# K-Means Clustering

## K-Means Clustering: Mathematical Procedure

### Algorithm Steps

1. **Initialization Step**
   - Choose $K$ initial centroids randomly from data points
   - Let centroids be: $\mu_1, \mu_2, \ldots, \mu_K$
2. **Assignment Step**
   - For each data point $\mathbf{x}_i$, calculate distance to all centroids:

$$d(\mathbf{x}_i, \mu_j) = \|\mathbf{x}_i - \mu_j\|^2$$

   - Assign point to closest centroid:

$$c_i = \arg\min_j \|\mathbf{x}_i - \mu_j\|^2$$

3. **Update Step**
   - Recalculate centroids as mean of assigned points:

$$\mu_j = \frac{1}{|C_j|} \sum_{\mathbf{x}_i \in C_j} \mathbf{x}_i$$

     where $|C_j|$ is number of points in cluster $j$
4. **Convergence Check**
   - Repeat Steps 2-3 until:
     - Centroids don't change significantly, OR
     - Cluster assignments remain stable, OR
     - Maximum iterations reached

## Mathematical Formulas

- **Euclidean Distance:**

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{(q_x - p_x)^2 + (q_y - p_y)^2}$$

- **Squared Euclidean Distance:**

$$d^2(\mathbf{p}, \mathbf{q}) = (q_x - p_x)^2 + (q_y - p_y)^2$$

- **Centroid Calculation:**

$$\mu = \left( \frac{\sum x_i}{n}, \frac{\sum y_i}{n} \right)$$

- **Sum of Squared Errors (SSE):**

$$\text{SSE} = \sum_{j=1}^{K} \sum_{\mathbf{x}_i \in C_j} \|\mathbf{x}_i - \mu_j\|^2$$

# Example 1: Student Performance Clustering

## Student Data (Exam Scores out of 100)

| Student | Math Score | English Score |
|---|---|---|
| A | 85 | 90 |
| B | 92 | 88 |
| C | 45 | 50 |
| D | 50 | 55 |
| E | 88 | 85 |
| F | 40 | 45 |
| G | 95 | 92 |
| H | 48 | 52 |

## K=2 Clusters: High Performers vs Low Performers

**Initial Centroids:** $\mu_1 = \text{A}(85, 90)$, $\mu_2 = \text{C}(45, 50)$

**Iteration 1 - Distance Calculations:**

$$
\begin{aligned}
d(\mathrm{A}, \mu_1) &= (85-85)^2 + (90-90)^2 = 0 \\
d(\mathrm{A}, \mu_2) &= (85-45)^2 + (90-50)^2 = 1600 + 1600 = 3200 \\
d(\mathrm{B}, \mu_1) &= (92-85)^2 + (88-90)^2 = 49 + 4 = 53 \\
d(\mathrm{B}, \mu_2) &= (92-45)^2 + (88-50)^2 = 2209 + 1444 = 3653 \\
d(\mathrm{C}, \mu_1) &= (45-85)^2 + (50-90)^2 = 1600 + 1600 = 3200 \\
d(\mathrm{C}, \mu_2) &= (45-45)^2 + (50-50)^2 = 0 \\
d(\mathrm{D}, \mu_1) &= (50-85)^2 + (55-90)^2 = 1225 + 1225 = 2450 \\
d(\mathrm{D}, \mu_2) &= (50-45)^2 + (55-50)^2 = 25 + 25 = 50 \\
d(\mathrm{E}, \mu_1) &= (88-85)^2 + (85-90)^2 = 9 + 25 = 34 \\
d(\mathrm{E}, \mu_2) &= (88-45)^2 + (85-50)^2 = 1849 + 1225 = 3074 \\
d(\mathrm{F}, \mu_1) &= (40-85)^2 + (45-90)^2 = 2025 + 2025 = 4050 \\
d(\mathrm{F}, \mu_2) &= (40-45)^2 + (45-50)^2 = 25 + 25 = 50 \\
d(\mathrm{G}, \mu_1) &= (95-85)^2 + (92-90)^2 = 100 + 4 = 104 \\
d(\mathrm{G}, \mu_2) &= (95-45)^2 + (92-50)^2 = 2500 + 1764 = 4264 \\
d(\mathrm{H}, \mu_1) &= (48-85)^2 + (52-90)^2 = 1369 + 1444 = 2813 \\
d(\mathrm{H}, \mu_2) &= (48-45)^2 + (52-50)^2 = 9 + 4 = 13
\end{aligned}
$$

