POWERING
POTENTIAL

Unsupervised Machine Learning: Clustering

Unsupervised Machine Learning: Clustering

→ Clustering - principles

→ K-means clustering

→ Finding the optimum number of clusters

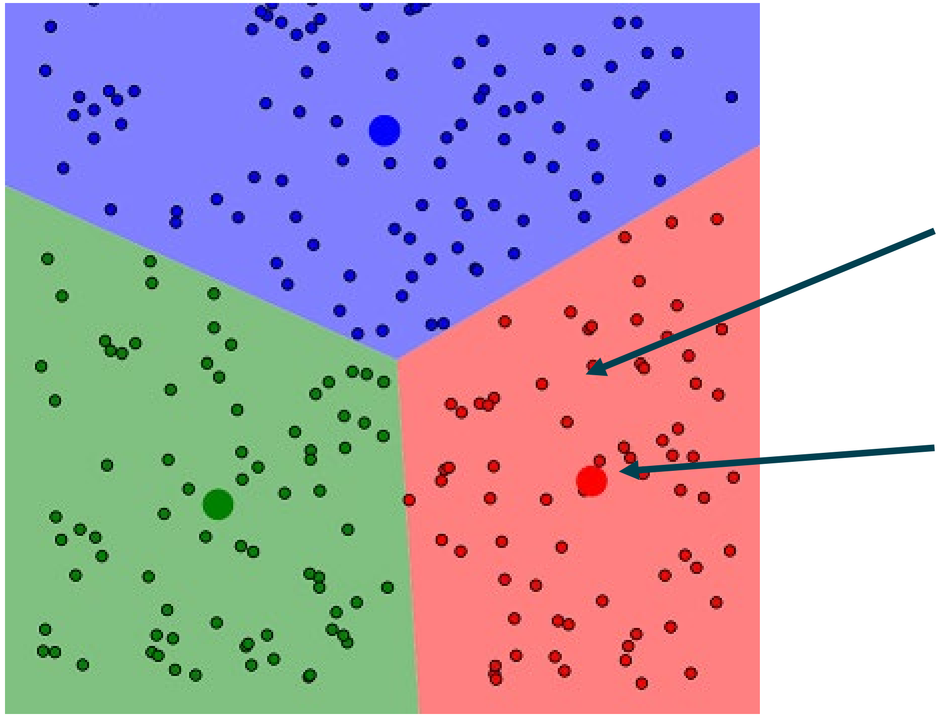
CLUSTERING

Clustering (Cluster Analysis)

- Analytical technique to ***group objects according to their attributes*** so that objects belonging to the same group are ***much more similar to objects in the same group*** than to objects in other groups.
- Finds **existing** patterns in data.
- **Unsupervised learning**.
- Once the groups are determined and labelled, the clustered data can be used to perform **classification**.
- One of the most popular clustering method is **k-means clustering.** K is the number of clusters (groups) that is an input to the method.

K-MEANS CLUSTERING

https://www.naftaliharris.com/blog/visualizing-k-means-clustering/



K-means clustering is based on:

- distances (this means only quantitative attributes can be used).
- centroids.

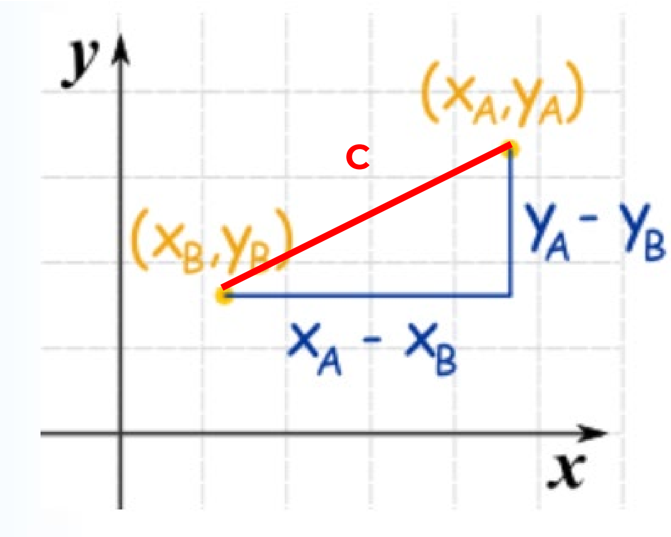
K-MEANS CLUSTERING

Clustering is based on objects (data points), centroids, and the distances between them.

