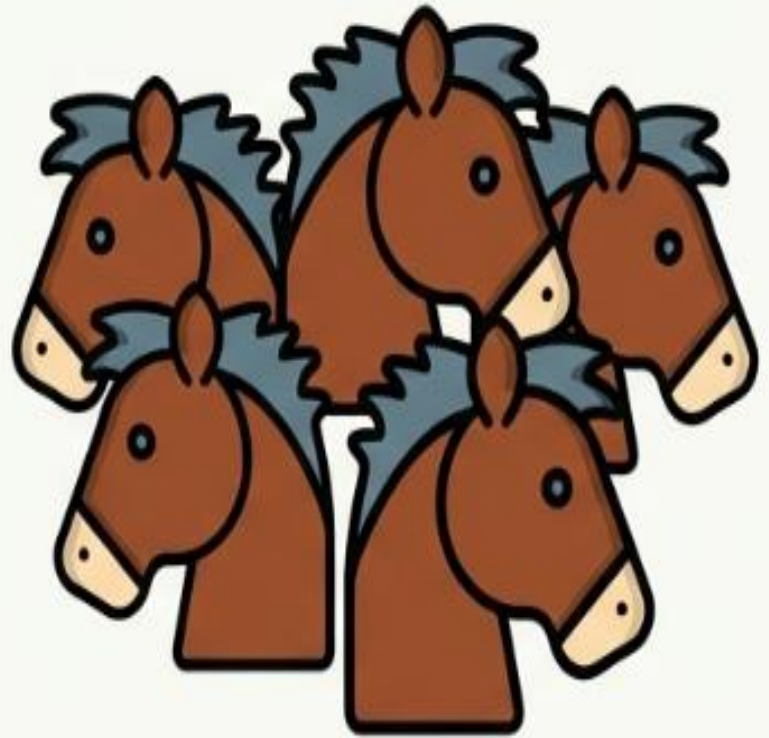


k-Nearest Neighbors

Giới thiệu



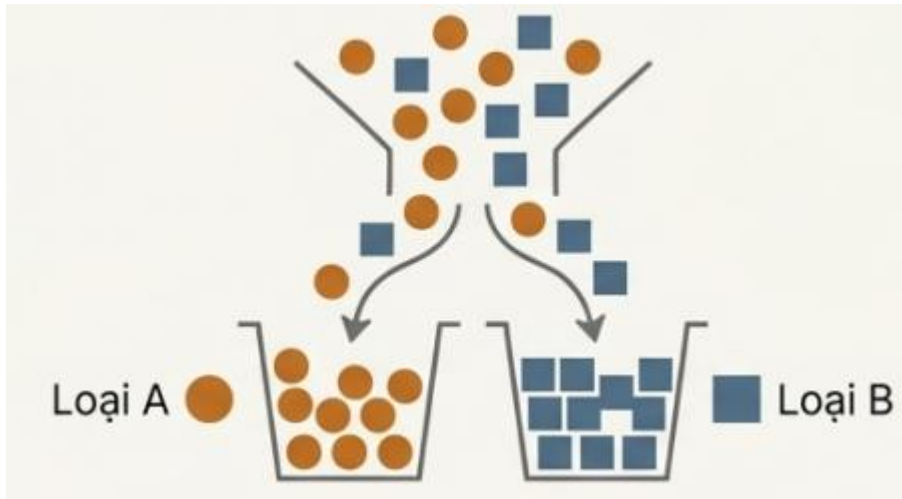
Ý tưởng cốt lõi của KNN

- Để hiểu bản chất của một điểm dữ liệu mới, hãy nhìn vào những “hàng xóm” gần nhất của điểm dữ liệu đó.

