Welcome to Advanced Data Analysis (PHYS 605)

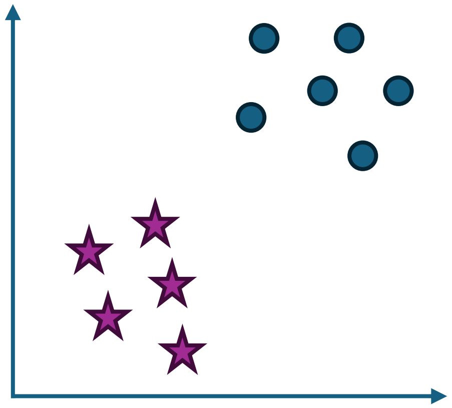
Prof. Claudia Gomes da Rocha

claudia.gomesdarocha@ucalgary.ca

Department of Physics and Astronomy
Faculty of Science, University of Calgary

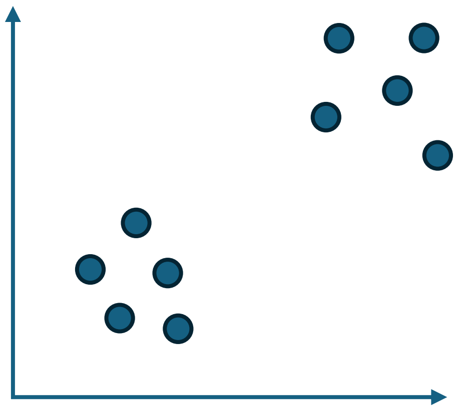
UNIVERSITY OF CALGARY

Supervised learning



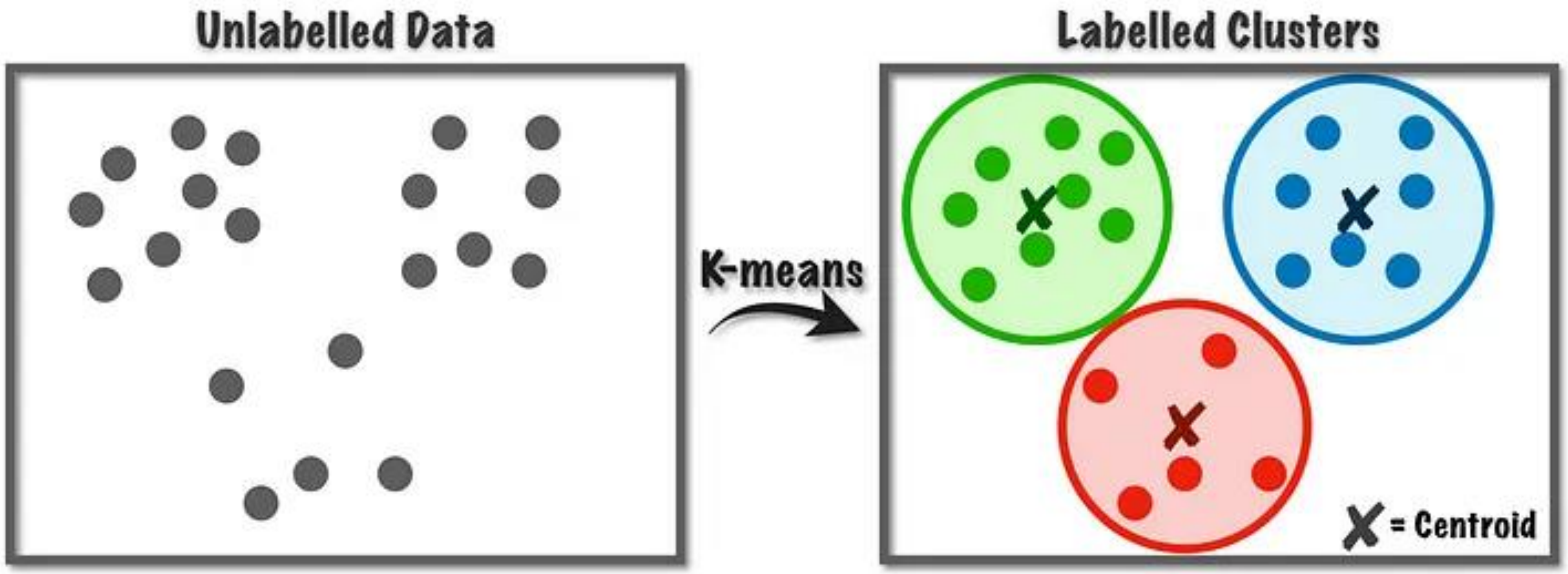
Labelled data

Unsupervised learning



Unlabelled data

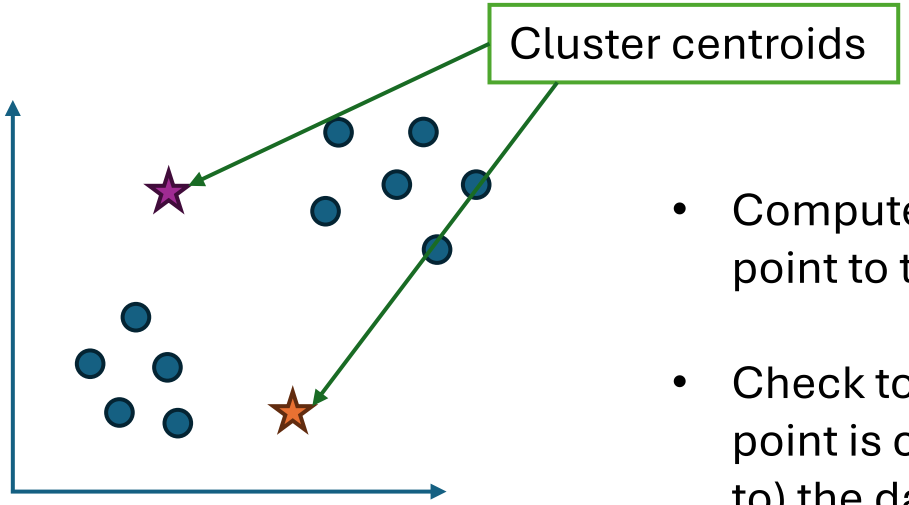
K-means clustering



https://towardsdatascience.com/k-means-a-complete-introduction-1702af9cd8c

K-means algorithm (illustration)

- Decide the number of centroids (equivalent to the number of clusters one wishes to identify, K);
- Randomly pick the coordinates of the position of the centroids;

