K-means clustering

Algorithm

Herman Kamper

http://www.kamperh.com/

K-means clustering algorithm and example

1. Randomly assign each item $\mathbf{x}^{(n)}$ to one of the K clusters.

2. repeat until cluster assignments stop changing:

   (a) for cluster $k = 1$ to K:

   Calculate the cluster centroid $\boldsymbol{\mu}_k$ as the mean of all the items assigned to cluster k.

   (b) for item $n = 1$ to N:

   Assign item $\mathbf{x}^{(n)}$ to the cluster with the closest centroid.

K-means clustering algorithm and example

Example: K=3

1. Randomly assign each item $\mathbf{x}^{(n)}$ to one of the K clusters.

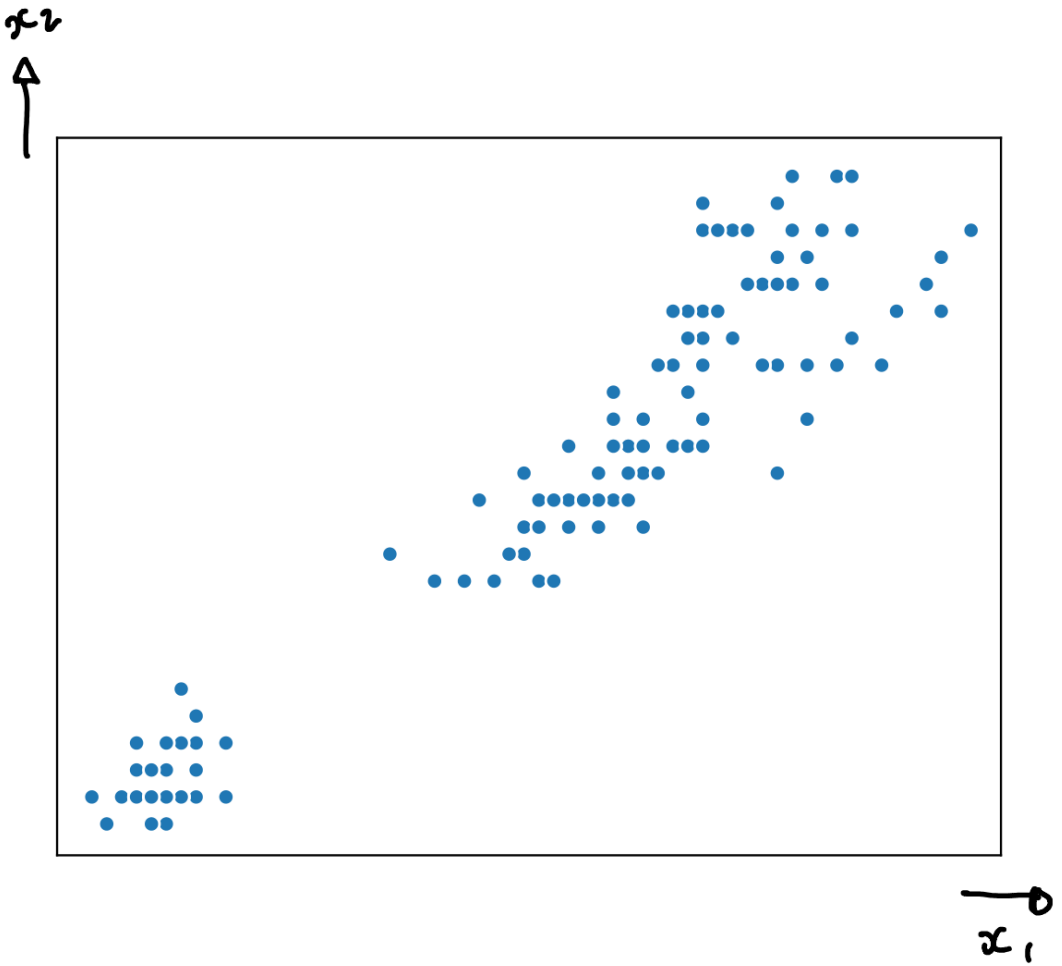
2. repeat until cluster assignments stop changing:

   (a) for cluster $k = 1$ to K:

   Calculate the cluster centroid $\boldsymbol{\mu}_k$ as the mean of all the items assigned to cluster k.