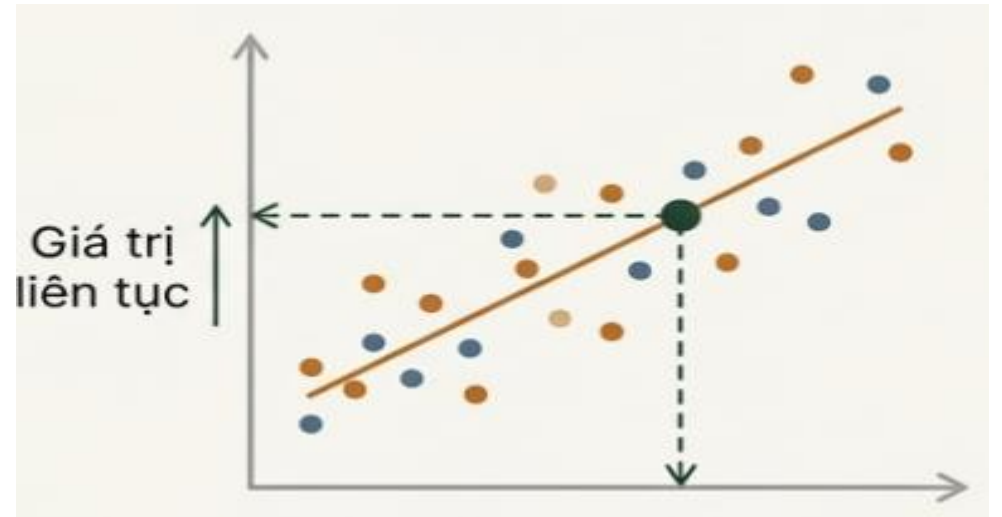
Tổng quan về KNN

- Học có giám sát: Yêu cầu dữ liệu đã được gán nhãn. Mục tiêu là học mối liên hệ giữa đầu vào và đầu ra đã biết
- Học lười: Không có giai đoạn huấn luyện thực sự. Chỉ đơn giản là ghi nhớ toàn bộ dữ liệu và trì hoãn tính toán cho đến khi có yêu cầu dự đoán

Hai nhiệm vụ chính của KNN



- Mục tiêu: Dự đoán một nhãn rời rạc
- Ví dụ: Phân loại email Spam hay Not Spam



- Mục tiêu: Dự đoán một giá trị liên tục
- Ví dụ: Dự đoán giá nhà, nhiệt độ ngày mai, doanh thu tháng tới

Quy trình xây dựng mô hình KNN

Thuật Toán K-Láng Giềng Gần Nhất (KNN): Toàn Tập Trong 4 Bước

Thuật toán K-Láng Giềng Gần Nhất (KNN) dự đoán đặc tính của một điểm dữ liệu mới bằng cách tham khảo ý kiến của 'k' điểm dữ liệu gần nó nhất trong tập huấn luyện.

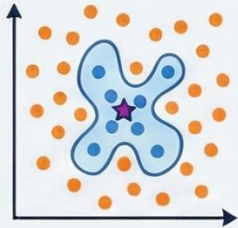
BƯỚC 1: Lựa chọn siêu tham số 'k'



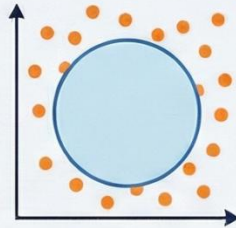
k (Số lượng hàng xóm)

'k' là số lượng "hàng xóm" được dùng để bỏ phiếu quyết định cho điểm dữ liệu mới.

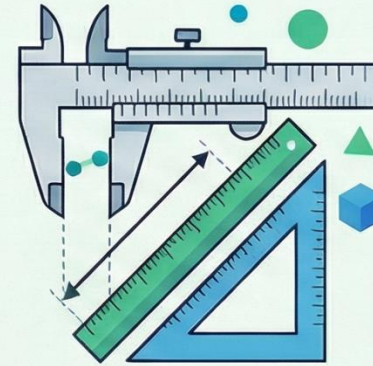
Thách thức khi chọn 'k'



'k' quá nhỏ
Dẫn đến quá khớp
(Overfitting)



'k' quá lớn
Dẫn đến dưới khớp
(Underfitting)

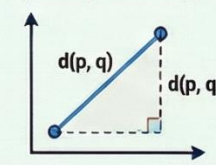


BƯỚC 2: Tính toán khoảng cách

Phải chuẩn hóa dữ liệu trước để các đặc trưng có thang đo tương đương và đóng góp đồng đều.

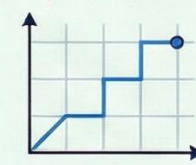
Các thước đo khoảng cách phổ biến

Euclidean
(Đường chim bay)



$$d(p, q) = \sqrt{\sum (p_i - q_i)^2}$$

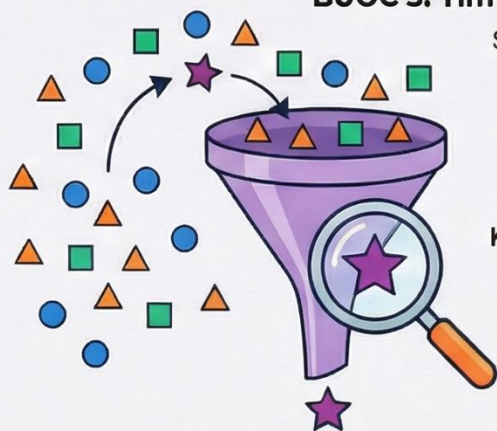
Manhattan
(Đường đi trong thành phố)



