K-means clustering

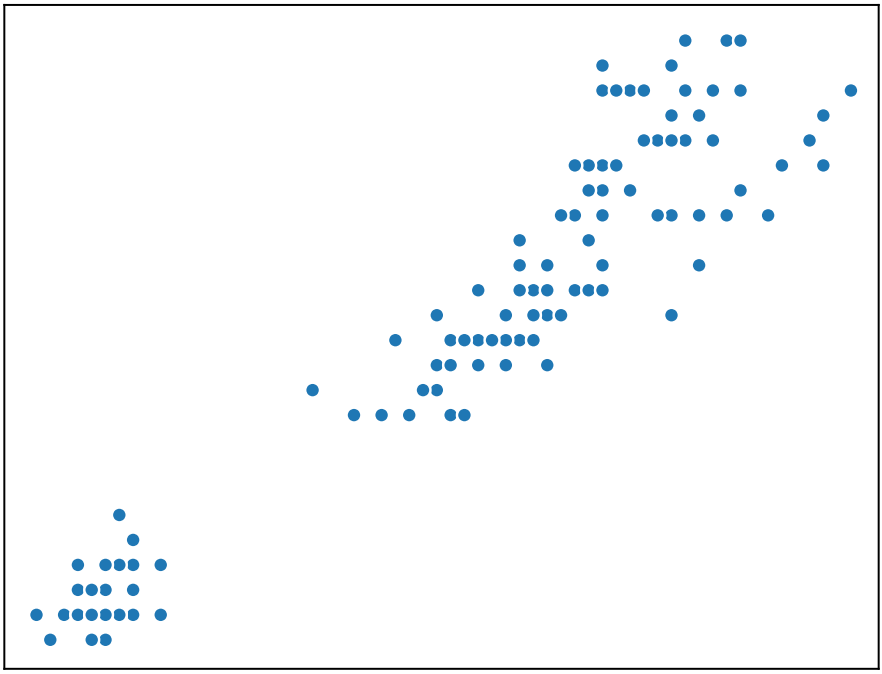
Herman Kamper

2023-03

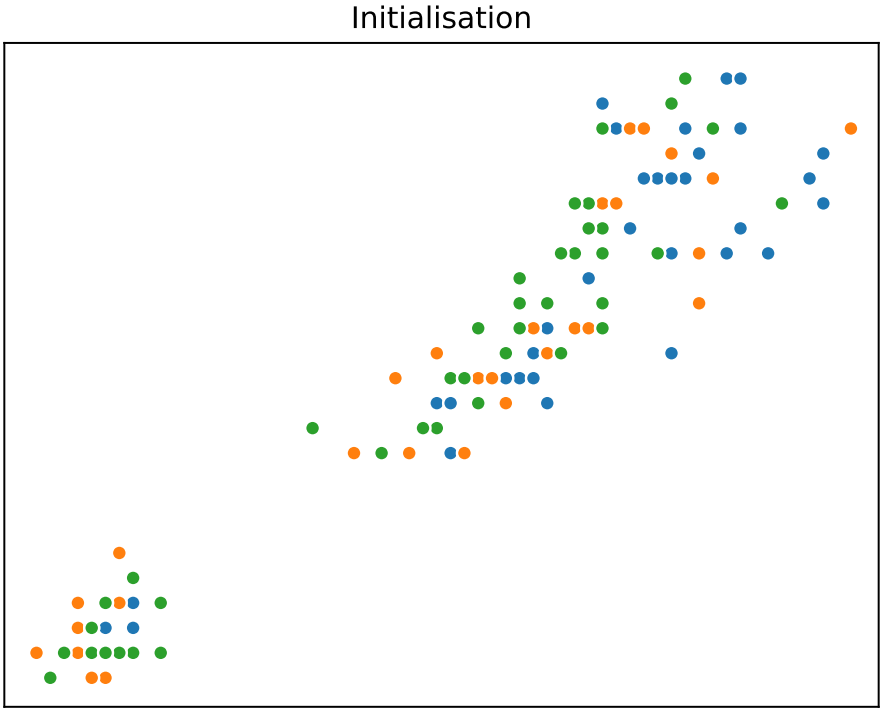
K-means clustering algorithm