   (b) for item $n = 1$ to N:

   Assign item $\mathbf{x}^{(n)}$ to the cluster with the closest centroid.



K-means clustering algorithm and example

1. Randomly assign each item $\mathbf{x}^{(n)}$ to one of the K clusters.

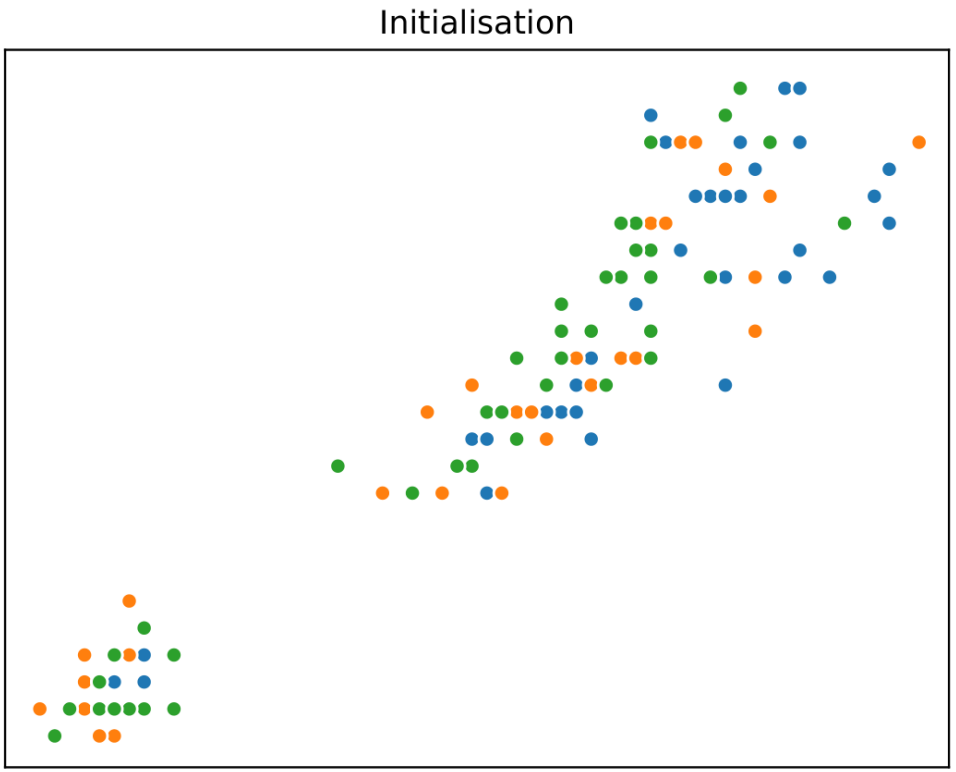
2. repeat until cluster assignments stop changing:

   (a) for cluster $k = 1$ to K:

   Calculate the cluster centroid $\boldsymbol{\mu}_k$ as the mean of all the items assigned to cluster k.

   (b) for item $n = 1$ to N:

   Assign item $\mathbf{x}^{(n)}$ to the cluster with the closest centroid.