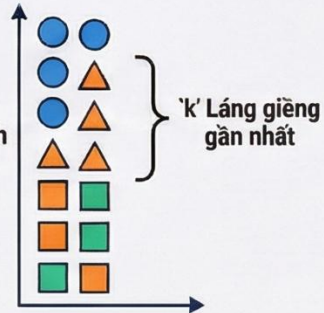
$$d(p, q) = \sum |p_i - q_i|$$

BƯỚC 3: Tìm 'k' láng giềng gần nhất

Sắp xếp tất cả các điểm dữ liệu theo khoảng cách tăng dần và chọn ra 'k' điểm đầu tiên.



Khoảng cách tăng dần



'k' Láng giềng gần nhất

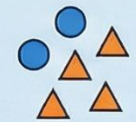
BƯỚC 4: Đưa ra dự đoán cuối cùng

Cách đưa ra quyết định phụ thuộc vào loại bài toán: phân loại hay hồi quy.

Phân loại



Bỏ phiếu theo đa số (Majority Vote):
Gán nhãn phổ biến nhất trong 'k' hàng xóm.



Dự đoán:
Nhãn Phổ Biến

Hồi quy



Lấy giá trị trung bình (Averaging):
Tính giá trị trung bình của 'k' hàng xóm.



$$\text{Mean} = \frac{10+12+15+11+13}{5} = 12.2$$

Dự đoán:
Giá Trị Trung Bình

Quy trình xây dựng mô hình KNN

- Bước 1: Lựa chọn k
 - Lựa chọn số lượng “hàng xóm” được dùng để tham khảo ý kiến
 - k có ảnh hưởng lớn đến độ phức tạp và khả năng tổng quát hóa của mô hình
 - => Làm thế nào để chọn k ?????

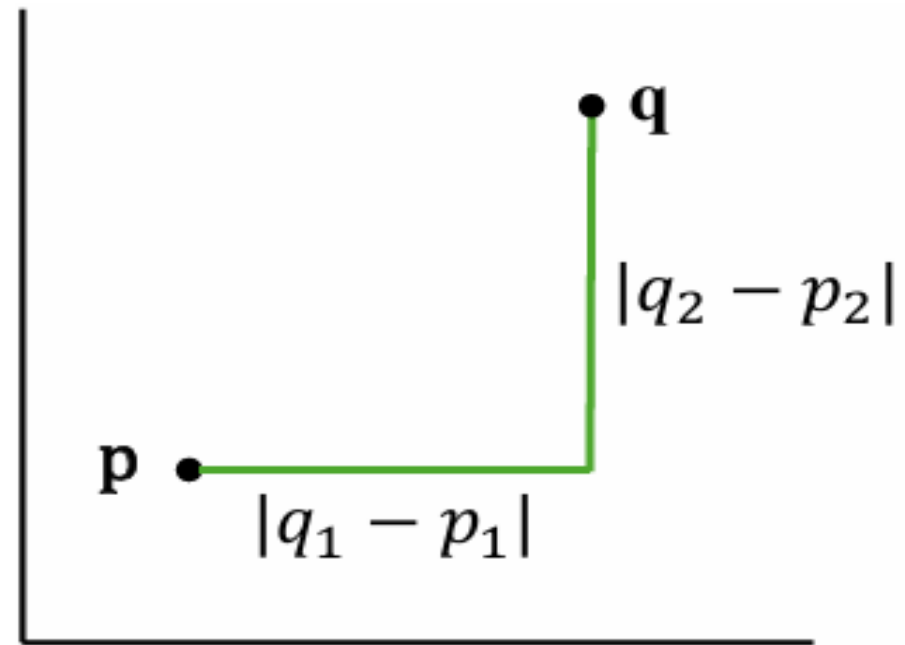
Quy trình xây dựng mô hình KNN

- Bước 1: Lựa chọn k
 - Không có một công thức toán học nào cho giá trị k tốt nhất.
 - Dựa trên kinh nghiệm:
 - Cross-Validation
 - Rule of Thumb
 - Chọn k là số lẻ cho bài toán phân loại

Quy trình xây dựng mô hình KNN

- Bước 2: Tính khoảng cách
 - Các cách đo khoảng cách phổ biến:
 - Manhattan:

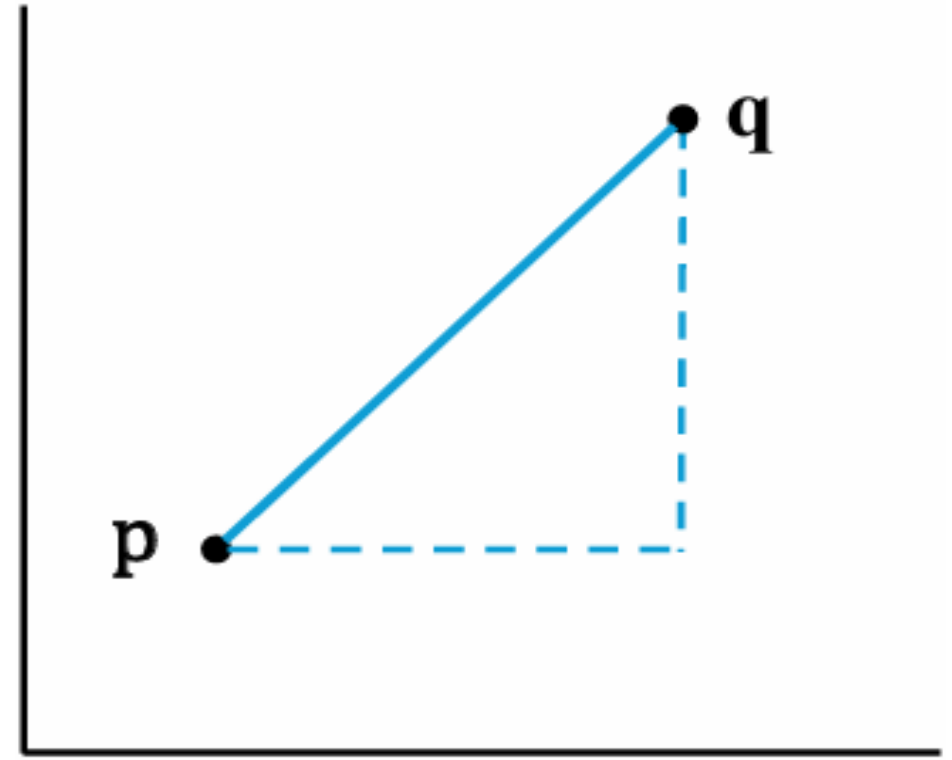
$$d(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^n |q_i - p_i|$$



Quy trình xây dựng mô hình KNN

- Bước 2: Tính khoảng cách
 - Các cách đo khoảng cách phổ biến:
 - Eclid:

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^n (q_i - p_i)^2}$$



Quy trình xây dựng mô hình KNN

- Bước 2: Tính khoảng cách
 - Lưu ý quan trọng
 - chuẩn hóa dữ liệu
 - 2 phương pháp phổ biến:
 - Min-MaxScaling (Normalization): Đưa tất cả giá trị của một đặc trưng về một khoảng $[0, 1]$

$$X_{\text{scaled}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

- Standardization: Biến đổi dữ liệu sao cho đặc trưng có giá trị trung bình (μ) là 0 và độ lệch chuẩn (σ) là 1

$$X_{\text{scaled}} = \frac{X - \mu}{\sigma}$$

Quy trình xây dựng mô hình KNN

- Bước 3: Tìm k láng giềng gần nhất
 - Sắp xếp các khoảng cách đã tính ở trên theo thứ tự từ nhỏ đến lớn.
 - Chọn k điểm dữ liệu ứng với k khoảng cách nhỏ nhất

Quy trình xây dựng mô hình KNN

- Bước 4: Ra quyết định, dự đoán kết quả
 - Đối với bài toán phân loại:
 - Nhãn của điểm dữ liệu mới sẽ được gán cho lớp chiếm đa số trong k láng giềng được chọn
 - Đối với bài toán hồi quy
 - Nhãn của điểm dữ liệu mới là giá trị trung bình của giá trị của k láng giềng được chọn

Quy trình xây dựng mô hình KNN

- Bước 4: Ra quyết định, dự đoán kết quả
 - Đối với bài toán phân loại:
 - Nhãn của điểm dữ liệu mới sẽ được gán cho lớp chiếm đa số trong k láng giềng được chọn
 - Đối với bài toán hồi quy
 - Nhãn của điểm dữ liệu mới là giá trị trung bình của giá trị của k láng giềng được chọn

Thực hành

HS	Giờ học	Giờ chơi	Kết quả
HS01	2	7	Trượt
HS02	3	6	Trượt
HS03	4	5	Đậu
HS04	5	5	Đậu
HS05	6	3	Đậu
HS06	5	4	?????