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

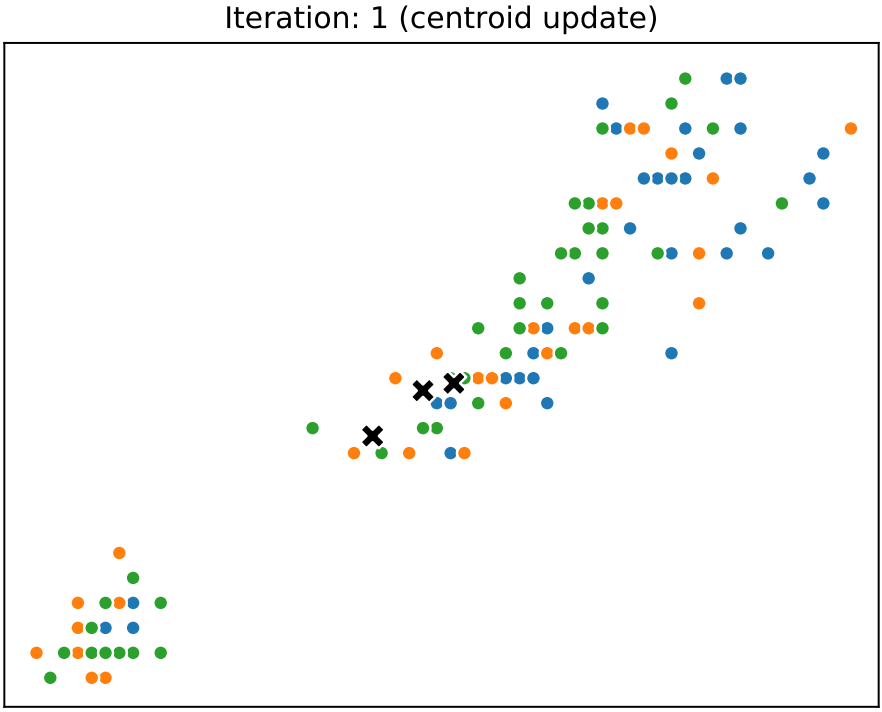
K-means example



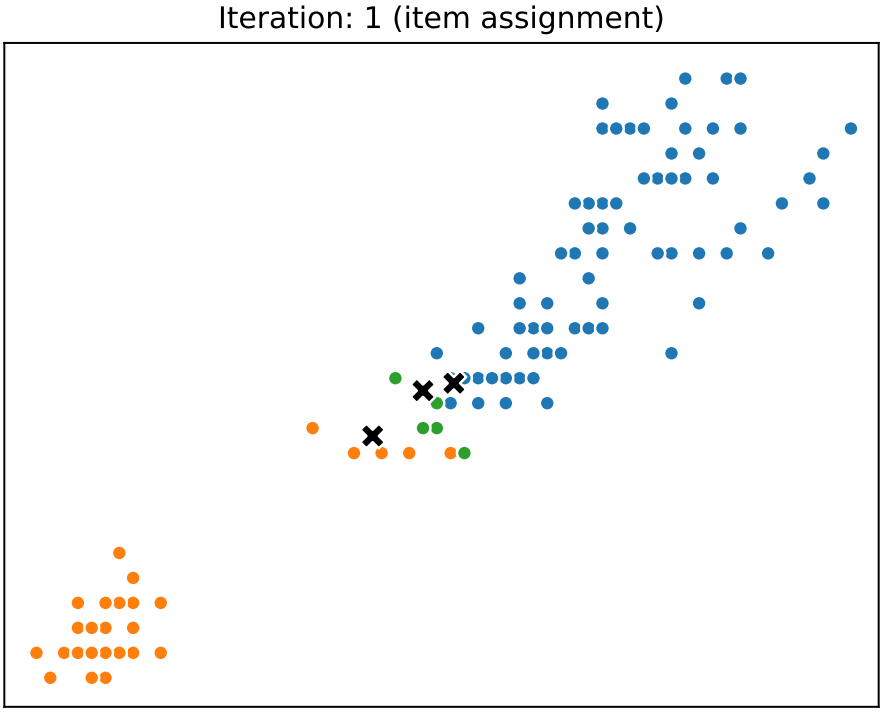
K-means example



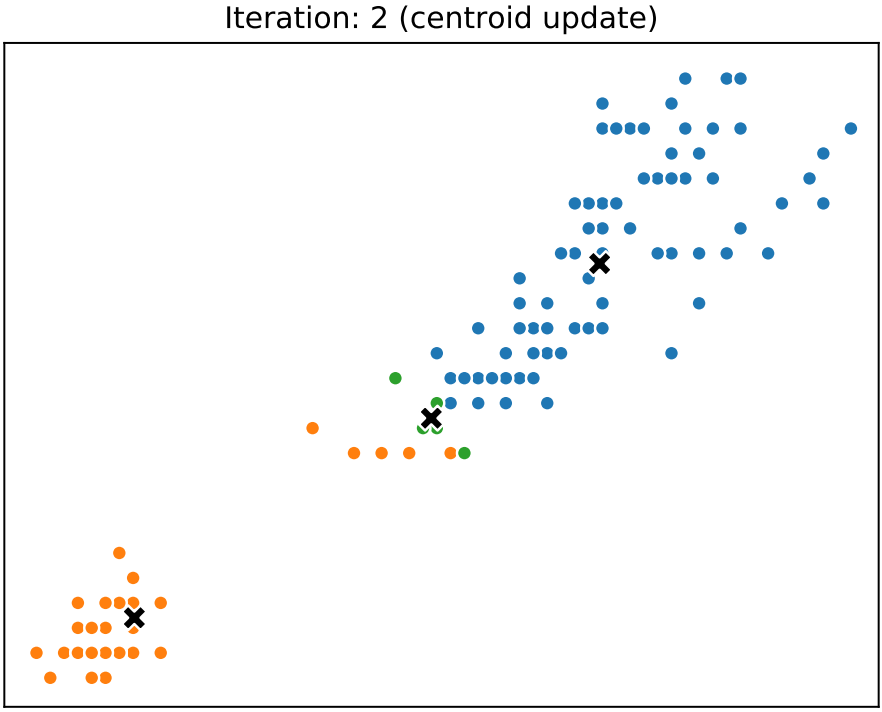
