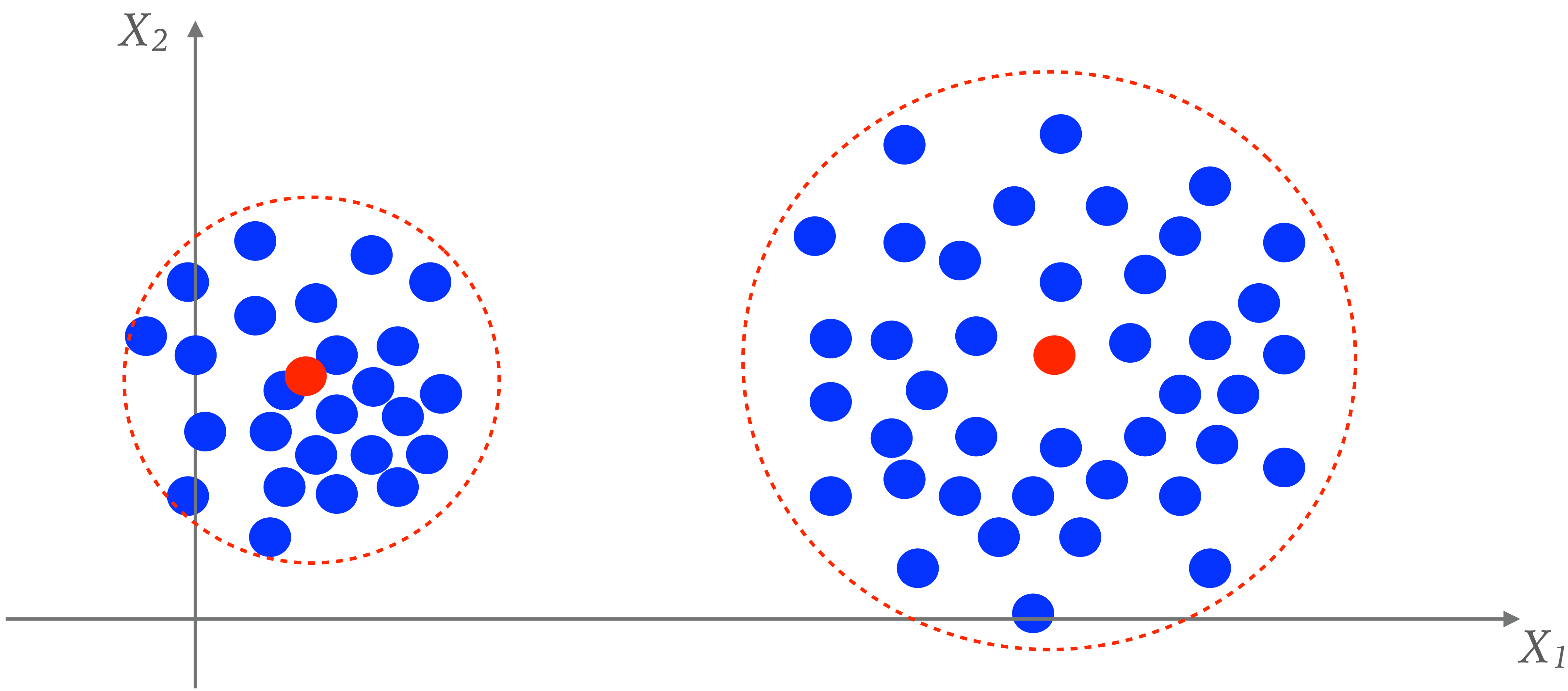
CSCI 347: Introduction to Data Mining

k-means

CLUSTERING

- Clustering is broadly and vaguely defined as finding groups of similar entities in a data set
- K-means is a representative-based algorithm that finds a specified number *k* of clusters



K-MEANS CLUSTERING

- Clustering is broadly and vaguely defined as finding groups of similar entities in a data set
- K-means is an algorithm that:

K-MEANS CLUSTERING

- Clustering is broadly and vaguely defined as finding groups of similar entities in a data set
- K-means is an algorithm that:
  - Requires the number of clusters to be found, k, as an input parameter
  - Iteratively updates cluster representatives (means) and cluster assignments (assignments of points to cluster means)
  - Converges when the updates to means are small enough
  - Finds a local minimum of the objective function:

$$J = \sum_{j=1}^{k} \sum_{x_i \in C_j} ||x_i - \mu_j||_2^2$$

K-MEANS CLUSTERING EXAMPLE

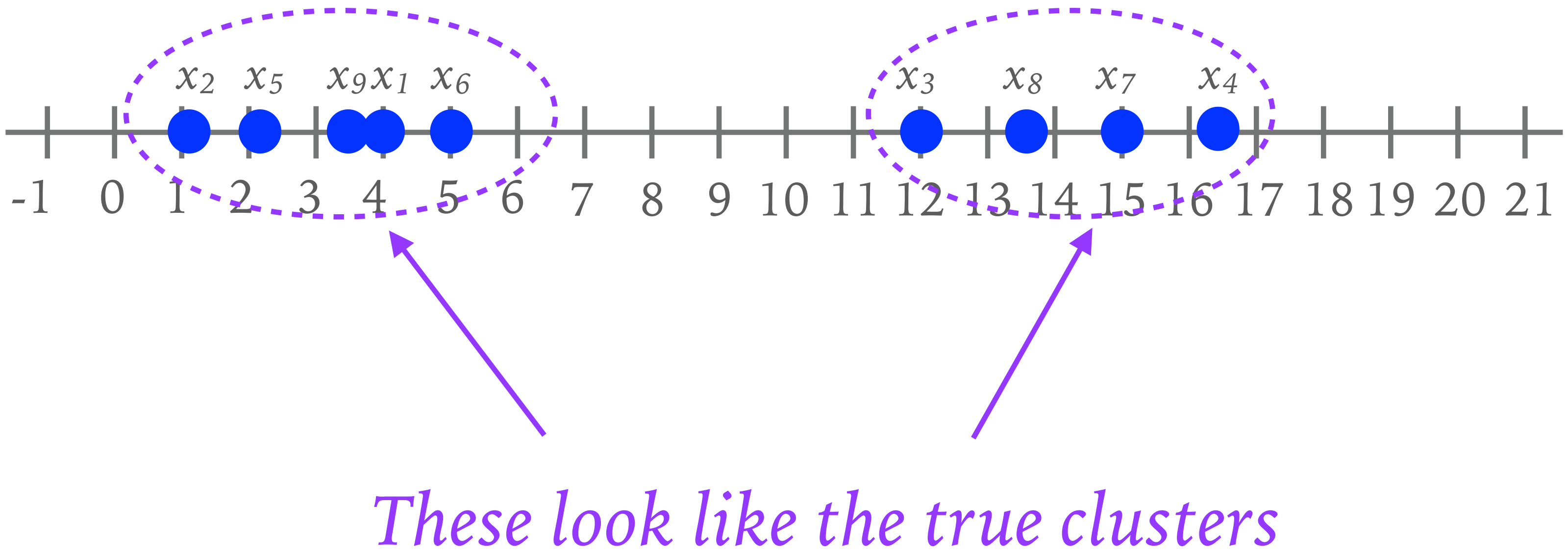
1-dimensional example with k=2

	X_1
x_1	4
x_2	1.1
x_3	12
x_4	16.4
x_5	2.3
x_6	5
x_7	15
x_8	13.7
x_9	3.5

K-MEANS CLUSTERING EXAMPLE

1-dimensional example with $k=2$

	X_1
x_1	4
x_2	1.1
x_3	12
x_4	16.4
x_5	2.3
x_6	5
x_7	15
x_8	13.7
x_9	3.5

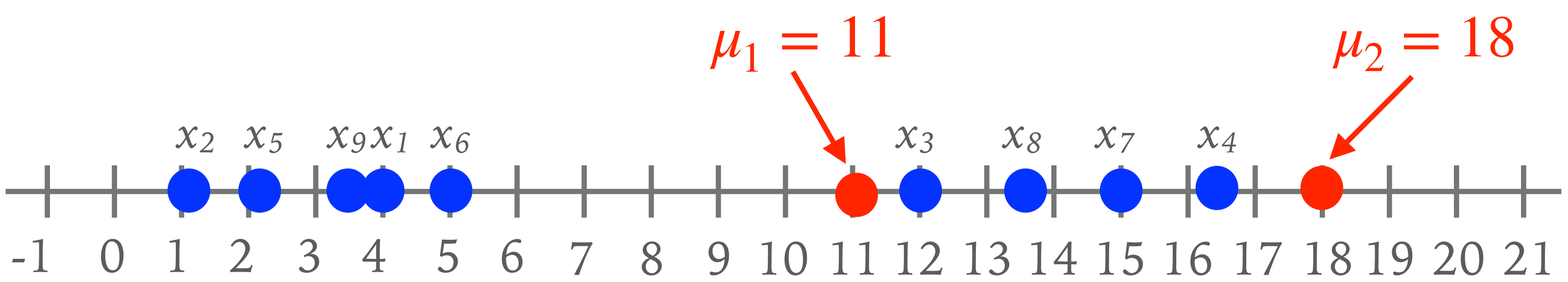


K-MEANS CLUSTERING EXAMPLE

1-dimensional example with k=2

	X_1
x_1	4
x_2	1.1