- Each **object** (data point) that we want to group (cluster) is represented as a **point with coordinates its attributes**.
- A **centroid** is the “average representative” of a group of objects. It is a calculated data point with value of **each attribute the average** of the attribute’s values of the members of the group.

  → If the group consists of people, and the attributes we consider are height and weight, the centroid will have the average height and the average weight of the group.

- Usually **Euclidean distance** between the centroids and the objects is used. For 2 attributes:

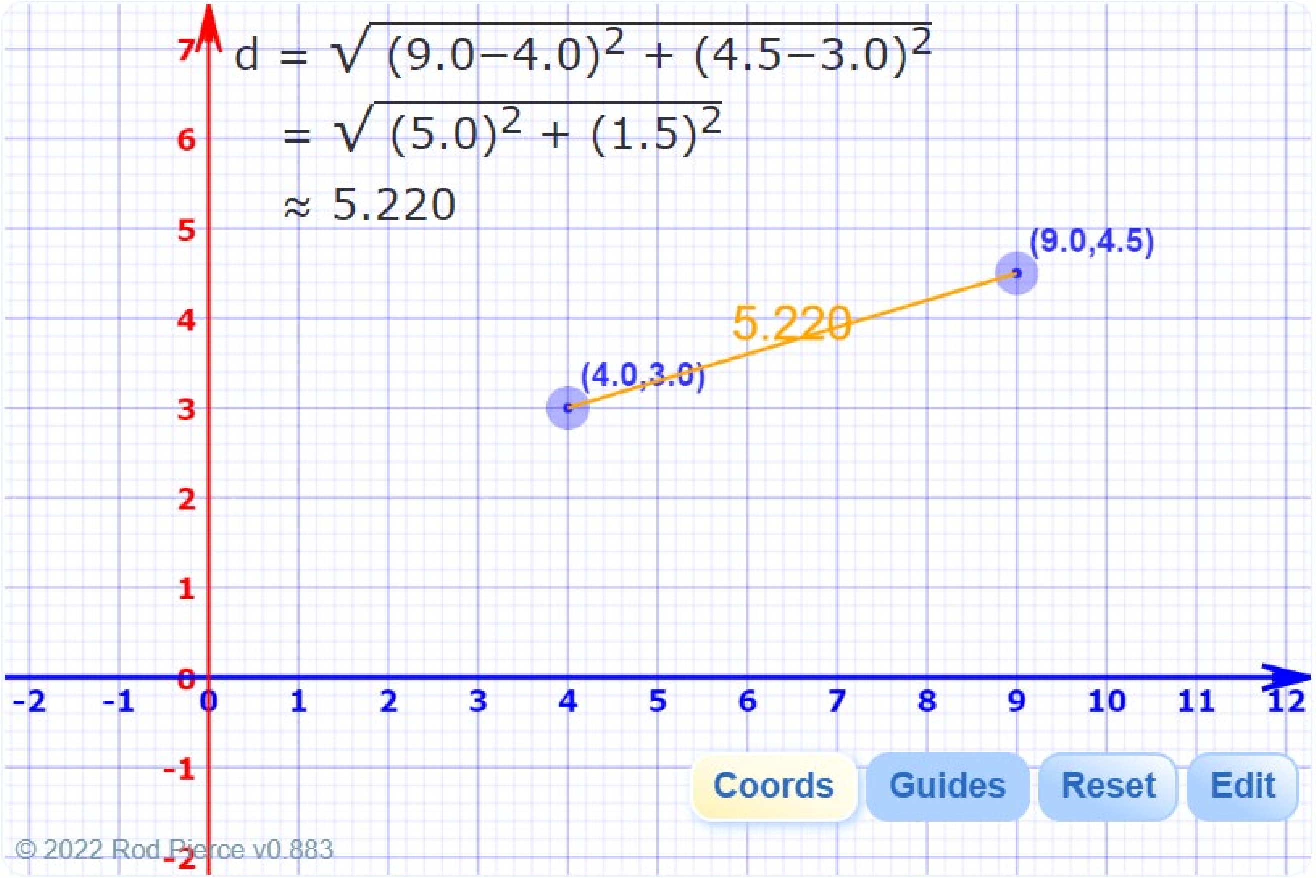


Remember Pythagoras theorem?

$$c = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}$$

ATTRIBUTES, COORDINATES, DISTANCES

Attributes (coordinates) determine distances (similarities)



$$
\begin{aligned}
d &= \sqrt{(9.0-4.0)^2 + (4.5-3.0)^2} \\
&= \sqrt{(5.0)^2 + (1.5)^2} \\
&\approx 5.220
\end{aligned}
$$