**Cluster Assignments:**

- **High Performers:** A, B, E, G
- **Low Performers:** C, D, F, H

## Iteration 2

**Update Centroids:**

$$
\begin{aligned}
\mu_1 &= \left( \frac{85+92+88+95}{4}, \frac{90+88+85+92}{4} \right) = (90, 88.75) \\
\mu_2 &= \left( \frac{45+50+40+48}{4}, \frac{50+55+45+52}{4} \right) = (45.75, 50.5)
\end{aligned}
$$

**Re-check Assignments:**

$$
\begin{aligned}
d(\mathrm{A}, \mu_1) &= (85-90)^2 + (90-88.75)^2 = 25 + 1.56 = 26.56 \\
d(\mathrm{A}, \mu_2) &= (85-45.75)^2 + (90-50.5)^2 = 1540.56 + 1560.25 = 3100.81 \\
d(\mathrm{D}, \mu_1) &= (50-90)^2 + (55-88.75)^2 = 1600 + 1139.06 = 2739.06 \\
d(\mathrm{D}, \mu_2) &= (50-45.75)^2 + (55-50.5)^2 = 18.06 + 20.25 = 38.31
\end{aligned}
$$

**Final Clusters:**

- **High Performers:** A, B, E, G
- **Low Performers:** C, D, F, H

# Example 2: Body Type Classification

## Person Data (Height in cm, Weight in kg)

| Person | Height | Weight |
|---|---|---|
| Person1 | 155 | 45 |
| Person2 | 160 | 50 |
| Person3 | 165 | 55 |
| Person4 | 185 | 85 |
| Person5 | 190 | 90 |
| Person6 | 195 | 95 |
| Person7 | 175 | 65 |
| Person8 | 172 | 68 |
| Person9 | 168 | 62 |
| Person10 | 162 | 58 |

## K=3 Clusters: Short, Normal, Tall

**Initial Centroids:** $\mu_1 = \text{Person1}(155, 45)$, $\mu_2 = \text{Person3}(165, 55)$, $\mu_3 = \text{Person5}(190, 90)$

**Iteration 1 - Sample Distance Calculations:**

$$d(\text{P1}, \mu_1) = (155 - 155)^2 + (45 - 45)^2 = 0$$
$$d(\text{P1}, \mu_2) = (155 - 165)^2 + (45 - 55)^2 = 100 + 100 = 200$$
$$d(\text{P1}, \mu_3) = (155 - 190)^2 + (45 - 90)^2 = 1225 + 2025 = 3250$$
$$d(\text{P3}, \mu_1) = (165 - 155)^2 + (55 - 45)^2 = 100 + 100 = 200$$
$$d(\text{P3}, \mu_2) = (165 - 165)^2 + (55 - 55)^2 = 0$$
$$d(\text{P3}, \mu_3) = (165 - 190)^2 + (55 - 90)^2 = 625 + 1225 = 1850$$
$$d(\text{P7}, \mu_1) = (175 - 155)^2 + (65 - 45)^2 = 400 + 400 = 800$$
$$d(\text{P7}, \mu_2) = (175 - 165)^2 + (65 - 55)^2 = 100 + 100 = 200$$
$$d(\text{P7}, \mu_3) = (175 - 190)^2 + (65 - 90)^2 = 225 + 625 = 850$$

**Cluster Assignments:**

- **Short:** Person1, Person2, Person10
- **Normal:** Person3, Person7, Person8, Person9
- **Tall:** Person4, Person5, Person6

### Iteration 2

**Update Centroids:**

$$\mu_1 = \left(\frac{155+160+162}{3}, \frac{45+50+58}{3}\right) = (159, 51)$$
$$\mu_2 = \left(\frac{165+175+172+168}{4}, \frac{55+65+68+62}{4}\right) = (170, 62.5)$$
$$\mu_3 = \left(\frac{185+190+195}{3}, \frac{85+90+95}{3}\right) = (190, 90)$$

**Re-check Critical Points:**

$$d(\text{P2}, \mu_1) = (160-159)^2 + (50-51)^2 = 1+1 = 2$$
$$d(\text{P2}, \mu_2) = (160-170)^2 + (50-62.5)^2 = 100 + 156.25 = 256.25$$
$$d(\text{P9}, \mu_1) = (168-159)^2 + (62-51)^2 = 81 + 121 = 202$$
$$d(\text{P9}, \mu_2) = (168-170)^2 + (62-62.5)^2 = 4 + 0.25 = 4.25$$

