K-means

CSCI 347

Adiesha Liyana Ralalage

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:
  - Greedy algorithm that minimizes the squared error of points to their respective cluster means.

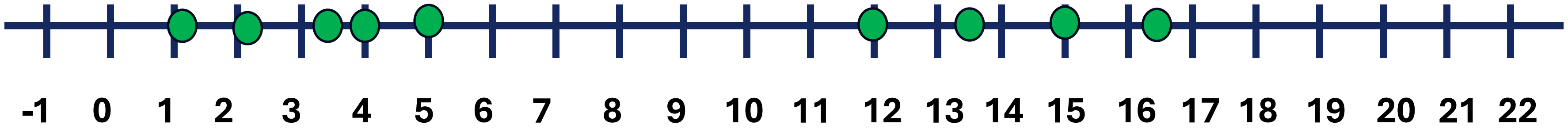
$$J = \sum_{i=1}^{k} \sum_{x \in C_j} \left\| x_i - \mu_j \right\|_2^2$$

  - Hard clustering method: each point is only assigned to one cluster.

K-means clustering example

- 1-dimensional example with $k = 2$

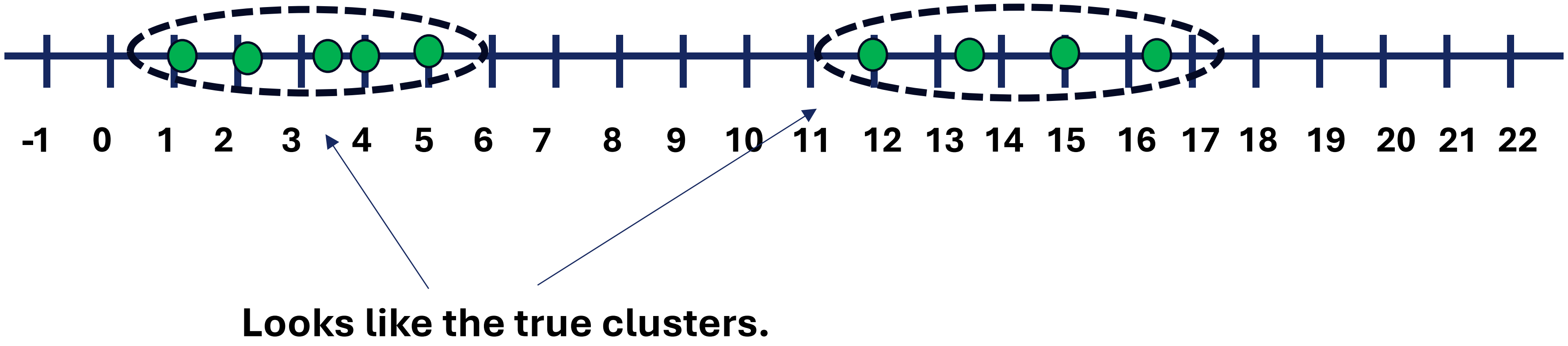
	X_1
x_1	4
x_2	1.1
x_3	12
x_4	16.4
x_5	2.3
x_6	5
x_7	15
x_8	13.7
x_9	3.5



K-means clustering example

- 1-dimensional example with $k = 2$

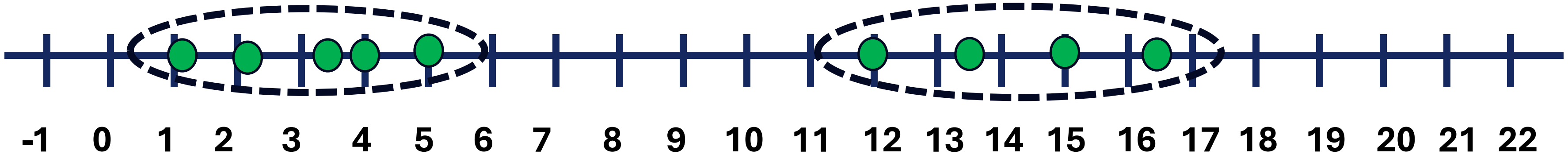
	X_1
x_1	4
x_2	1.1
x_3	12
x_4	16.4
x_5	2.3
x_6	5
x_7	15
x_8	13.7
x_9	3.5