Have a go yourself – interactively:

https://www.mathsisfun.com/algebra/distance-2-points.html

K-MEANS CLUSTERING ALGORITHM

How k-means clustering works:

(0) The number of clusters k must be specified

(1) Select k random points for initial centroids

(2) Assign each data point to its nearest centroid

(3) Re-calculate the centroid for each group (average of the points in the group)

(4) Repeat (2) and (3) until the centroids and the groups stabilise

→ **Try a very good interactive animation of k-means clustering:**

https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

K-MEANS CLUSTERING – OPTIMUM NUMBER OF CLUSTERS

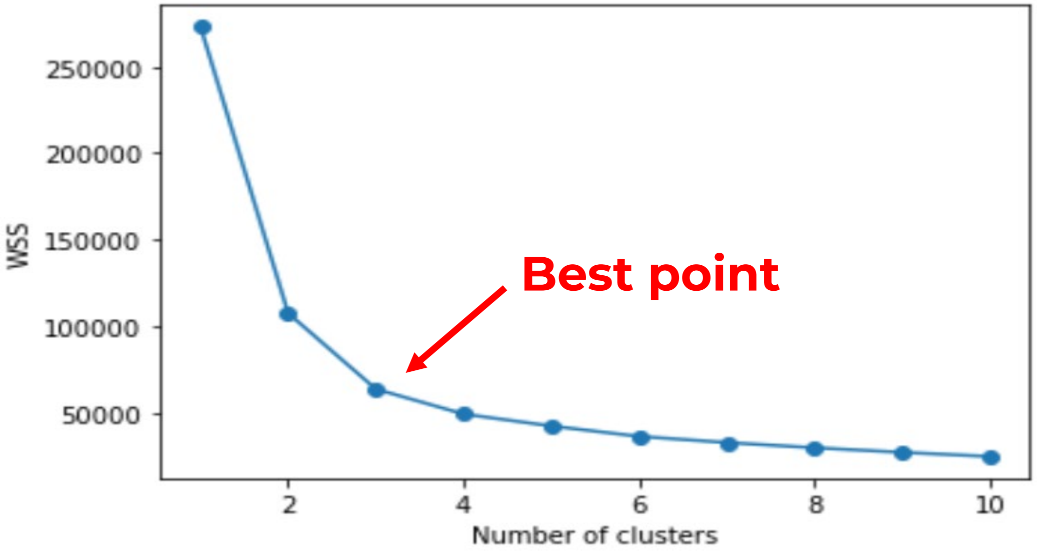
What is the best (optimum) number of clusters ?

Elbow method:

→ **k-means clustering is run for a series of numbers of clusters** (e.g., 1:10).

→ In each run, a score is computed, usually **Within Sum of Squares (WSS)** – the sum of squared distances from each data point to each centroid.

