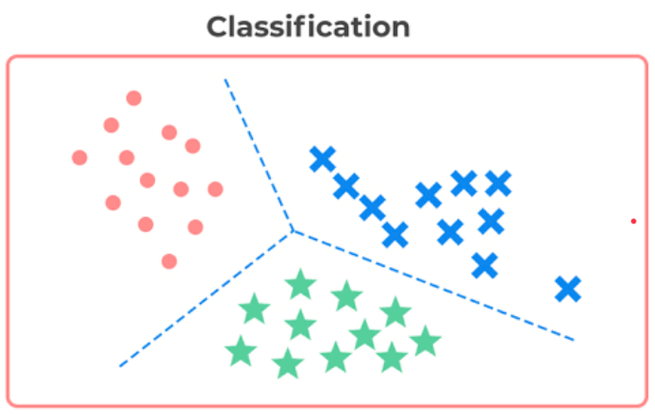
Ibrahim radwan

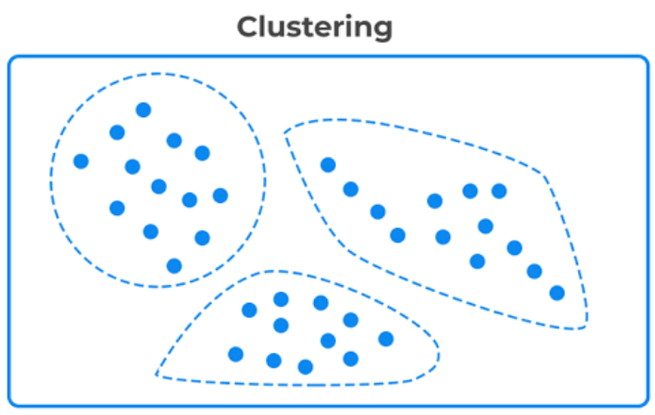
Clustering

20/09/2024

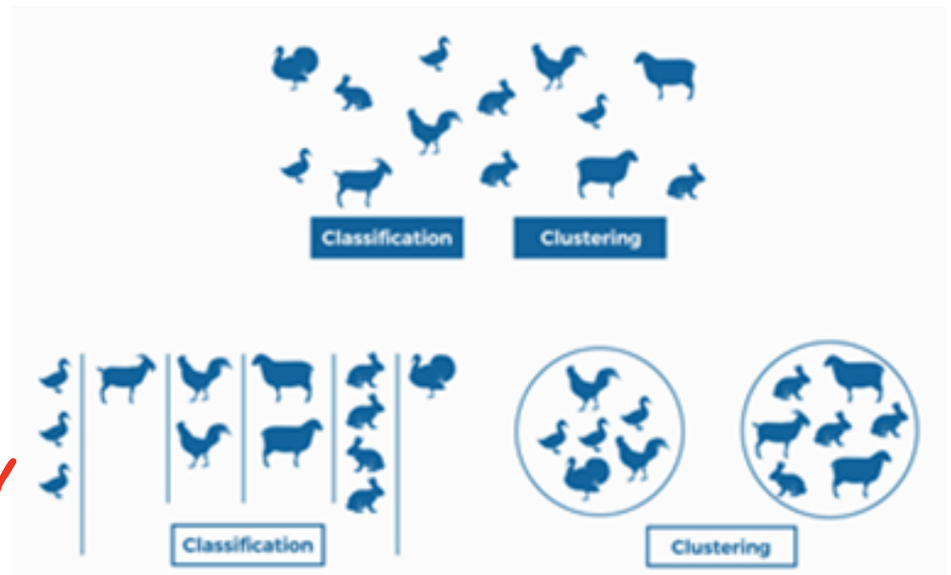
⇒ Grouping or partitioning data based on their attributes