x_3	12
x_4	16.4
x_5	2.3
x_6	5
x_7	15
x_8	13.7
x_9	3.5

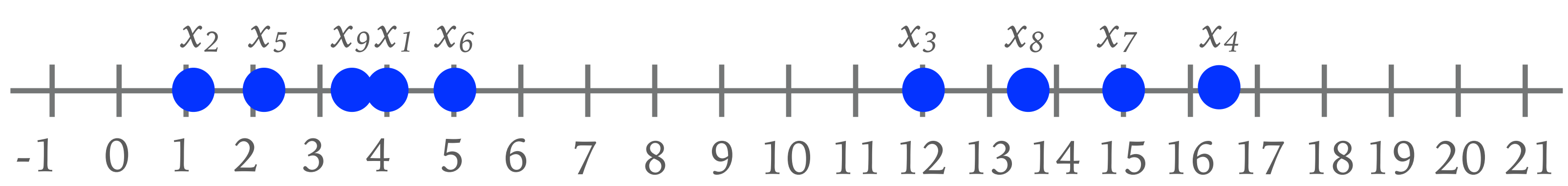
Step 1: randomly initialize 2 means



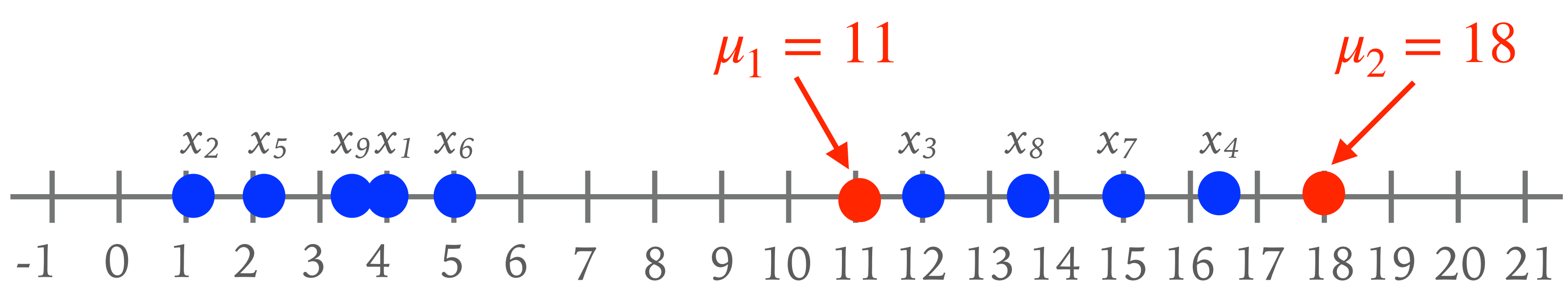
K-MEANS CLUSTERING EXAMPLE

1-dimensional example with k=2

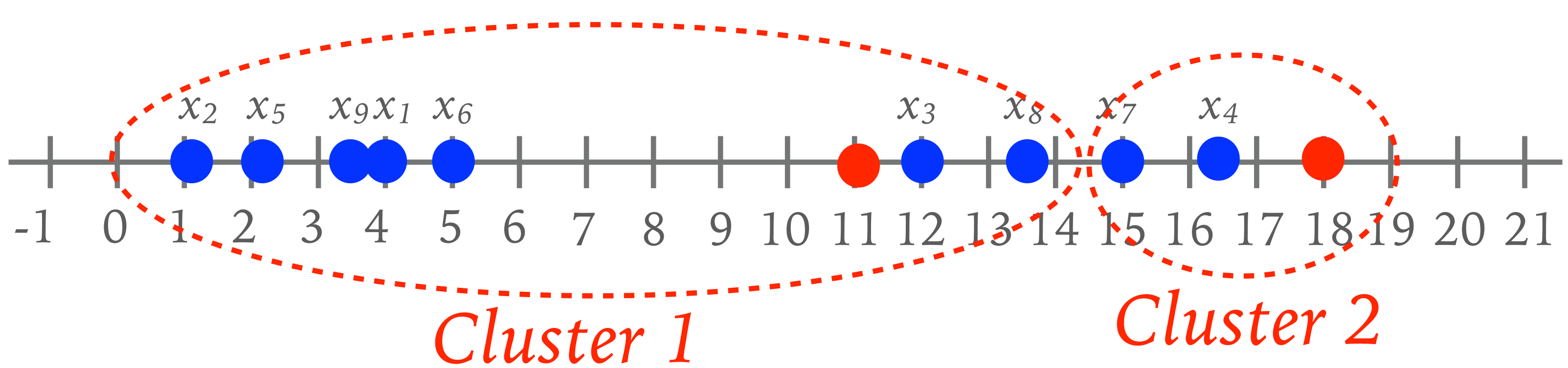
	X_1
x_1	4
x_2	1.1
x_3	12
x_4	16.4
x_5	2.3
x_6	5
x_7	15
x_8	13.7
x_9	3.5



Step 1: randomly initialize 2 means



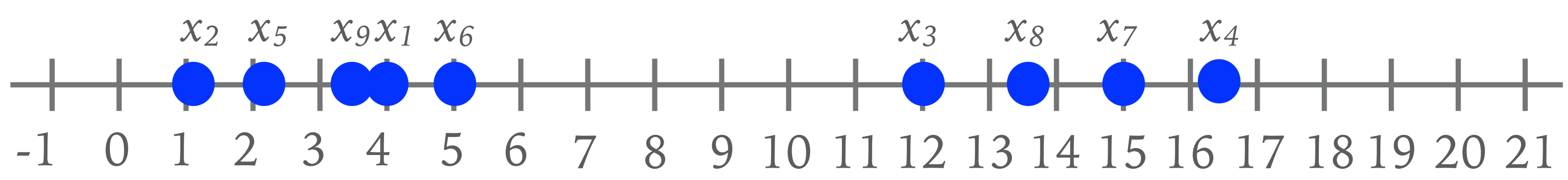
Step 2: assign each point to the cluster with the closest mean