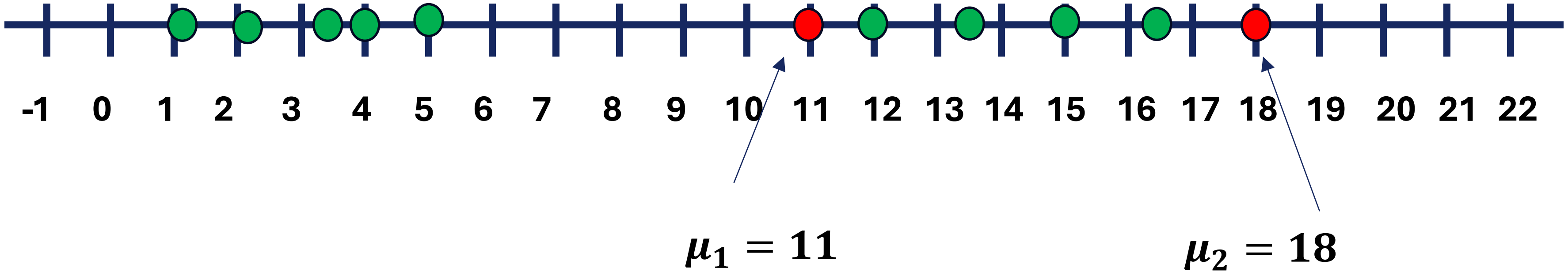
K-means clustering example

- 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



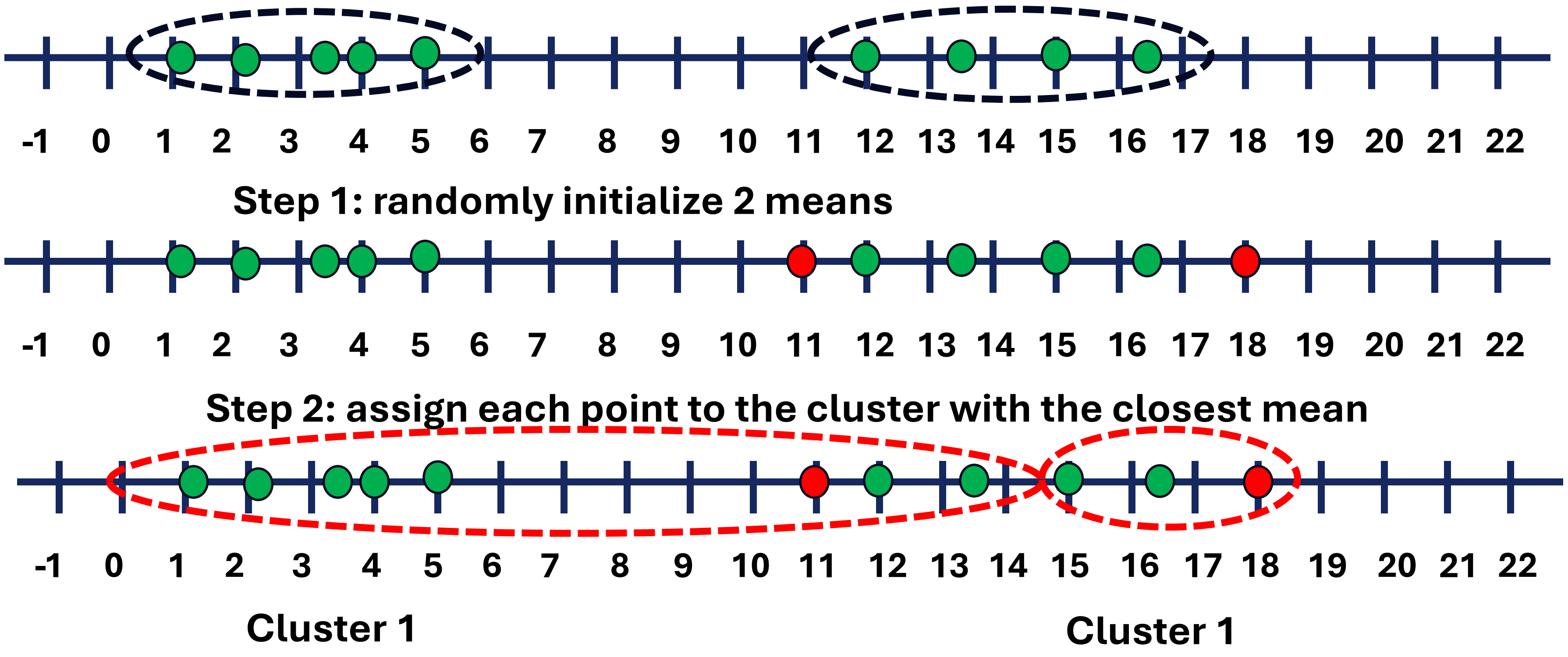
Step 1: randomly initialize 2 means



K-means clustering example

- 1-dimensional example with $k = 2$

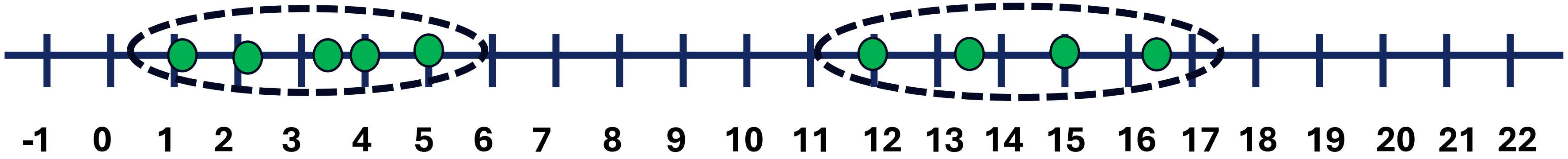
	X_1
x_1	4
x_2	1.1
x_3	12
x_4	16.4
x_5	2.3
x_6	5
x_7	15
x_8	13.7
x_9	3.5