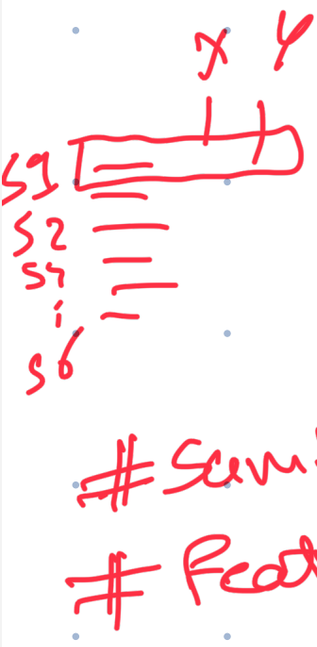
label - based



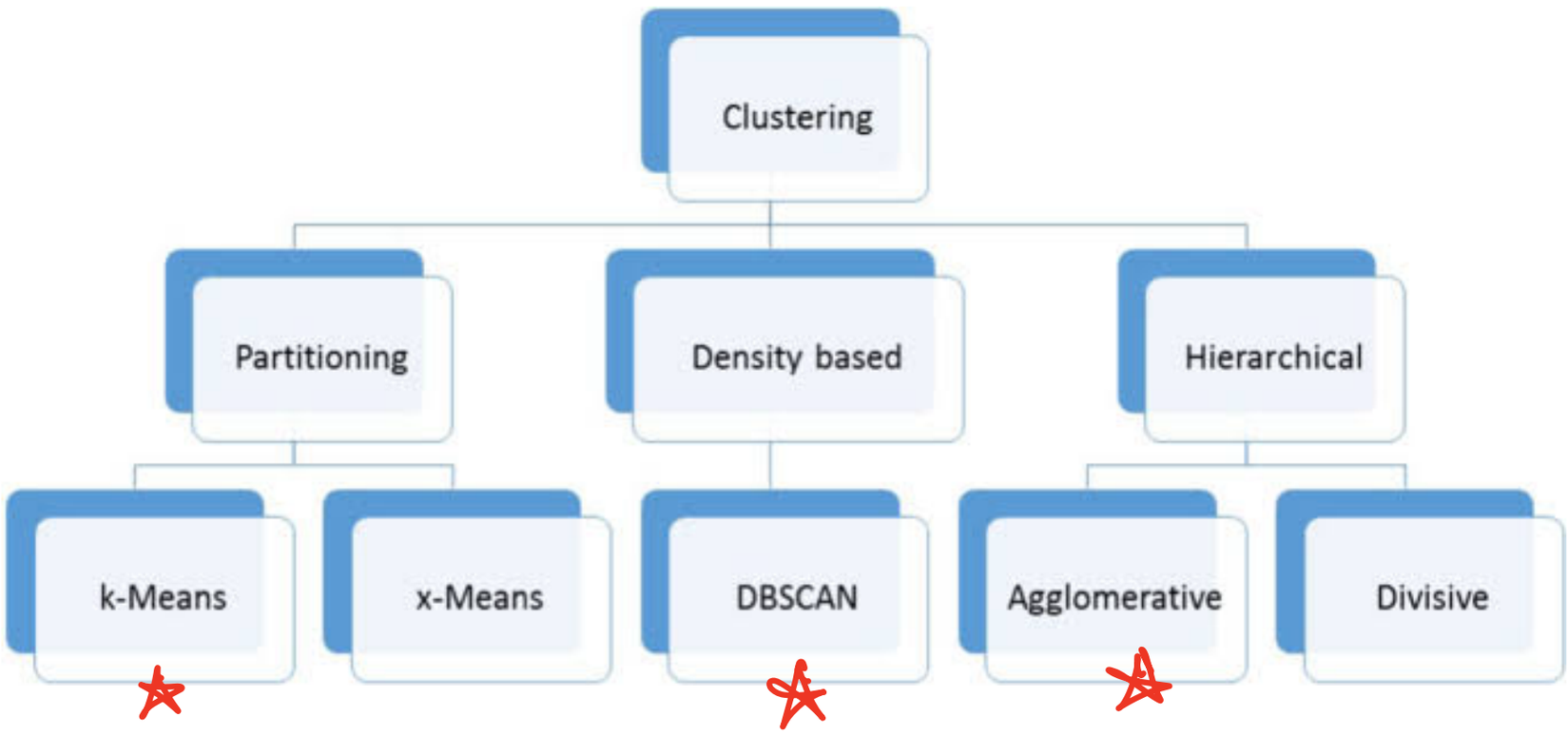
Attribute - based



⇒ Types of clustering



#samples
#features



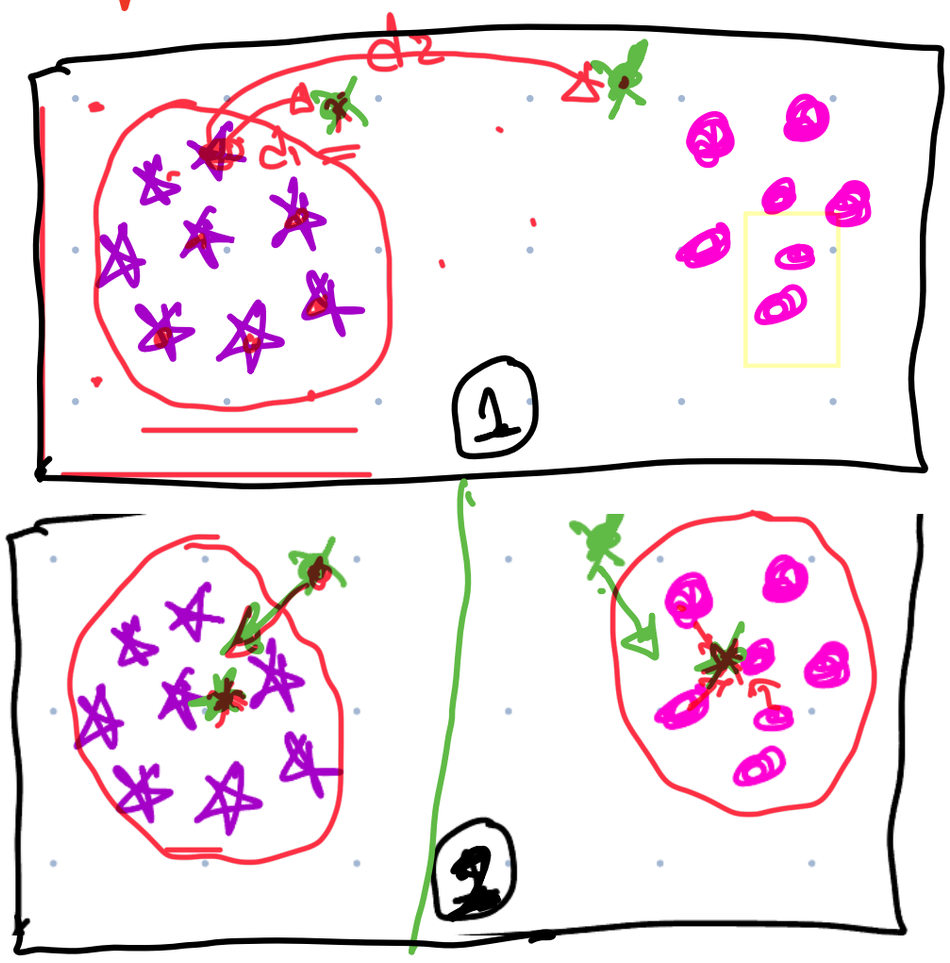
* Common ones

⇒ K-Means

General steps

- An iterative clustering algorithm
  - Initialize: Pick *K* random points as cluster centers
  - Alternate:
    1. Assign data points to closest cluster center
    2. Change the cluster center to the average of its assigned points
  - Stop when no points' assignments change

Visual Explation



K = 2

K-Means

Number of clusters ← K

Centroids of clusters

K-Means Mathematically

K-means Solution

A proposal: Minimise the *distortion function*, i.e., the sum of the squared distances of each data point to its closest vector $\boldsymbol{\mu}_k$.