HS	Giờ học	Giờ chơi	Điểm thi
HS01	2	7	4.0
HS02	3	6	4.5
HS03	4	5	6.0
HS04	5	5	7.5
HS05	6	3	8.0
HS06	5	4	?????

Thực hành

- Bước 1: Lựa chọn k
 - Chọn $k=3$
- Bước 2: Tính khoảng cách Euclidean
 - K/c từ $H6 = (5, 4)$ đến $H1 = (2, 7)$
 - K/c từ $H6 = (5, 4)$ đến $H2 = (3, 6)$
 - K/c từ $H6 = (5, 4)$ đến $H3 = (4, 5)$
 - K/c từ $H6 = (5, 4)$ đến $H4 = (5, 5)$
 - K/c từ $H6 = (5, 4)$ đến $H4 = (6, 3)$

HS	Giờ học	Giờ chơi	Kết quả
HS01	2	7	Trượt
HS02	3	6	Trượt
HS03	4	5	Đậu
HS04	5	5	Đậu
HS05	6	3	Đậu
HS06	5	4	?????

HS	Giờ học	Giờ chơi	Điểm thi
HS01	2	7	4.0
HS02	3	6	4.5
HS03	4	5	6.0
HS04	5	5	7.5
HS05	6	3	8.0
HS06	5	4	?????

Thực hành

- Bước 2: Tính khoảng cách Euclidean
 - K/c từ $H6 = (5, 4)$ đến $H1 = (2, 7)$
 - K/c từ $H6 = (5, 4)$ đến $H2 = (3, 6)$
 - K/c từ $H6 = (5, 4)$ đến $H3 = (4, 5)$
 - K/c từ $H6 = (5, 4)$ đến $H4 = (5, 5)$
 - K/c từ $H6 = (5, 4)$ đến $H4 = (6, 3)$
- Bước 3: Tìm k láng giềng gần nhất

Điểm DL	K/C	Kết quả	Điểm
H1			
H2			
H3			
H4			
H5			

HS	Giờ học	Giờ chơi	Kết quả
HS01	2	7	Trượt
HS02	3	6	Trượt
HS03	4	5	Đậu
HS04	5	5	Đậu
HS05	6	3	Đậu
HS06	5	4	?????

HS	Giờ học	Giờ chơi	Điểm thi
HS01	2	7	4.0
HS02	3	6	4.5
HS03	4	5	6.0
HS04	5	5	7.5
HS05	6	3	8.0
HS06	5	4	?????

Thực hành

- Bước 2: Tính khoảng cách Euclidean

Điểm DL	K/C	Kết quả	Điểm
H1			

=> 3 láng giềng gần nhất của H6 là:

- Bước 4: Ra quyết định – Dự đoán kết quả

HS	Giờ học	Giờ chơi	Kết quả
HS01	2	7	Trượt
HS02	3	6	Trượt
HS03	4	5	Đậu
HS04	5	5	Đậu
HS05	6	3	Đậu
HS06	5	4	?????

HS	Giờ học	Giờ chơi	Điểm thi
HS01	2	7	4.0
HS02	3	6	4.5
HS03	4	5	6.0
HS04	5	5	7.5
HS05	6	3	8.0
HS06	5	4	?????

Độ sáng	Độ bão hòa	Lớp học	Khoảng cách
1	25	Màu đỏ	10
4	20	Màu đỏ	25
5	50	Màu xanh da trời	33,54
7	55	Màu xanh da trời	45
6	10	Màu đỏ	47,17
3	18	?????	?????

K-MEANS

Giới thiệu



Ý tưởng cốt lõi của K-MEANS

- Tìm ra các phân cụm tự nhiên trong dữ liệu bằng cách tối thiểu tổng bình phương khoảng cách từ các điểm tới trọng tâm cụm của chúng