**Final Clusters:**

- **Short:** Person1, Person2, Person10 (Height: 155-162cm)
- **Normal:** Person3, Person7, Person8, Person9 (Height: 165-175cm)
- **Tall:** Person4, Person5, Person6 (Height: 185-195cm)

# Example 3: Customer Spending Behavior

## Monthly Spending Data ($)

| Customer | Food | Entertainment | Shopping |
|---|---|---|---|
| C1 | 200 | 50 | 100 |
| C2 | 180 | 40 | 80 |
| C3 | 500 | 200 | 400 |
| C4 | 450 | 180 | 350 |
| C5 | 220 | 60 | 120 |
| C6 | 480 | 190 | 380 |
| C7 | 190 | 45 | 90 |
| C8 | 510 | 210 | 420 |

## K=2 Clusters: Conservative vs Big Spenders

**Initial Centroids:** $\mu_1 = C1(200, 50, 100)$, $\mu_2 = C3(500, 200, 400)$

**Distance Calculations (3D Euclidean Squared):**

$d(C1, \mu_1) = (200-200)^2 + (50-50)^2 + (100-100)^2 = 0$
$d(C1, \mu_2) = (200-500)^2 + (50-200)^2 + (100-400)^2 = 90000 + 22500 + 90000 = 202500$
$d(C2, \mu_1) = (180-200)^2 + (40-50)^2 + (80-100)^2 = 400 + 100 + 400 = 900$
$d(C2, \mu_2) = (180-500)^2 + (40-200)^2 + (80-400)^2 = 102400 + 25600 + 102400 = 230400$
$d(C5, \mu_1) = (220-200)^2 + (60-50)^2 + (120-100)^2 = 400 + 100 + 400 = 900$
$d(C5, \mu_2) = (220-500)^2 + (60-200)^2 + (120-400)^2 = 78400 + 19600 + 78400 = 176400$

**Cluster Assignments:**

- **Conservative Spenders:** C1, C2, C5, C7
- **Big Spenders:** C3, C4, C6, C8

# Example 4: Simple 2D Points

## Data Points

| Point | Coordinates |
|---|---|
| A | (1, 1) |
| B | (1, 2) |
| C | (3, 4) |
| D | (3, 5) |
| E | (6, 1) |
| F | (6, 2) |

## Iteration 1

**Initial Centroids:** $\mu_1 = A(1,1)$, $\mu_2 = C(3,4)$

**Distance Calculations:**

$$
\begin{aligned}
d(A, \mu_1) &= (1-1)^2 + (1-1)^2 = 0 \\
d(A, \mu_2) &= (1-3)^2 + (1-4)^2 = 4 + 9 = 13 \\
d(B, \mu_1) &= (1-1)^2 + (2-1)^2 = 0 + 1 = 1 \\
d(B, \mu_2) &= (1-3)^2 + (2-4)^2 = 4 + 4 = 8 \\
d(C, \mu_1) &= (3-1)^2 + (4-1)^2 = 4 + 9 = 13 \\
d(C, \mu_2) &= (3-3)^2 + (4-4)^2 = 0 \\
d(D, \mu_1) &= (3-1)^2 + (5-1)^2 = 4 + 16 = 20 \\
d(D, \mu_2) &= (3-3)^2 + (5-4)^2 = 0 + 1 = 1 \\
d(E, \mu_1) &= (6-1)^2 + (1-1)^2 = 25 + 0 = 25 \\
d(E, \mu_2) &= (6-3)^2 + (1-4)^2 = 9 + 9 = 18 \\
d(F, \mu_1) &= (6-1)^2 + (2-1)^2 = 25 + 1 = 26 \\
d(F, \mu_2) &= (6-3)^2 + (2-4)^2 = 9 + 4 = 13
\end{aligned}
$$

**Cluster Assignments:**

- **Cluster 1:** A, B (closer to $\mu_1$)
- **Cluster 2:** C, D, E, F (closer to $\mu_2$)

## Iteration 2

**Update Centroids:**

$$
\begin{aligned}
\mu_1 &= \left( \frac{1+1}{2}, \frac{1+2}{2} \right) = (1, 1.5) \\
\mu_2 &= \left( \frac{3+3+6+6}{4}, \frac{4+5+1+2}{4} \right) = (4.5, 3)
\end{aligned}
$$