K-means clustering algorithm and example

1. Randomly assign each item $\mathbf{x}^{(n)}$ to one of the K clusters.

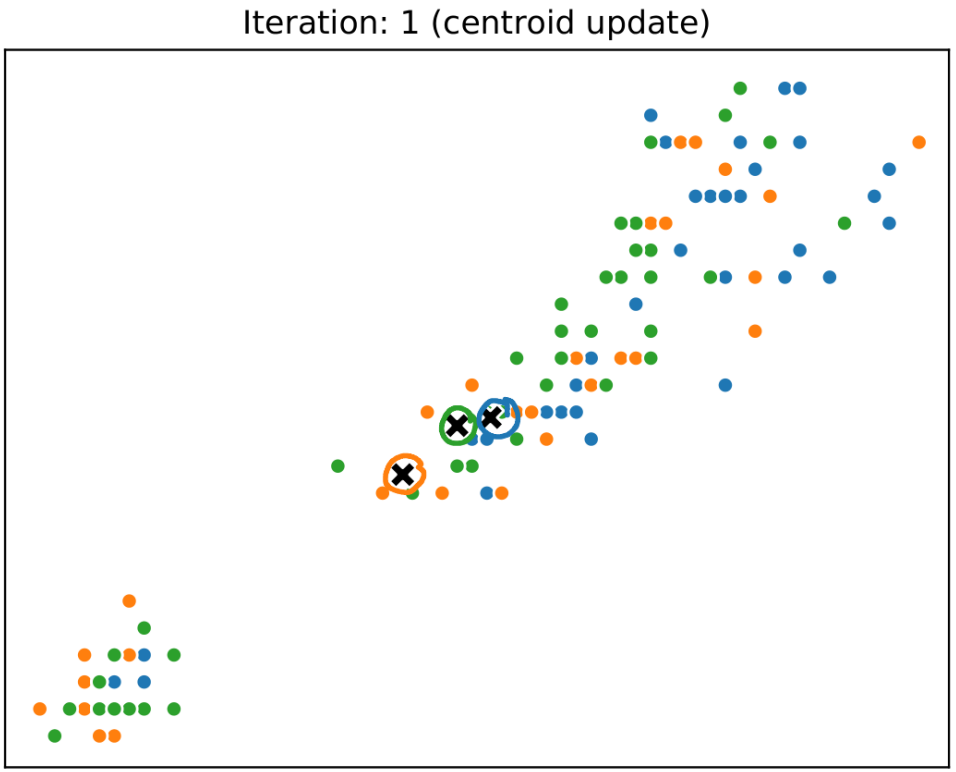
2. repeat until cluster assignments stop changing:

   (a) for cluster $k = 1$ to K:

   Calculate the cluster centroid $\boldsymbol{\mu}_k$ as the mean of all the items assigned to cluster k.

   (b) for item $n = 1$ to N:

   Assign item $\mathbf{x}^{(n)}$ to the cluster with the closest centroid.



K-means clustering algorithm and example

1. Randomly assign each item $\mathbf{x}^{(n)}$ to one of the K clusters.

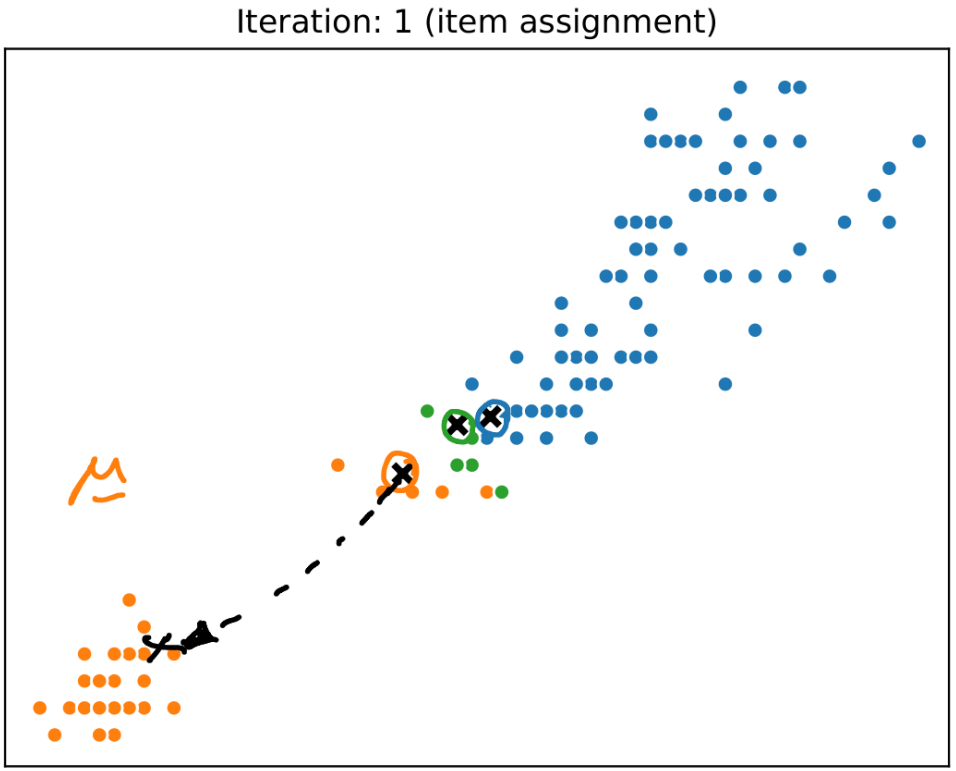
2. repeat until cluster assignments stop changing:

   (a) for cluster $k = 1$ to K:

   Calculate the cluster centroid $\boldsymbol{\mu}_k$ as the mean of all the items assigned to cluster k.