Tổng quan về K-MEANS

- Học không giám sát: Dữ liệu không cần được gán nhãn. Mục tiêu là chia dữ liệu thành các cụm sao cho các điểm trong từng cụm càng đồng nhất càng tốt.
- Phân hoạch (partitioning): Mỗi điểm dữ liệu được gán vào đúng một cụm duy nhất. Số lượng cụm do người dùng quyết định. Có một số phương pháp như elbow, silhouette score, gap statistic,...
- Trọng tâm cụm (Centroid): Được xác định bằng cách lấy trung bình các điểm dữ liệu trong cụm.
- Hàm mục tiêu: Là hàm số đo tổng bình phương khoảng cách từ mỗi điểm dữ liệu đến tâm cụm chứa nó

$$J(k) = \sum_{i=1}^k \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \mu_i\|^2$$

$$S = \arg \min_S \sum_{i=1}^k \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \mu_i\|^2$$

Quy trình xây dựng mô hình K-MEANS

- Bước 1: Chọn số cụm k
 - Bước 2: Khởi tạo trọng tâm ban đầu
 - Bước 3: Gán các điểm dữ liệu vào các cụm
 - Bước 4: Cập nhật trọng tâm
 - Bước 5: Kiểm tra hội tụ
-
- Lưu ý: Thuật toán lặp lại 2 bước 3 và 4 cho tới khi các trọng tâm không thay đổi hoặc sự thay đổi nhỏ hơn một ngưỡng cho trước.

Quy trình xây dựng mô hình KNN

- Lưu ý quan trọng
 - chuẩn hóa dữ liệu
 - 2 phương pháp phổ biến:
 - Min-MaxScaling (Normalization): Đưa tất cả giá trị của một đặc trưng về một khoảng [0, 1]

$$X_{\text{scaled}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

- Standardization: Biến đổi dữ liệu sao cho đặc trưng có giá trị trung bình (μ) là 0 và độ lệch chuẩn (σ) là 1

$$X_{\text{scaled}} = \frac{X - \mu}{\sigma}$$

Quy trình xây dựng mô hình K-MEANS

- Bước 1: Lựa chọn k
 - K có ảnh hưởng lớn đến độ phức tạp và khả năng tổng quát hóa của mô hình:
 - => Làm thế nào để chọn k?????