K-MEANS CLUSTERING EXAMPLE

1-dimensional example with $k=2$

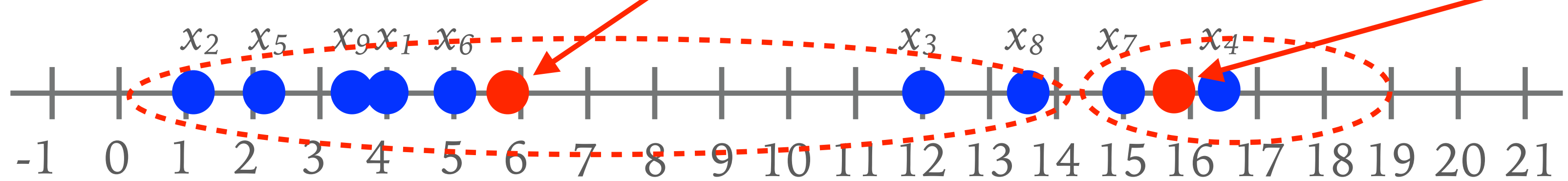
	X_1
x_1	4
x_2	1.1
x_3	12
x_4	16.4
x_5	2.3
x_6	5
x_7	15
x_8	13.7
x_9	3.5



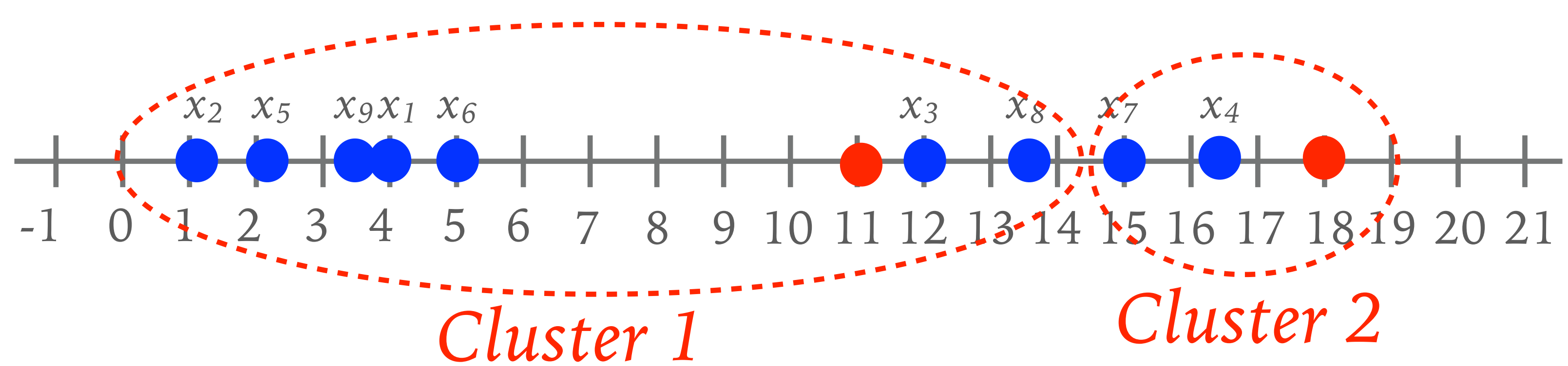
Step 3: re-compute the means based on cluster membership

$$\mu_1 = \frac{x_1 + x_2 + x_3 + x_5 + x_6 + x_8 + x_9}{8} = \frac{41.6}{8} = 5.94$$

$$\mu_2 = \frac{x_4 + x_7}{2} = \frac{31.4}{2} = 15.7$$



Step 2: assign each point to the cluster with the closest mean



K-MEANS CLUSTERING EXAMPLE

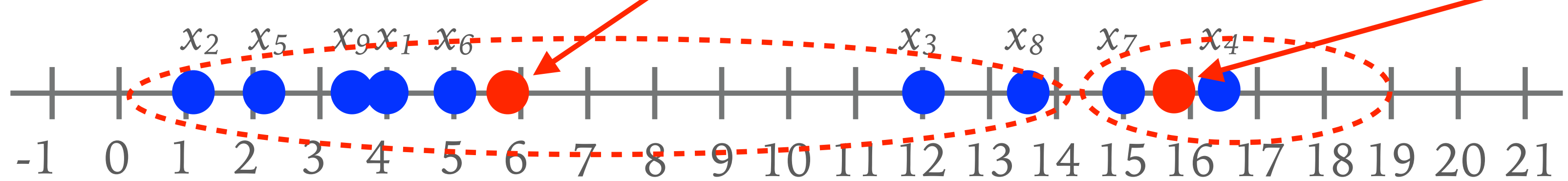
1-dimensional example with k=2

	X_1
x_1	4
x_2	1.1
x_3	12
x_4	16.4
x_5	2.3
x_6	5
x_7	15
x_8	13.7
x_9	3.5

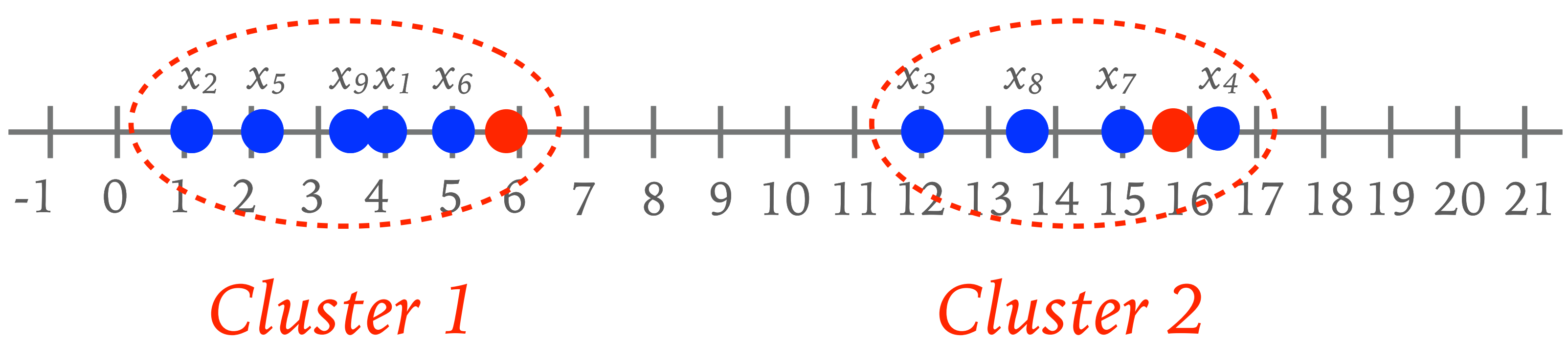
Step 3: re-compute the means based on cluster membership