   (b) for item $n = 1$ to N:

   Assign item $\mathbf{x}^{(n)}$ to the cluster with the closest centroid.



K-means clustering algorithm and example

1. Randomly assign each item $\mathbf{x}^{(n)}$ to one of the K clusters.

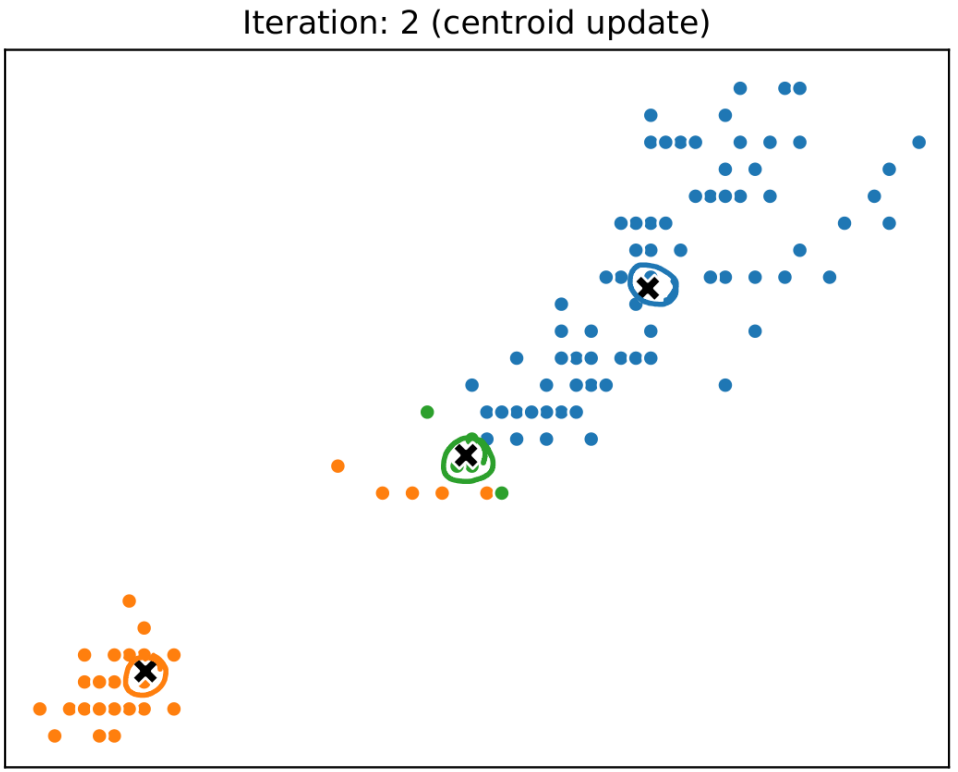
2. repeat until cluster assignments stop changing:

   (a) for cluster $k = 1$ to K:

   Calculate the cluster centroid $\boldsymbol{\mu}_k$ as the mean of all the items assigned to cluster k.

   (b) for item $n = 1$ to N:

   Assign item $\mathbf{x}^{(n)}$ to the cluster with the closest centroid.