Quy trình xây dựng mô hình K-MEANS

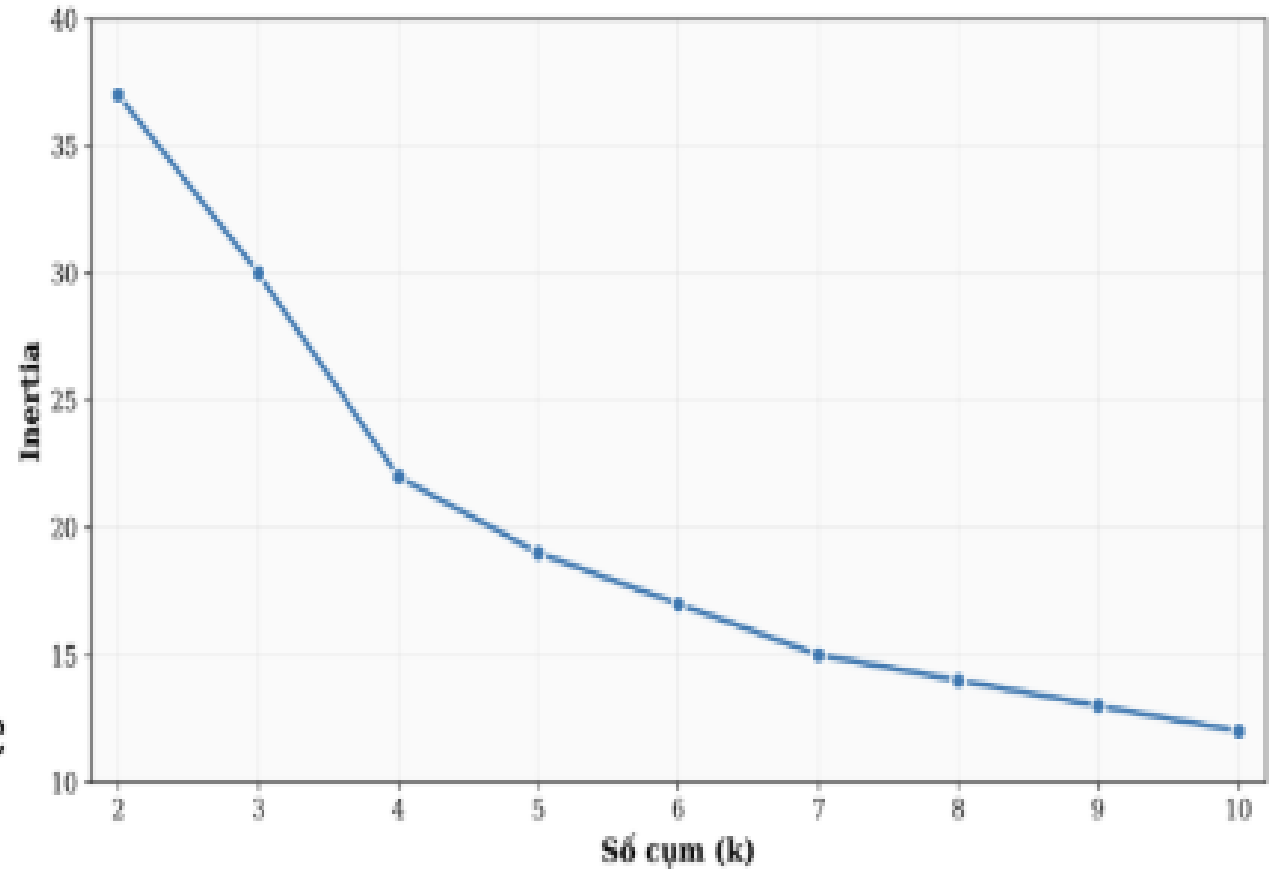
- Bước 1: Lựa chọn k

$$Inertia = \sum_{k=1}^K \sum_{x_i \in C_k} \|x_i - \mu_k\|^2$$

Diagram illustrating the formula for Inertia:

- $\sum_{k=1}^K$ is labeled "Cluster" (Cluster).
- $\sum_{x_i \in C_k}$ is labeled "Data point" (Data point).
- μ_k is labeled "Centroid of cluster k" (Centroid of cluster k).

Plot **Inertia vs K** and look for the "elbow" where reduction slows down



Quy trình xây dựng mô hình K-MEANS

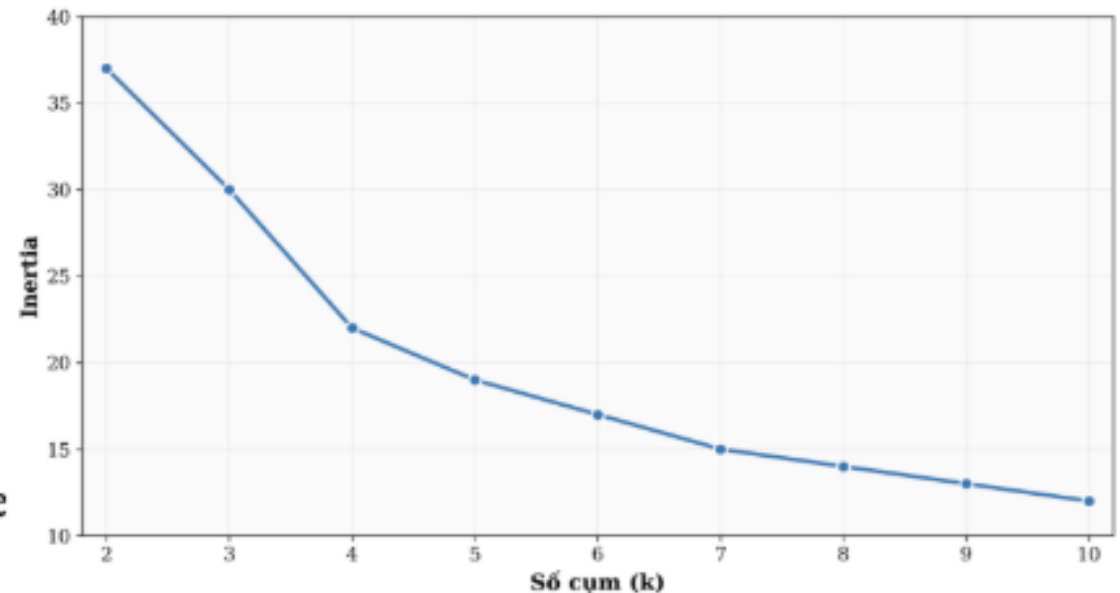
- Bước 1: Lựa chọn k
 - Không có một công thức toán học nào cho giá trị k tốt nhất.
 - Dựa trên kinh nghiệm:
 - Elbow
 - Điểm Silhouette
 - Gap statistic

$$Inertia = \sum_{k=1}^K \sum_{x_i \in C_k} \|x_i - \mu_k\|^2$$

Diagram illustrating the formula for Inertia:

- $\sum_{k=1}^K$ is labeled "Cluster" (with an arrow pointing to the summation index k).
- $\sum_{x_i \in C_k}$ is labeled "Data point" (with an arrow pointing to the summation index x_i).
- μ_k is labeled "Centroid of cluster k" (with an arrow pointing to the symbol μ_k).