K-means example



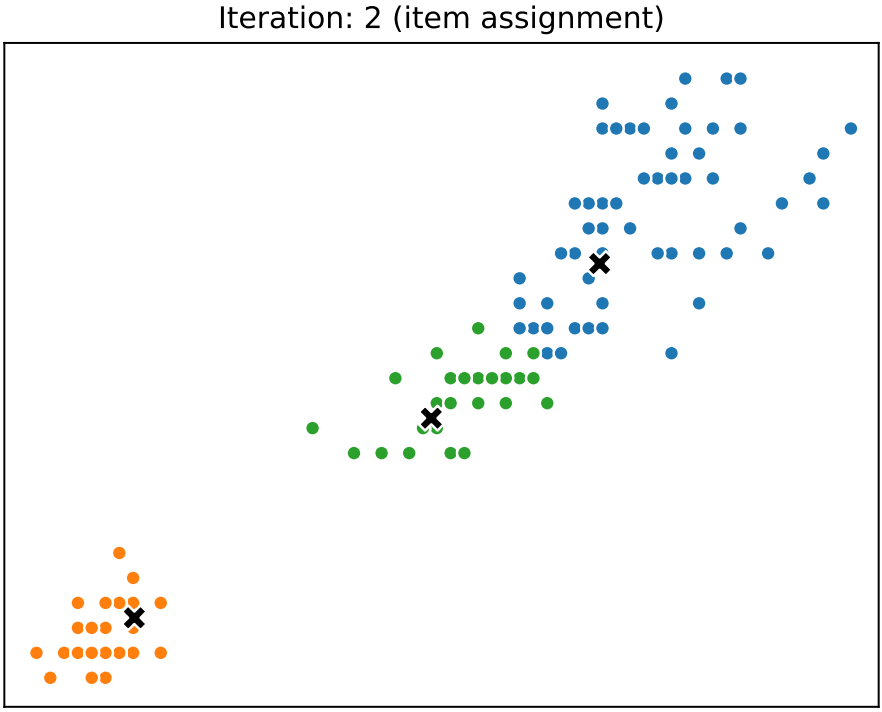
K-means example



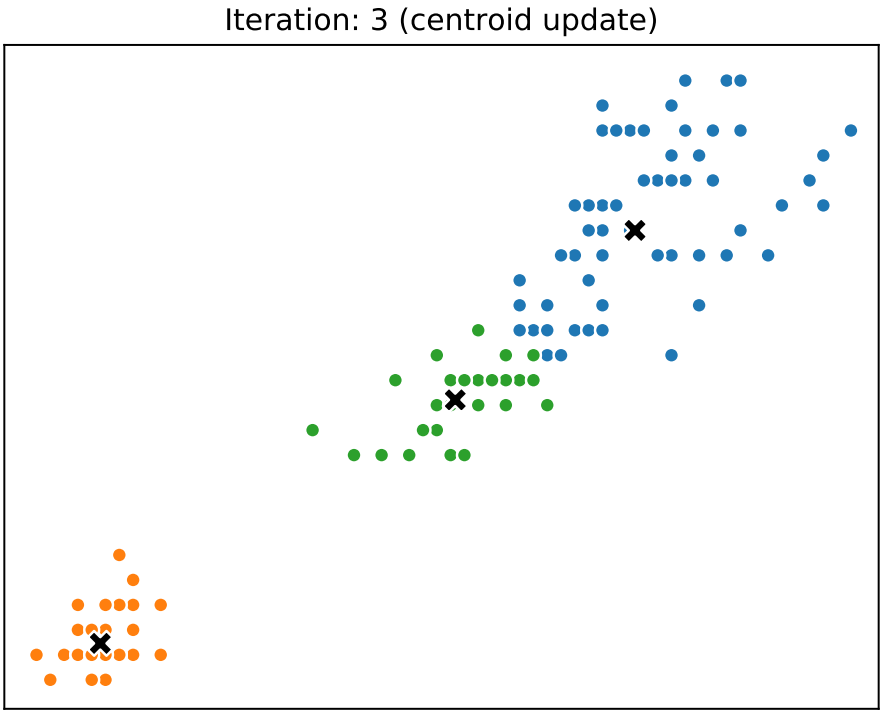
K-means example



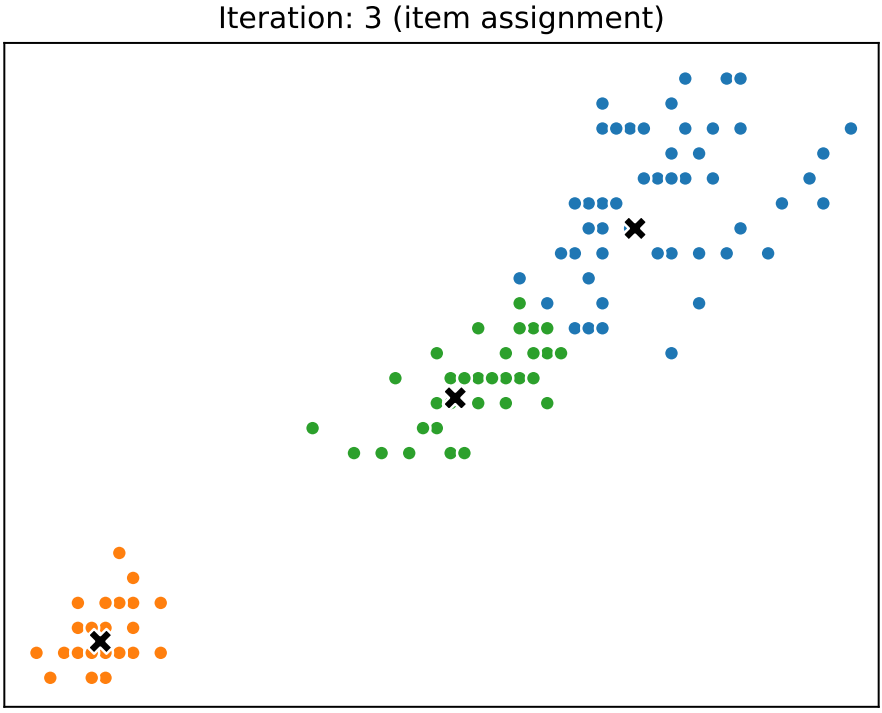
K-means example



K-means example



K-means example



K-means clustering algorithm details

Notation

C_k denotes the set of indices of items assigned to cluster k.