$$\mu_1 = \frac{x_1 + x_2 + x_3 + x_5 + x_6 + x_8 + x_9}{8} = \frac{41.6}{7} = 5.94$$

$$\mu_2 = \frac{x_4 + x_7}{2} = \frac{31.4}{2} = 15.7$$



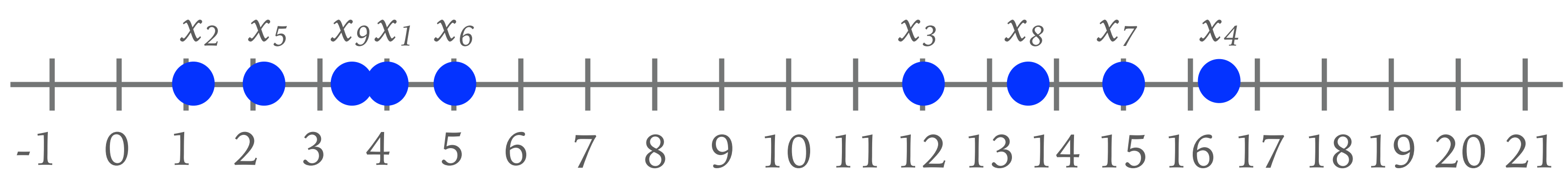
Step 4: assign each point to the cluster with the closest mean



K-MEANS CLUSTERING EXAMPLE

1-dimensional example with k=2

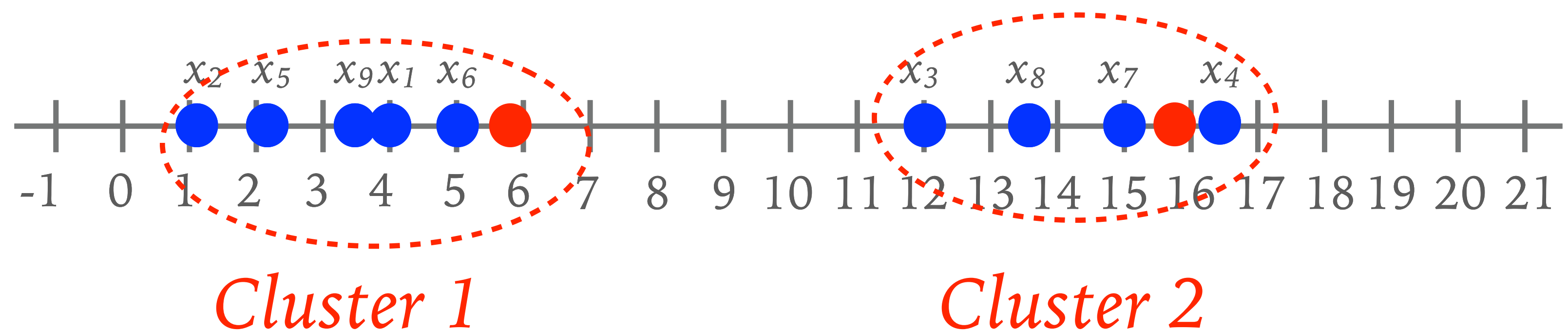
	X_1
x_1	4
x_2	1.1
x_3	12
x_4	16.4
x_5	2.3
x_6	5
x_7	15
x_8	13.7
x_9	3.5



Step 5: re-compute the means based on cluster membership

$$\mu_1 = \frac{x_1 + x_2 + x_5 + x_6 + x_9}{5} = \frac{15.9}{5} = 3.18 \qquad \mu_2 = \frac{x_3 + x_4 + x_7 + x_8}{4} = \frac{57.1}{4} = 14.3$$

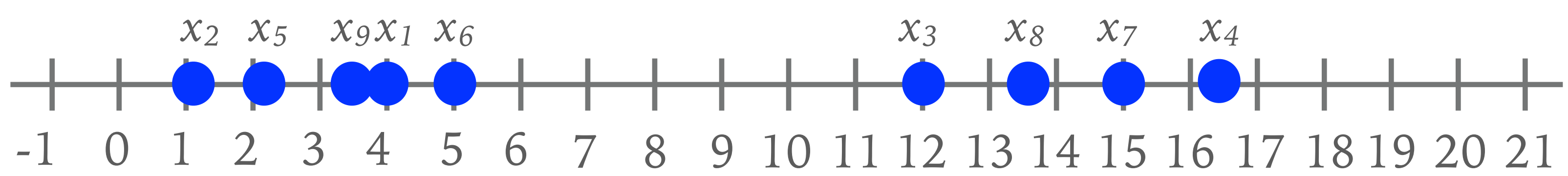
Step 4: assign each point to the cluster with the closest mean