**Re-check Critical Points:**

$$
\begin{aligned}
d(E, \mu_1) &= (6-1)^2 + (1-1.5)^2 = 25 + 0.25 = 25.25 \\
d(E, \mu_2) &= (6-4.5)^2 + (1-3)^2 = 2.25 + 4 = 6.25 \\
d(F, \mu_1) &= (6-1)^2 + (2-1.5)^2 = 25 + 0.25 = 25.25 \\
d(F, \mu_2) &= (6-4.5)^2 + (2-3)^2 = 2.25 + 1 = 3.25
\end{aligned}
$$

**Final Assignments:**

- **Cluster 1:** A, B
- **Cluster 2:** C, D, E, F

# Student's Work: Customer Segmentation

## Customer Data (Annual Spending in $1000s)

| Customer | Groceries | Electronics |
|---|---|---|
| C1 | 15 | 2 |
| C2 | 12 | 1 |
| C3 | 3 | 20 |
| C4 | 2 | 25 |
| C5 | 18 | 3 |

## Iteration 1

**Initial Centroids:** $\mu_1 = C1(15, 2)$, $\mu_2 = C3(3, 20)$

**Distance Calculations (Squared Euclidean):**

$$
\begin{aligned}
d(C1, \mu_1) &= (15-15)^2 + (2-2)^2 = 0 \\
d(C1, \mu_2) &= (15-3)^2 + (2-20)^2 = 144 + 324 = 468 \\
d(C2, \mu_1) &= (12-15)^2 + (1-2)^2 = 9 + 1 = 10 \\
d(C2, \mu_2) &= (12-3)^2 + (1-20)^2 = 81 + 361 = 442 \\
d(C3, \mu_1) &= (3-15)^2 + (20-2)^2 = 144 + 324 = 468 \\
d(C3, \mu_2) &= (3-3)^2 + (20-20)^2 = 0 \\
d(C4, \mu_1) &= (2-15)^2 + (25-2)^2 = 169 + 529 = 698 \\
d(C4, \mu_2) &= (2-3)^2 + (25-20)^2 = 1 + 25 = 26 \\
d(C5, \mu_1) &= (18-15)^2 + (3-2)^2 = 9 + 1 = 10 \\
d(C5, \mu_2) &= (18-3)^2 + (3-20)^2 = 225 + 289 = 514
\end{aligned}
$$

**Cluster Assignments:**

- **Cluster 1:** C1, C2, C5
- **Cluster 2:** C3, C4

## Iteration 2

**Update Centroids:**

$$
\begin{aligned}
\mu_1 &= \left( \frac{15+12+18}{3}, \frac{2+1+3}{3} \right) = (15, 2) \\
\mu_2 &= \left( \frac{3+2}{2}, \frac{20+25}{2} \right) = (2.5, 22.5)
\end{aligned}
$$

**Re-check Distances:**

$$d(C1, \mu_1) = 0, \qquad d(C1, \mu_2) = (15 - 2.5)^2 + (2 - 22.5)^2 = 156.25 + 420.25 = 576.5$$
$$d(C2, \mu_1) = 10, \qquad d(C2, \mu_2) = (12 - 2.5)^2 + (1 - 22.5)^2 = 90.25 + 462.25 = 552.5$$
$$d(C5, \mu_1) = 10, \qquad d(C5, \mu_2) = (18 - 2.5)^2 + (3 - 22.5)^2 = 240.25 + 380.25 = 620.5$$

**Final Result:**

- **Budget Shoppers (Cluster 1):** C1, C2, C5
- **Tech Enthusiasts (Cluster 2):** C3, C4

# Convergence Analysis

## SSE Calculation for Student Example

$$\begin{aligned}
\text{SSE}_{\text{High Performers}} &= d(A, \mu_1)^2 + d(B, \mu_1)^2 + d(E, \mu_1)^2 + d(G, \mu_1)^2 \\
&= 26.56 + 8 + 13 + 29 = 76.56 \\
\text{SSE}_{\text{Low Performers}} &= d(C, \mu_2)^2 + d(D, \mu_2)^2 + d(F, \mu_2)^2 + d(H, \mu_2)^2 \\
&= 13.56 + 38.31 + 63.06 + 13 = 127.93 \\
\text{Total SSE} &= 76.56 + 127.93 = 204.49
\end{aligned}$$

## Key Observations

- K-means effectively groups similar students by performance
- Body type clustering reveals natural height-based categories
- Customer segmentation identifies distinct spending patterns
- Algorithm converges quickly with well-separated data
- Feature scaling may be needed for different measurement units