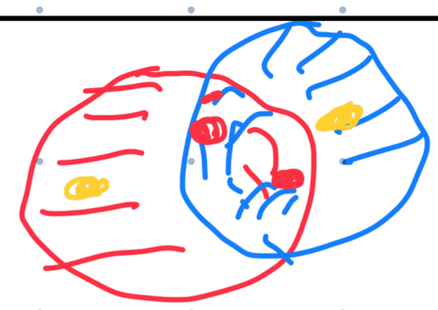
$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$$

← objective function

1. Given K, randomly select $\boldsymbol{\mu}_{k=1,\cdots,K}$
2. Minimise J with respect to r_{nk}, keeping the $\boldsymbol{\mu}_k$ fixed. ← Equivalent to Expectation step.
3. Minimise J with respect to $\boldsymbol{\mu}_k$, keeping the r_{nk} fixed. ← Equivalent to Maximization step.
4. Repeat steps 2 (*Expectation*) and 3 (*Maximisation*) steps until convergence, that is, $\Delta J < \epsilon$.

⇒ Initialization of K-Means

- Randomly



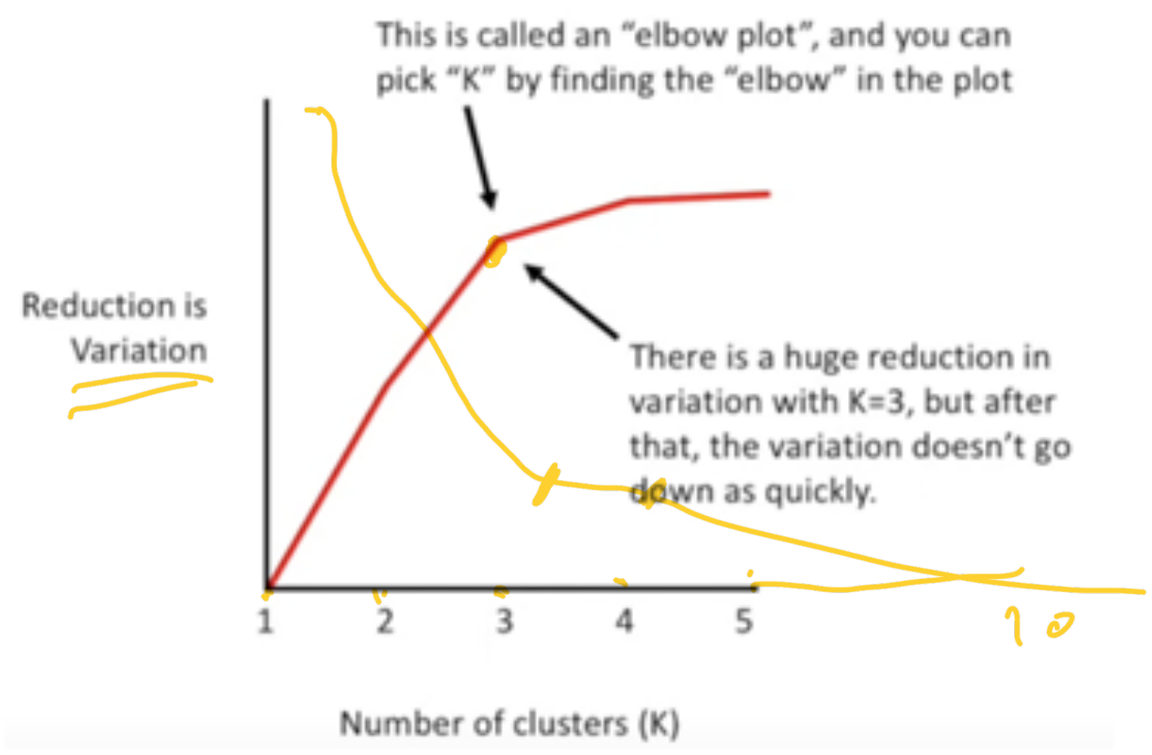
- Manually by selecting specific positions to the centroids.

- Can be done smartly ⇒ kMeans++

⇒ How to find the best value for k ??

- Elbow Method

1 - 10



Try different values of K & choose the point that shows significant reduction in variation

- Calculate WCSS for different Ks
- Plot
- Identify

How to Evaluate the clustering outcome?