K-means clustering algorithm and example

1. Randomly assign each item $\mathbf{x}^{(n)}$ to one of the K clusters.

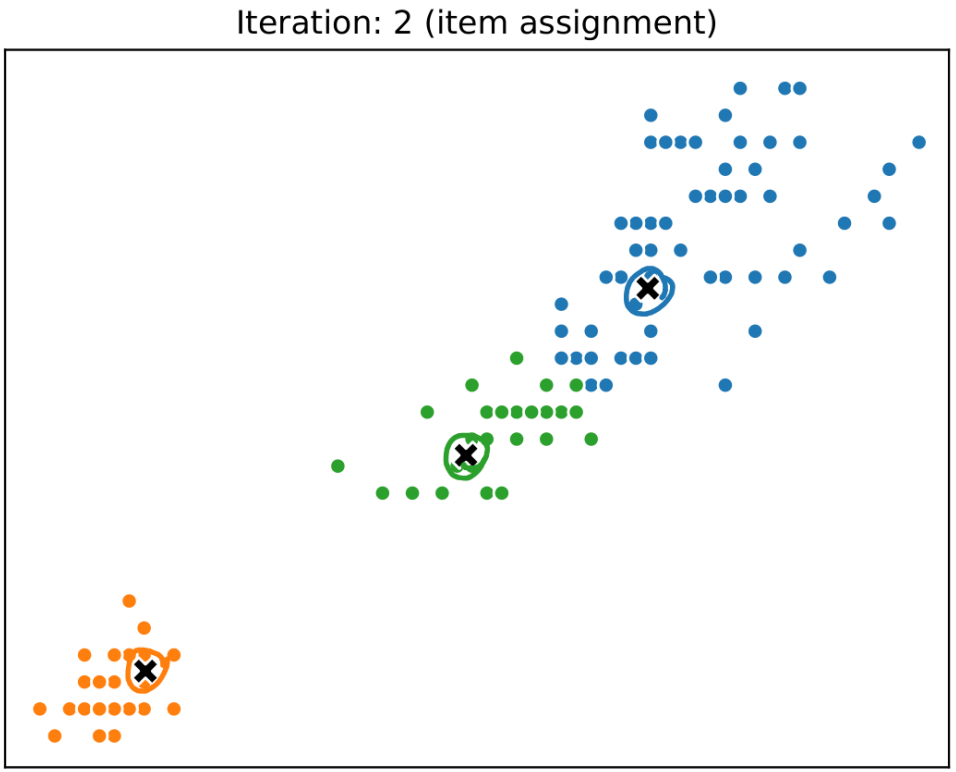
2. repeat until cluster assignments stop changing:

   (a) for cluster $k = 1$ to K:

   Calculate the cluster centroid $\boldsymbol{\mu}_k$ as the mean of all the items assigned to cluster k.

   (b) for item $n = 1$ to N:

   Assign item $\mathbf{x}^{(n)}$ to the cluster with the closest centroid.



K-means clustering algorithm and example

1. Randomly assign each item $\mathbf{x}^{(n)}$ to one of the K clusters.

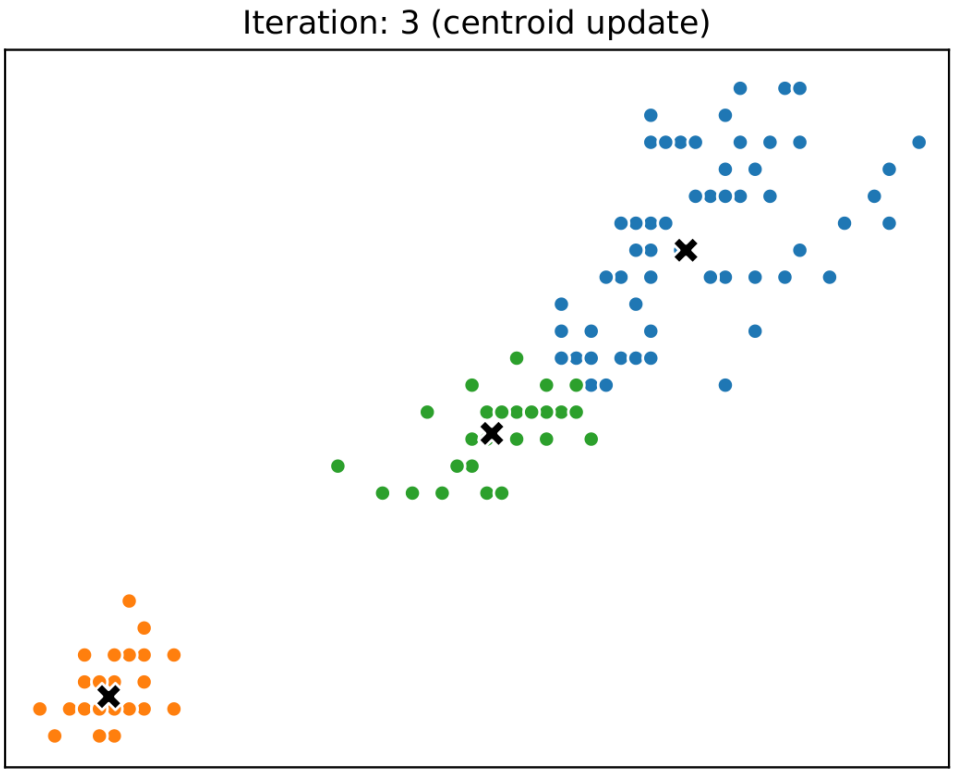
2. repeat until cluster assignments stop changing:

   (a) for cluster $k = 1$ to K:

   Calculate the cluster centroid $\boldsymbol{\mu}_k$ as the mean of all the items assigned to cluster k.

   (b) for item $n = 1$ to N:

   Assign item $\mathbf{x}^{(n)}$ to the cluster with the closest centroid.



K-means clustering algorithm and example

1. Randomly assign each item $\mathbf{x}^{(n)}$ to one of the K clusters.

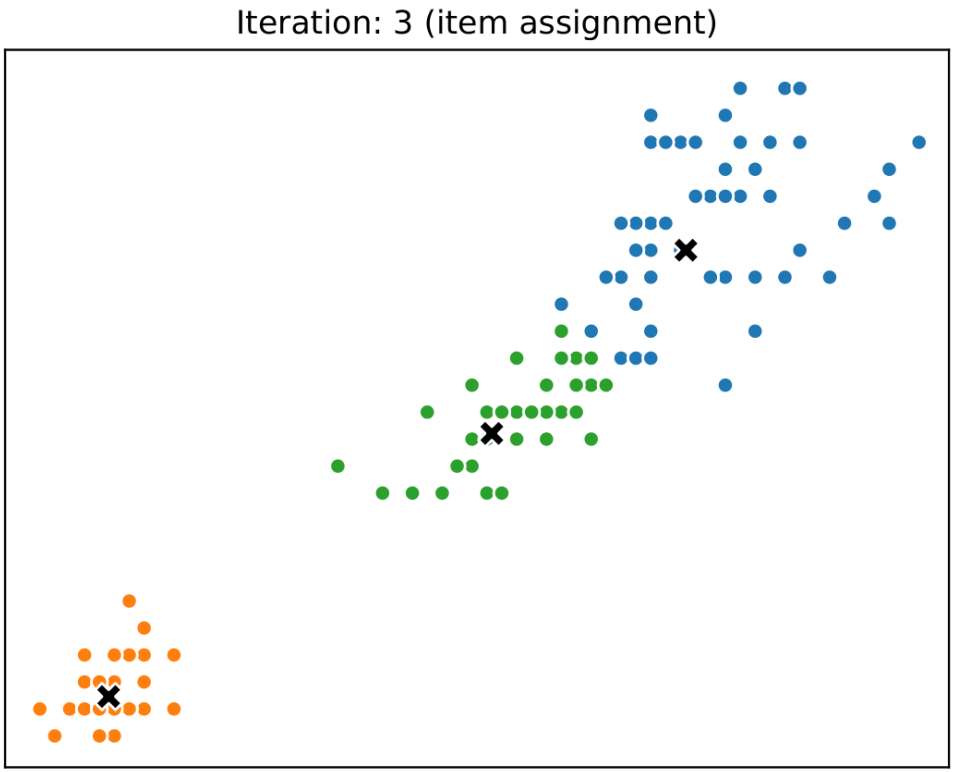
2. repeat until cluster assignments stop changing:

   (a) for cluster $k = 1$ to K:

   Calculate the cluster centroid $\boldsymbol{\mu}_k$ as the mean of all the items assigned to cluster k.

   (b) for item $n = 1$ to N:

   Assign item $\mathbf{x}^{(n)}$ to the cluster with the closest centroid.