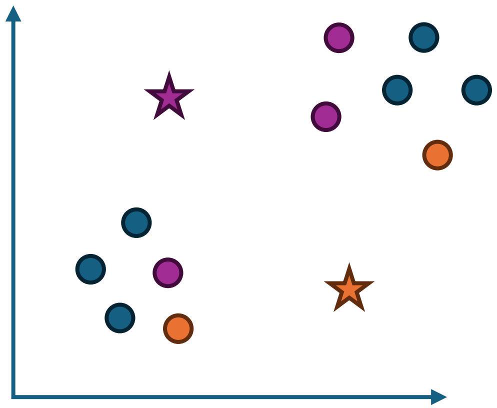


- Compute the distance between each data point to the cluster centroids;
- Check to which centroid cluster a given data point is closer. Paint (or assign a cluster label to) the data point in the same colour as the colour of the cluster centroid.

K-means algorithm (illustration)

- Decide the number of centroids (equivalent to the number of clusters one wishes to identify, K);

- Randomly pick the coordinates of the position of the centroids;



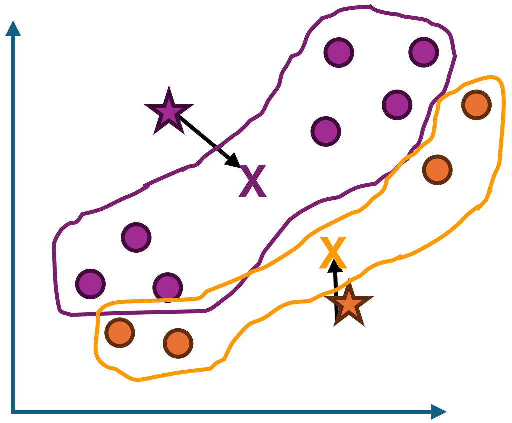
- Compute the distance between each data point to the cluster centroids;

- Check to which centroid cluster a given data point is closer. Paint (or assign a cluster label to) the data point in the same colour as the colour of the cluster centroid.