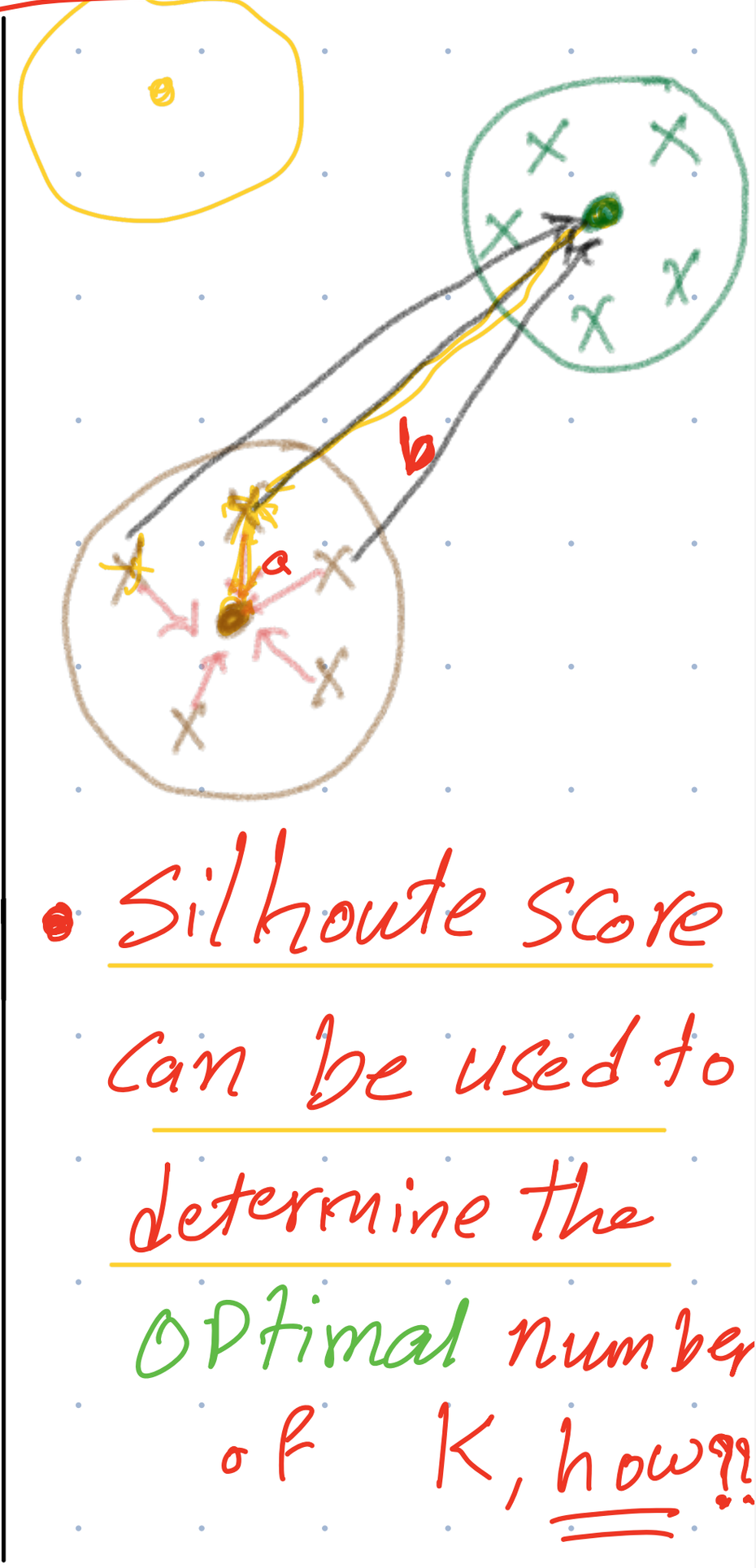
- ## Silhoute score

SSI ⇔ Silhoute score index

$$SSI_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

- a_i: distance between point i and the corresponding centroid.
- b_i: distance between Point i and the closest centroid that this Point does not belong to.
- SSI_i: can be between -1 to 1
  - -1 ↗ very bad
  - 1 ↖ very good

The main idea is to assign each Point with a score, that measures the quality of being belonging to a specific cluster and far enough from other clusters



- Silhoute score can be used to determine the optimal number of K, how??