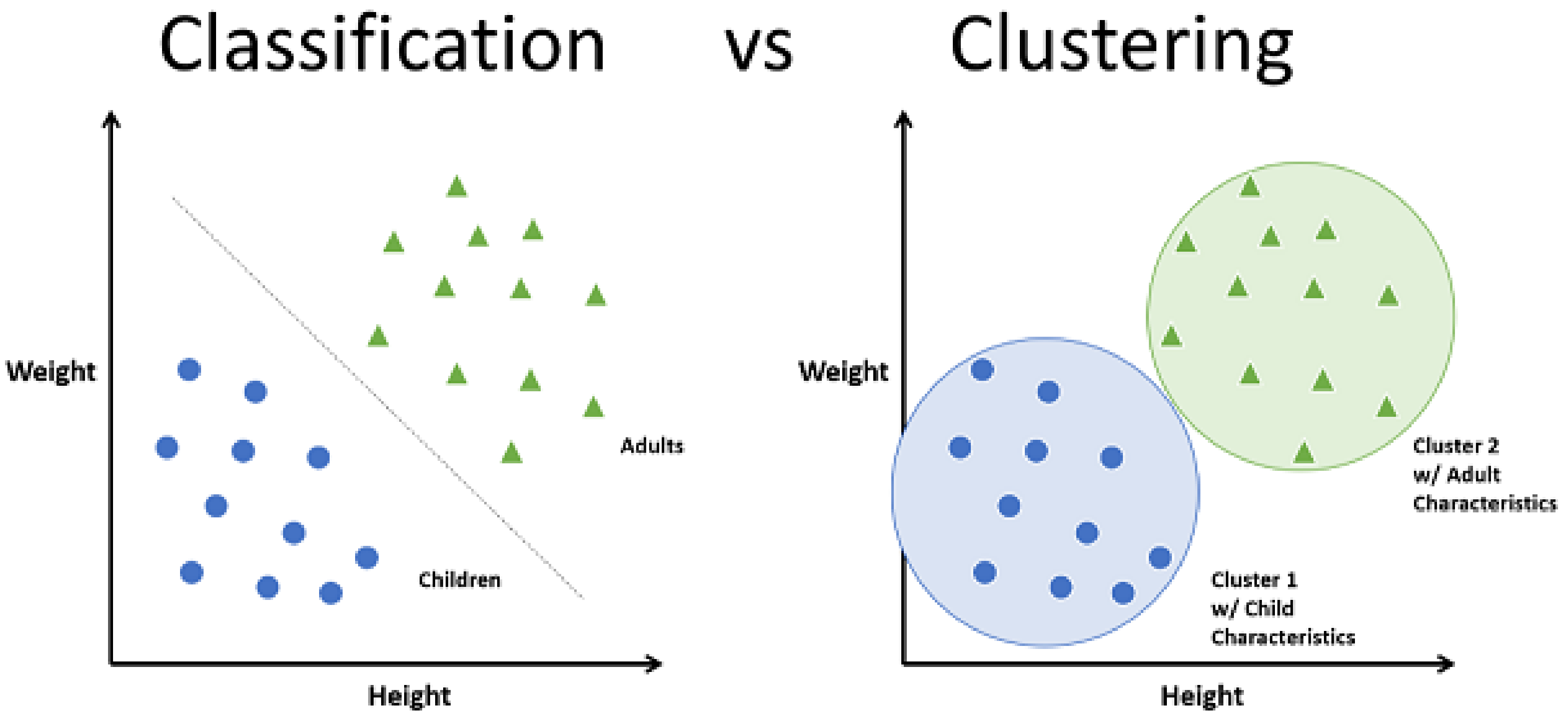
→ Plotting WSS for each k produces a graph looking like an elbow. The “elbow point” indicates the optimum number of clusters.



Best point

CLUSTERING VS CLASSIFICATION

Clustering determines and labels the groups. This data can then be used to perform **classification – determine to which labelled group new data points belong.**


Classification
vs
Clustering
Weight
Height
Adults
Children
Weight
Height
Cluster 2 w/ Adult Characteristics
Cluster 1 w/ Child Characteristics

Source:

https://www.analyticsvidhya.com/blog/2021/05/what-why-and-how-of-spectral-clustering/