$|C_k|$ denotes the number of items in cluster k.

Example: $C_4 = \{205, 12, 303\}$, $|C_4| = 3$

Inner loop

(a) Centroid update:

Update the centroids $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K$ while keeping the cluster assignments $C_1, C_2, \ldots, C_K$ fixed.

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

(b) Cluster assignment update:

Update the cluster assignments $C_1, C_2, \ldots, C_K$ while keeping the centroids $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K$ fixed.

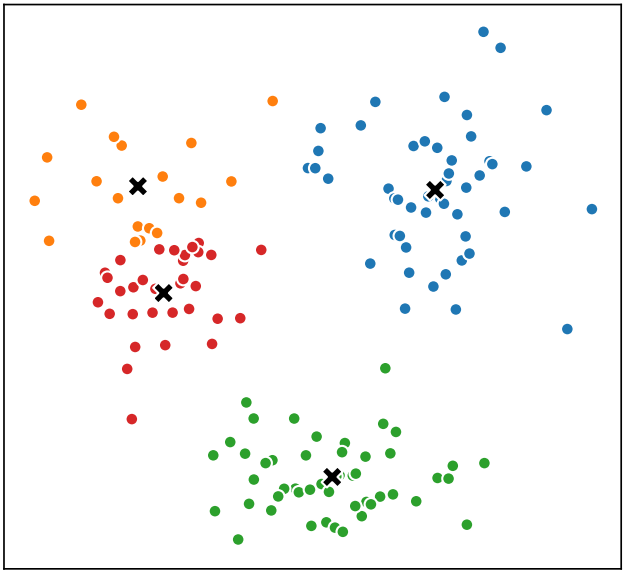
$$\arg\min_k \left\| \mathbf{x}^{(n)} - \boldsymbol{\mu}_k \right\|^2$$

Loss

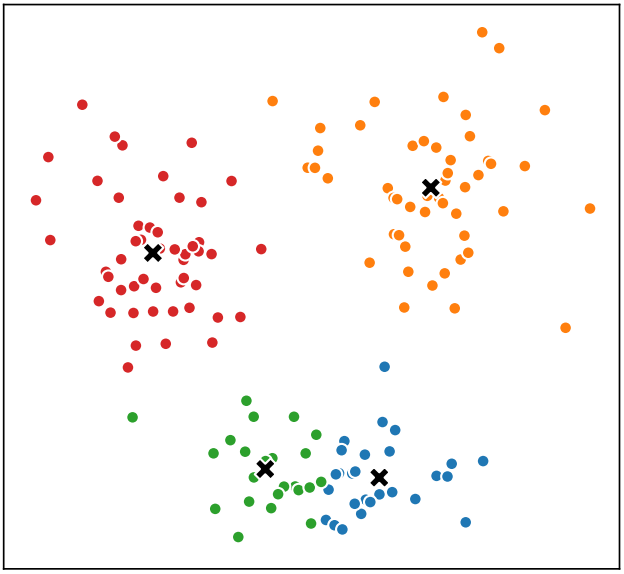
$$J(C_1, C_2, \ldots, C_K, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K) = \sum_{k=1}^{K} \sum_{i \in C_k} \left\| \mathbf{x}^{(i)} - \boldsymbol{\mu}_k \right\|^2$$

Effect of random initialisation

Sum of squared distances to centroids: 68.26



Sum of squared distances to centroids: 66.97



Videos covered in this note

- K-means clustering 1 - Algorithm (16 min)
- K-means clustering 2 - Details (14 min)

Reading

- ISLR 12.4.1
- ISLR 12.4.3 - Only the content regarding K-means clustering is examinable (not hierarchical clustering).