K-means clustering example

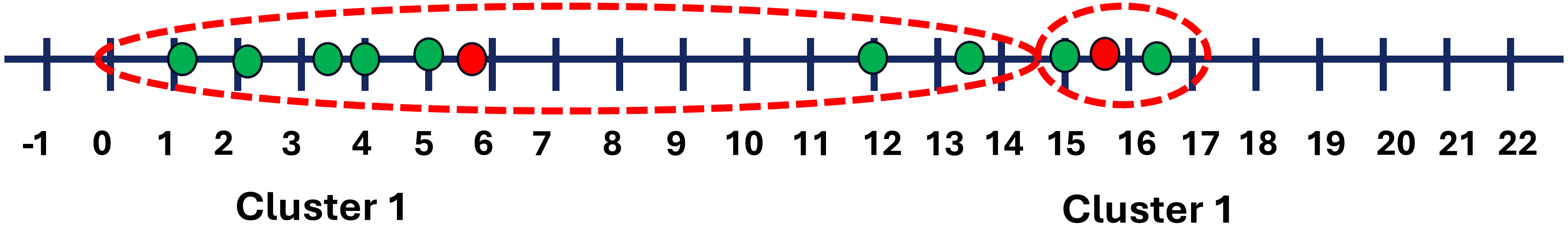
- 1-dimensional example with $k = 2$

	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: recompute the means based on the cluster membership

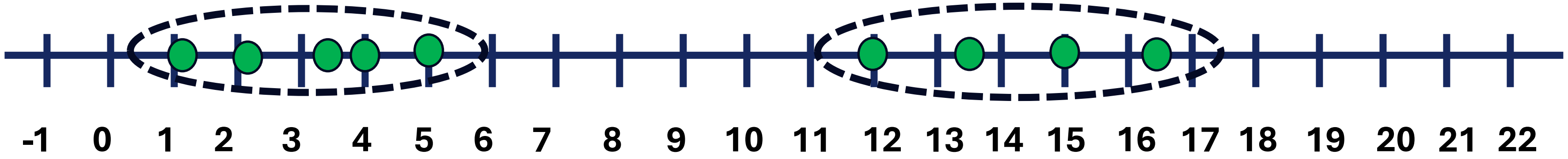
$$\boldsymbol{\mu_1} = \frac{x_1+x_2+x_3+x_5+x_6+x_8+x_9}{8} = \frac{41.6}{8} = 5.94 \qquad \boldsymbol{\mu_2} = \frac{x_4+x_7}{2} = \frac{31.4}{2} = 15.7$$