Plot **Inertia vs K** and look for the "elbow" where reduction slows down



Quy trình xây dựng mô hình K-MEANS

- Bước 2: Khởi tạo trọng tâm ban đầu
 - Phương pháp Forgy: Chọn ngẫu nhiên k điểm dữ liệu làm tâm
 - Phương pháp Random Partition: Gán ngẫu nhiên từng điểm dữ liệu vào các cụm, sau đó tính trọng tâm ban đầu của từng cụm

Quy trình xây dựng mô hình K-MEANS

- Bước 3: Gán các điểm dữ liệu vào các cụm
 - Với từng trọng tâm:
 - Tính khoảng cách của các điểm dữ liệu đến nó
 - Gán điểm dữ liệu vào cụm có khoảng cách nhỏ nhất

Quy trình xây dựng mô hình KNN

- Bước 4: Cập nhật trọng tâm

Quy trình xây dựng mô hình KNN

- Bước 5: Kiểm tra hội tụ
 - Thuật toán lặp lại 2 bước 3 và 4 cho tới khi các trọng tâm không thay đổi hoặc sự thay đổi nhỏ hơn một ngưỡng cho trước.

Thực hành

SV	Điểm học tập	Điểm rèn luyện
S01	85	83
S02	70	59
S03	90	50
S04	50	85
S05	50	50
S06	90	85

Thực hành

- Bước 1: Lựa chọn k
 - Chọn $k=3$
- Bước 2: Khởi tạo trong tâm ban đầu
 - Theo phương pháp Forgy:
 - $\mu_1^0 = (85, 83)$
 - $\mu_2^0 = (70, 59)$
 - $\mu_3^0 = (90, 50)$

SV	Điểm học tập	Điểm rèn luyện
S01	85	83
S02	70	59
S03	90	50
S04	50	85
S05	50	50
S06	90	85

Thực hành

- Bước 2: Khởi tạo trong tâm ban đầu
 - Theo phương pháp Forgy:
 - $\mu_1^0 = (85, 83)$
 - $\mu_2^0 = (70, 59)$
 - $\mu_3^0 = (90, 50)$
- Bước 3: Gán điểm dữ liệu vào cụm gần nhất

SV	Điểm học tập	Điểm rèn luyện
S01	85	83
S02	70	59
S03	90	50
S04	50	85
S05	50	50
S06	90	85

- Cụm 1: S1, S6
- Cụm 2: S2, S4, S5
- Cụm 3: S3

Điểm DL	K/C đến tâm 1	K/C đến tâm 2	K/C đến tâm 3
S1	0		
S2		0	
S3			0
S4			
S5			
S6			

Thực hành

- Bước 3: Gán điểm dữ liệu vào cụm gần nhất
 - Cụm 1: S1, S6
 - Cụm 2: S2, S4, S5
 - Cụm 3: S3
- Bước 4: Cập nhật trọng tâm

$$\begin{aligned}\mu_1^1 &= (87,5; 84) \\ \mu_2^1 &= (56,7; 64,7) \\ \mu_3^1 &= (90, 50)\end{aligned}$$

SV	Điểm học tập	Điểm rèn luyện
S01	85	83
S02	70	59
S03	90	50
S04	50	85
S05	50	50
S06	90	85

Thực hành

- Bước 4: Cập nhật trọng tâm

$$\mu_1^1 =$$

$$\mu_2^1 =$$

$$\mu_3^1 =$$

- Bước 5: Kiểm tra hội tụ và lặp lại

HS	Giờ học	Giờ chơi	Kết quả
HS01	2	7	Trượt
HS02	3	6	Trượt
HS03	4	5	Đậu
HS04	5	5	Đậu
HS05	6	3	Đậu
HS06	5	4	?????

HS	Giờ học	Giờ chơi	Điểm thi
HS01	2	7	4.0
HS02	3	6	4.5
HS03	4	5	6.0
HS04	5	5	7.5
HS05	6	3	8.0
HS06	5	4	?????

Thực hành

Điểm	Feature 1	Feature 2	Feature 3
p_1	2.1	3.1	1.6
p_2	3.2	3.6	2.1
p_3	3.6	3.1	2.6
p_4	7.9	8.1	7.6
p_5	8.6	8.7	8.2
p_6	9.1	8.1	8.6
p_7	1.2	2.1	1.7