K-means algorithm (illustration)

- Decide the number of centroids (equivalent to the number of clusters one wishes to identify, K);

- Randomly pick the coordinates of the position of the centroids;



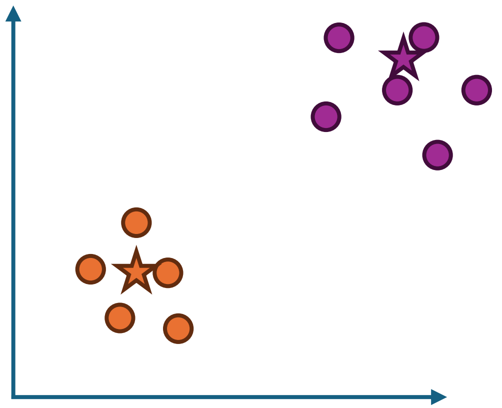
- Compute the distance between each data point to the cluster centroids;

- Check to which centroid cluster a given data point is closer. Paint (or assign a cluster label to) the data point in the same colour as the colour of the cluster centroid.

- Once all labels/colours are assigned to all data points, prepare to recompute the centroids! We will calculate the average coordinate for each of the clusters (purple and orange clusters);
- The centroids will be relocated to these new (average) coordinates.
- Then repeat the steps from the third bullet item until convergence.

K-means algorithm (illustration)

- Decide the number of centroids (equivalent to the number of clusters one wishes to identify, K);
- Randomly pick the coordinates of the position of the centroids;



- Compute the distance between each data point to the cluster centroids;
- Check to which centroid cluster a given data point is closer. Paint (or assign a cluster label to) the data point in the same colour as the colour of the cluster centroid.

- Once all labels/colours are assigned to all data points, prepare to recompute the centroids! We will calculate the average coordinate for each of the clusters (purple and orange clusters);
- The centroids will be relocated to these new (average) coordinates.
- Then repeat the steps from the third bullet item until convergence.

K-means mathematically

DEFINITION (Partition) A partition of $[n] = \{1, \ldots, n\}$ of size k is a collection of non-empty subsets $C_1, \ldots, C_k \subseteq [n]$ that:

- are pairwise disjoint, i.e., $C_i \cap C_j = \emptyset$, $\forall i \neq j$
- cover all of $[n]$, i.e., $\cup_{i=1}^k C_i = [n]$.

Wikipedia: *"In centroid-based clustering, clusters are represented by a central vector, which may not necessarily be a member of the data set. When the number of clusters is fixed to k, k-means clustering gives a formal definition as* ***an optimization problem****: find the k cluster centers and assign the objects to the nearest cluster center, such that the squared distances from the cluster are minimized."*

Under the k-means objective, the "cost" of $C_1, \ldots, C_k$ is defined as

$$\mathcal{G}(C_1, \ldots, C_k) = \min_{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k \in \mathbb{R}^d} \sum_{i=1}^{k} \sum_{j \in C_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2.$$

Here $\boldsymbol{\mu}_i \in \mathbb{R}^d$ is the representative – or center – of cluster C_i. Note that $\boldsymbol{\mu}_i$ need not be one of the $\mathbf{x}_j$'s.

Our goal is to find a partition $C_1, \ldots, C_k$ that minimizes $\mathcal{G}(C_1, \ldots, C_k)$, i.e., solves the problem

$$\min_{C_1, \ldots, C_k} \mathcal{G}(C_1, \ldots, C_k)$$

over all partitions of $[n]$ of size k. This is a finite optimization problem, as there are only a finite number of such partitions. Note, however, that the objective function itself is an optimization problem over $\mathbb{R}^d \times \cdots \times \mathbb{R}^d$, that is, k copies of $\mathbb{R}^d$.

https://mmids-textbook.github.io/chap01_intro/03_clustering/roch-mmids-intro-clustering.html