K-means clustering example

- 1-dimensional example with $k = 2$

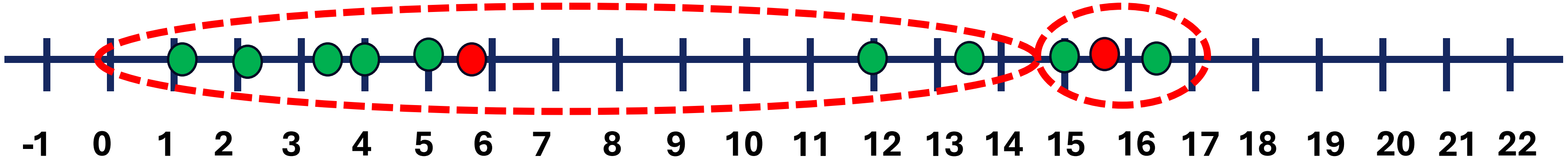
	X_1
x_1	4
x_2	1.1
x_3	12
x_4	16.4
x_5	2.3
x_6	5
x_7	15
x_8	13.7
x_9	3.5



Step 3: recompute the means based on the cluster membership

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

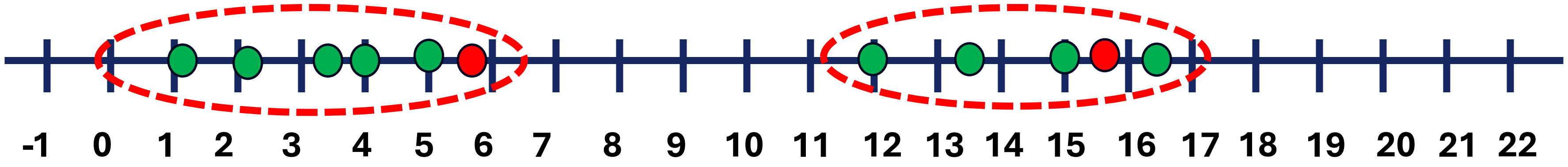
$$\boldsymbol{\mu_2} = \frac{x_4+x_7}{2} = \frac{31.4}{2} = 15.7$$



Cluster 1

Cluster 1

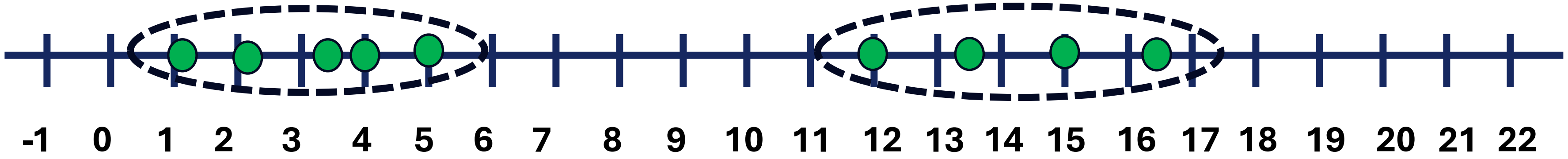
Step 2: assign each point to the cluster with the closest mean



K-means clustering example

- 1-dimensional example with $k = 2$

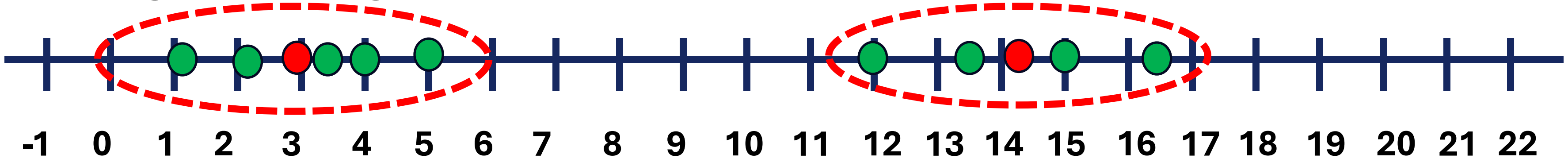
	X_1
x_1	4
x_2	1.1
x_3	12
x_4	16.4
x_5	2.3
x_6	5
x_7	15
x_8	13.7
x_9	3.5