K-means clustering

Details and loss

Herman Kamper

http://www.kamperh.com/

K-means clustering algorithm details

C_k: Set of indices of items assigned to cluster k

e.g. $C_4 = \{205, 12, 303\}$; $|C_4| = 3$

$x^{(205)}$

$|C_k|$: # items in cluster k

1. Randomly assign each item $\mathbf{x}^{(n)}$ to one of the K clusters.

2. repeat until cluster assignments stop changing:

   (a) for cluster $k = 1$ to K:

   Calculate the cluster centroid $\boldsymbol{\mu}_k$ as the mean of all the items assigned to cluster k.

$$\underline{\mu}_k = \frac{1}{|C_k|} \sum_{i \in C_k} \underline{x}^{(i)}$$

   (b) for item $n = 1$ to N:

   Assign item $\mathbf{x}^{(n)}$ to the cluster with the closest centroid.

$$\arg\min_k \| \underline{x}^{(n)} - \underline{\mu}_k \|^2$$

K-means clustering algorithm details

Loss: $J(c_1, \ldots, c_K, \underline{\mu}_1, \ldots \underline{\mu}_K) = \sum_{k=1}^{K} \sum_{i \in c_k} \| \underline{x}^{(i)} - \underline{\mu}_k \|^2$

1. Randomly assign each item $\mathbf{x}^{(n)}$ to one of the K clusters.

2. repeat until cluster assignments stop changing:

   (a) for cluster $k = 1$ to K:

   Calculate the cluster centroid $\boldsymbol{\mu}_k$ as the mean of all the items assigned to cluster k.

   (b) for item $n = 1$ to N:

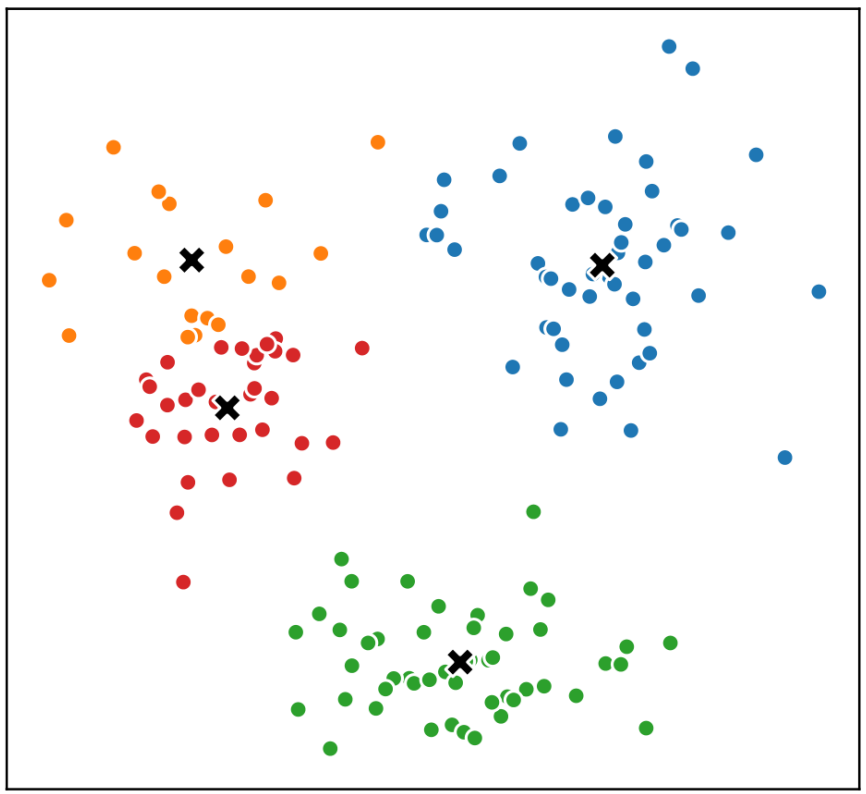
   Assign item $\mathbf{x}^{(n)}$ to the cluster with the closest centroid.

Centroid update:
Update $\underline{\mu}_1, \ldots, \underline{\mu}_K$ while keeping $c_1, \ldots, c_K$

Cluster assignments:
Update $c_1, \ldots, c_K$ while keeping $\underline{\mu}_1, \ldots \underline{\mu}_K$ fixed

Random initialisation leads to different local optima

Sum of squared distances to centroids: 68.26



Sum of squared distances to centroids: 66.97