K-MEANS CLUSTERING EXAMPLE

1-dimensional example with k=2

	X_1
x_1	4
x_2	1.1
x_3	12
x_4	16.4
x_5	2.3
x_6	5
x_7	15
x_8	13.7
x_9	3.5

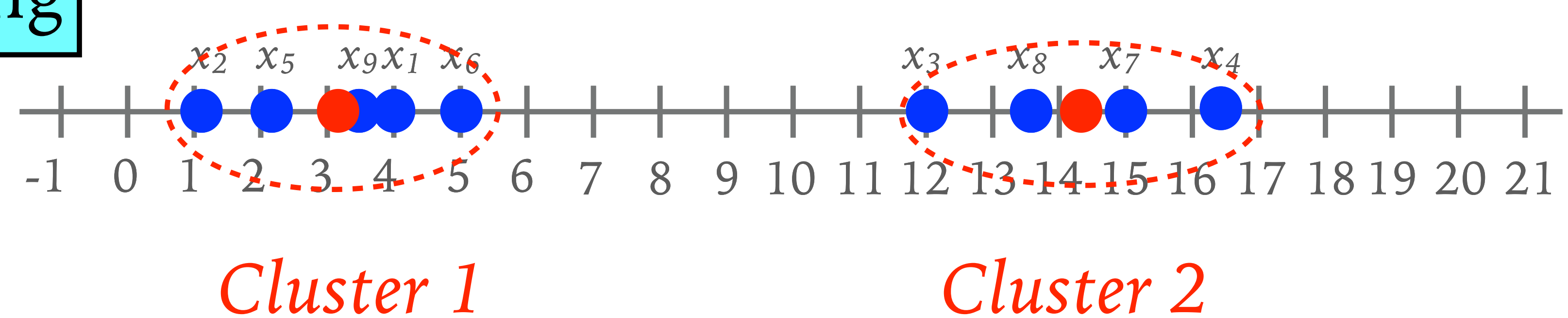


Step 5: re-compute the means based on cluster membership

$$\mu_1 = \frac{x_1 + x_2 + x_5 + x_6 + x_9}{5} = \frac{15.9}{5} = 3.18 \qquad \mu_2 = \frac{x_3 + x_4 + x_7 + x_8}{4} = \frac{57.1}{4} = 14.3$$

Step 6: assign each point to the cluster with the closest mean

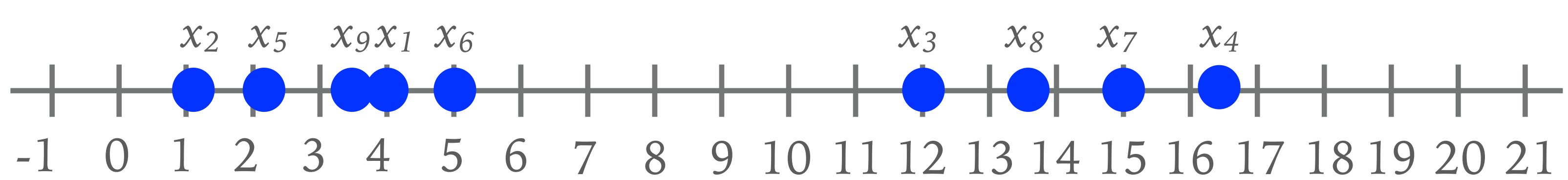
No change: stop iterating



2-MEANS CLUSTERING EXAMPLE

K-means iterates 2 steps until convergence:

	X_1
x_1	4
x_2	1.1
x_3	12
x_4	16.4
x_5	2.3
x_6	5
x_7	15
x_8	13.7
x_9	3.5



Mean computation step: re-compute the means based on cluster membership

$$\mu_1 = \frac{x_1 + x_2 + x_5 + x_6 + x_9}{5} = \frac{15.9}{5} = 3.18 \qquad \mu_2 = \frac{x_3 + x_4 + x_7 + x_8}{4} = \frac{57.1}{4} = 14.3$$

Re-assignment step: assign each point to the cluster with the closest mean

No change: stop iterating

Cluster 1 *Cluster 2*

THE K-MEANS CLUSTERING ALGORITHM

k-means$(D \in R^{n \times m}, k, \epsilon)$:

$t = 0$

Randomly initialize k representatives $\mu_1, \ldots, \mu_k \in R^m$

repeat:

$t = t + 1$ // iteration count

$C_j = \varnothing$ for $j = 1, \ldots, k$ //re-initialize clusters to be empty

for each $x_p \in D$: //cluster assignment step

$j^* = \text{argmin}_{i \in \{1,\ldots,k\}} \{ ||x_p - \mu_i||_2^2 \}$ // find cluster representative with smallest distance to x_p

$C_{j^*} = C_{j^*} \cup \{x_p\}$ // add x_p to C_{j^*}

for each $i = 1, \ldots, k$: // representative update step

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

until:

$$\sum_{i=1}^{k} ||\mu_i^t - \mu_i^{t-1}||^2 \leq \epsilon$$

K-MEANS CLUSTERING OBJECTIVE

- Want to minimize of the following objective function wrt means:

	X_1	X_2
x_1	4	1
x_2	1.1	-0.2
x_3	12	5.4
x_4	16.4	11.2
x_5	2.3	1.1
x_6	5	2
x_7	15	17.2
x_8	13.7	11.1
x_9	3.5	1.2

$$J = \sum_{j=1}^{k} \sum_{x_i \in C_j} ||x_i - \mu_j||_2^2$$

$$J = \sum_{j=1}^{k} \sum_{x_i \in C_j} \left((x_i - \mu_j)^T (x_i - \mu_j) \right) = \sum_{j=1}^{k} \sum_{x_i \in C_j} \left(x_i^T x_i - x_i^T \mu_j - \mu_j^T x_i + \mu_j^T \mu_j \right) = \sum_{j=1}^{k} \sum_{x_i \in C_j} \left(x_i^T x_i - 2 x_i^T \mu_j + \mu_j^T \mu_j \right)$$

K-MEANS CLUSTERING OBJECTIVE

➤ Want to minimize of the following objective function wrt means:

	X_1	X_2
x_1	4	1
x_2	1.1	-0.2
x_3	12	5.4
x_4	16.4	11.2
x_5	2.3	1.1
x_6	5	2
x_7	15	17.2
x_8	13.7	11.1
x_9	3.5	1.2

$$J = \sum_{j=1}^{k} \sum_{x_i \in C_j} ||x_i - \mu_j||_2^2$$

$$J = \sum_{j=1}^{k} \sum_{x_i \in C_j} \left((x_i - \mu_j)^T (x_i - \mu_j) \right) = \sum_{j=1}^{k} \sum_{x_i \in C_j} \left(x_i^T x_i - x_i^T \mu_j - \mu_j^T x_i + \mu_j^T \mu_j \right) = \sum_{j=1}^{k} \sum_{x_i \in C_j} \left(x_i^T x_i - 2x_i^T \mu_j + \mu_j^T \mu_j \right)$$

$$J = \sum_{x_i \in C_1} \left(x_i^T x_i - 2x_i^T \mu_1 + \mu_1^T \mu_1 \right) + \sum_{x_i \in C_2} \left(x_i^T x_i - 2x_i^T \mu_2 + \mu_2^T \mu_2 \right) + \ldots + \sum_{x_i \in C_k} \left(x_i^T x_i - 2x_i^T \mu_k + \mu_k^T \mu_k \right)$$

K-MEANS CLUSTERING OBJECTIVE

- Want to minimize of the following objective function wrt means:

	X_1	X_2
x_1	4	1
x_2	1.1	-0.2
x_3	12	5.4
x_4	16.4	11.2
x_5	2.3	1.1
x_6	5	2
x_7	15	17.2
x_8	13.7	11.1
x_9	3.5	1.2

$$J = \sum_{j=1}^{k} \sum_{x_i \in C_j} ||x_i - \mu_j||_2^2$$

$$J = \sum_{j=1}^{k} \sum_{x_i \in C_j} \left((x_i - \mu_j)^T (x_i - \mu_j) \right) = \sum_{j=1}^{k} \sum_{x_i \in C_j} \left(x_i^T x_i - x_i^T \mu_j - \mu_j^T x_i + \mu_j^T \mu_j \right) = \sum_{j=1}^{k} \sum_{x_i \in C_j} \left(x_i^T x_i - 2x_i^T \mu_j + \mu_j^T \mu_j \right)$$

$$J = \sum_{x_i \in C_1} \left(x_i^T x_i - 2x_i^T \mu_1 + \mu_1^T \mu_1 \right) + \sum_{x_i \in C_2} \left(x_i^T x_i - 2x_i^T \mu_2 + \mu_2^T \mu_2 \right) + \ldots + \sum_{x_i \in C_k} \left(x_i^T x_i - 2x_i^T \mu_k + \mu_k^T \mu_k \right)$$

$$\frac{\delta J}{\delta \mu_j} = \frac{\delta}{\delta \mu_j} \left(\sum_{x_i \in C_j} \left(x_i^T x_i - 2x_i^T \mu_j + \mu_j^T \mu_j \right) \right) = \sum_{x_i \in C_j} \frac{\delta}{\delta \mu_j} (x_i^T x_i - 2x_i^T \mu_j + \mu_j^T \mu_j) = \sum_{x_i \in C_j} \left(\frac{\delta}{\delta \mu_j} \left(-2x_i^T \mu_j \right) + \frac{\delta}{\delta \mu_j} \left(\mu_j^T \mu_j \right) \right)$$

K-MEANS CLUSTERING OBJECTIVE

- Want to minimize of the following objective function wrt means:

	X1	X2
x1	4	1
x2	1.1	-0.2
x3	12	5.4
x4	16.4	11.2
x5	2.3	1.1
x6	5	2
x7	15	17.2
x8	13.7	11.1
x9	3.5	1.2

$$J = \sum_{j=1}^{k} \sum_{x_i \in C_j} ||x_i - \mu_j||_2^2$$

$$J = \sum_{j=1}^{k} \sum_{x_i \in C_j} \left((x_i - \mu_j)^T (x_i - \mu_j)\right) = \sum_{j=1}^{k} \sum_{x_i \in C_j} \left(x_i^T x_i - x_i^T \mu_j - \mu_j^T x_i + \mu_j^T \mu_j\right) = \sum_{j=1}^{k} \sum_{x_i \in C_j} \left(x_i^T x_i - 2x_i^T \mu_j + \mu_j^T \mu_j\right)$$

$$J = \sum_{x_i \in C_1} \left(x_i^T x_i - 2x_i^T \mu_1 + \mu_1^T \mu_1\right) + \sum_{x_i \in C_2} \left(x_i^T x_i - 2x_i^T \mu_2 + \mu_2^T \mu_2\right) + \ldots + \sum_{x_i \in C_k} \left(x_i^T x_i - 2x_i^T \mu_k + \mu_k^T \mu_k\right)$$

$$\frac{\delta J}{\delta \mu_j} = \frac{\delta}{\delta \mu_j} \left(\sum_{x_i \in C_j} \left(x_i^T x_i - 2x_i^T \mu_j + \mu_j^T \mu_j\right) \right) = \sum_{x_i \in C_j} \frac{\delta}{\delta \mu_j} (x_i^T x_i - 2x_i^T \mu_j + \mu_j^T \mu_j) = \sum_{x_i \in C_j} \left(\frac{\delta}{\delta \mu_j} \left(-2x_i^T \mu_j\right) + \frac{\delta}{\delta \mu_j} \left(\mu_j^T \mu_j\right) \right)$$

$$\frac{\delta J}{\delta \mu_j} = \sum_{x_i \in C_j} \left(-2x_i^T + 2\mu_j\right)$$

K-MEANS CLUSTERING OBJECTIVE

- Want to minimize of the following objective function wrt means:

	X_1	X_2
x_1	4	1
x_2	1.1	-0.2
x_3	12	5.4
x_4	16.4	11.2
x_5	2.3	1.1
x_6	5	2
x_7	15	17.2
x_8	13.7	11.1
x_9	3.5	1.2

$$J = \sum_{j=1}^{k} \sum_{x_i \in C_j} ||x_i - \mu_j||_2^2$$

$$J = \sum_{j=1}^{k} \sum_{x_i \in C_j} \left((x_i - \mu_j)^T (x_i - \mu_j)\right) = \sum_{j=1}^{k} \sum_{x_i \in C_j} \left(x_i^T x_i - x_i^T \mu_j - \mu_j^T x_i + \mu_j^T \mu_j\right) = \sum_{j=1}^{k} \sum_{x_i \in C_j} \left(x_i^T x_i - 2x_i^T \mu_j + \mu_j^T \mu_j\right)$$

$$J = \sum_{x_i \in C_1} \left(x_i^T x_i - 2x_i^T \mu_1 + \mu_1^T \mu_1\right) + \sum_{x_i \in C_2} \left(x_i^T x_i - 2x_i^T \mu_2 + \mu_2^T \mu_2\right) + \ldots + \sum_{x_i \in C_k} \left(x_i^T x_i - 2x_i^T \mu_k + \mu_k^T \mu_k\right)$$

$$\frac{\delta J}{\delta \mu_j} = \frac{\delta}{\delta \mu_j} \left(\sum_{x_i \in C_j} \left(x_i^T x_i - 2x_i^T \mu_j + \mu_j^T \mu_j\right) \right) = \sum_{x_i \in C_j} \frac{\delta}{\delta \mu_j} (x_i^T x_i - 2x_i^T \mu_j + \mu_j^T \mu_j) = \sum_{x_i \in C_j} \left(\frac{\delta}{\delta \mu_j} \left(-2x_i^T \mu_j\right) + \frac{\delta}{\delta \mu_j} \left(\mu_j^T \mu_j\right) \right)$$

$$\frac{\delta J}{\delta \mu_j} = \sum_{x_i \in C_j} \left(-2x_i^T + 2\mu_j\right) = 0 \quad \Rightarrow \sum_{x_i \in C_j} 2\mu_j = \sum_{x_i \in C_j} 2x_i^T \quad \Rightarrow \quad |C_j|\, \mu_j = \sum_{x_i \in C_j} x_i^T \quad \Rightarrow \mu_j = \frac{\sum_{x_i \in C_j} x_i^T}{|C_j|}$$

K-MEANS CLUSTERING OBJECTIVE

- Want to minimize of the following objective function wrt to cluster assignments:

	X_1	X_2
x_1	4	1
x_2	1.1	-0.2
x_3	12	5.4
x_4	16.4	11.2
x_5	2.3	1.1
x_6	5	2
x_7	15	17.2
x_8	13.7	11.1
x_9	3.5	1.2

$$J = \sum_{j=1}^{k} \sum_{x_i \in C_j} ||x_i - \mu_j||_2^2$$

$$J = \sum_{j=1}^{k} \sum_{x_i \in C_j} \left((x_i - \mu_j)^T (x_i - \mu_j) \right) = \sum_{j=1}^{k} \sum_{x_i \in C_j} \left(x_i^T x_i - x_i^T \mu_j - \mu_j^T x_i + \mu_j^T \mu_j \right) = \sum_{j=1}^{k} \sum_{x_i \in C_j} \left(x_i^T x_i - 2 x_i^T \mu_j + \mu_j^T \mu_j \right)$$

What cluster assignments will minimize J?

For a particular x_i, which assignment will minimize J?