Step 3: recompute the means based on the cluster membership

$$\boldsymbol{\mu_1} = \frac{x_1+x_2+x_5+x_6+x_9}{5} = \frac{15.9}{5} = \mathbf{3.18}$$

$$\boldsymbol{\mu_2} = \frac{x_3+x_4+x_7+x_8}{4} = \frac{57.1}{4} = \mathbf{14.3}$$

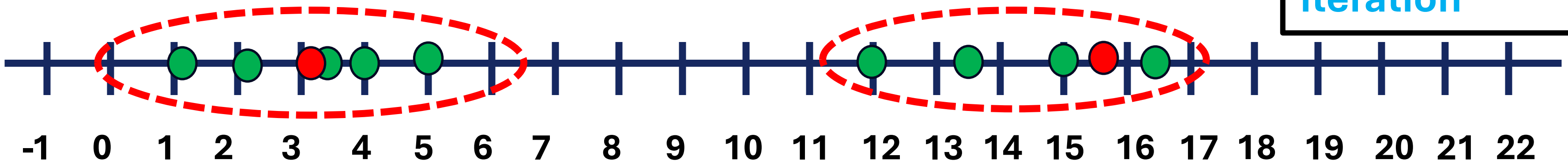


Cluster 1

Cluster 1

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

No change: stop the iteration



K-means algorithm

Algorithm 13.1: K-means Algorithm

$\textbf{K-}\textsc{means}$ $(\mathbf{D}, k, \epsilon)$:

$t = 0$

Randomly initialize k centroids: $\boldsymbol{\mu}_1^t, \boldsymbol{\mu}_2^t, \ldots, \boldsymbol{\mu}_k^t \in \mathbb{R}^d$

repeat

  $t \leftarrow t + 1$

  `// Cluster Assignment Step`

  foreach $\mathbf{x}_j \in \mathbf{D}$ **do**

    $j^* \leftarrow \arg\min_i \left\{ \left\| \mathbf{x}_j - \boldsymbol{\mu}_i^t \right\|^2 \right\}$ `// Assign` $\mathbf{x}_j$ `to closest centroid`

    $C_{j^*} \leftarrow C_{j^*} \cup \{\mathbf{x}_j\}$

  `// Centroid Update Step`

  foreach $i = 1$ *to* k **do**

    $\boldsymbol{\mu}_i^t \leftarrow \frac{1}{|C_i|} \sum_{\mathbf{x}_j \in C_i} \mathbf{x}_j$

until $\sum_{i=1}^{k} \left\| \boldsymbol{\mu}_i^t - \boldsymbol{\mu}_i^{t-1} \right\| \leq \epsilon$

K-means

- Given n points and a number d, is there a partition into k clusters such that the sum of squared distances between each point and the centroid of its cluster is at most d?
  - This problem is NP-hard
  - **NP-hardness of Euclidean sum-of-squares clustering**
  - https://link.springer.com/article/10.1007/s10994-009-5103-0

- K-means algorithm finds a local optimum.

K-means objective

- We want to minimize the following objective function w.r.t the means:

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

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

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

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

Taking the partial derivative with respect to a specific centroid μ_j:

$$\frac{\delta J}{\delta \mu_j} = \sum_{x \in C_j} (-2x_i + 2\mu_j)$$

K-means objective

- We want to minimize the following objective function wrt to means:

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

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

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

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

$$\frac{\delta J}{\delta \mu_j} = \sum_{x \in C_j} \left(-2 x_i + 2 \mu_j \right) = 0 \to \sum_{x \in C_j} 2 \mu_j = \sum_{x \in C_j} 2 x_i \to \left| x_i \in C_j \right| \mu_j = \sum_{x \in C_j} x_i \to \mu_j = \frac{\sum_{x \in C_j} x_i}{\left| x \in C_j \right|}$$

K-means objective

- $\mu_j = \frac{\sum_{x_i \in C_j} x_i}{|x \in C_j|}$
- Each centroid should be the average of the points in its cluster.