K-Means Clustering

datasciencedojo
unleash the data scientist in you

Unsupervised Learning

- trying to find hidden structure in unlabeled data
- no error or reward signal to evaluate a potential solution
- Common techniques: K-Means clustering, Hierarchical clustering, hidden Markov models, etc.
- It has a long history, and used in almost every field, e.g., medicine, psychology, botany, sociology, biology, archeology, marketing, insurance, libraries, etc.

Unsupervised learning

Example 1: Clothing size

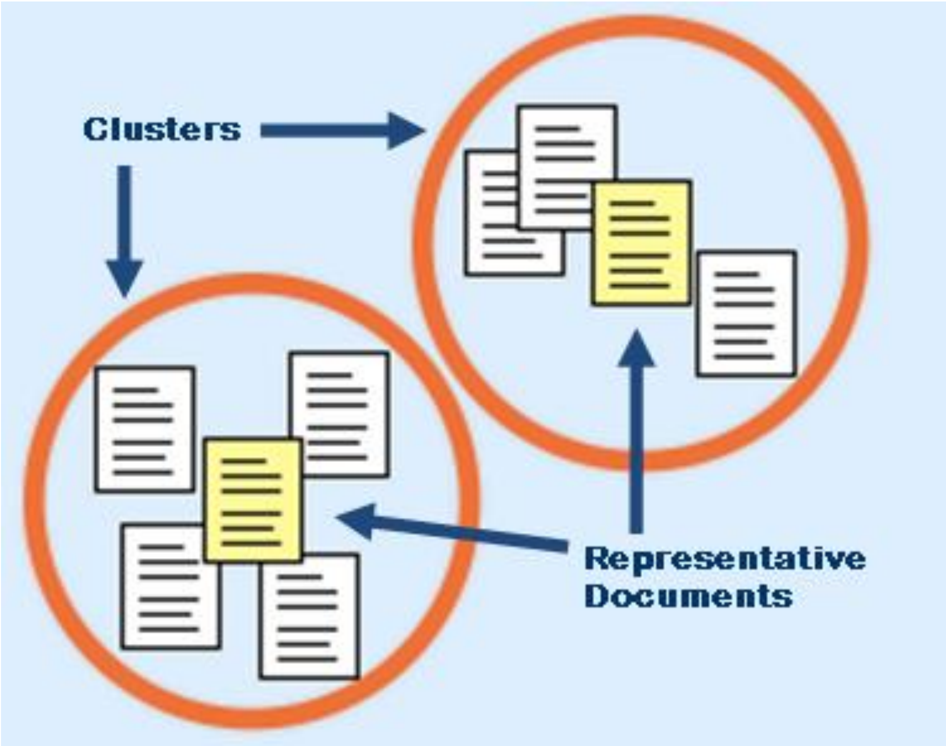
- Tailor-make for each person is too expensive
- One-size-fit-all: do not work!!
- groups people of similar sizes together to make “small”, “medium” and “large” T-Shirts.

datasciencedojo

Unsupervised learning

Example 2: Text document organization

- To find groups of documents that are similar to each other based on the important terms appearing in them.

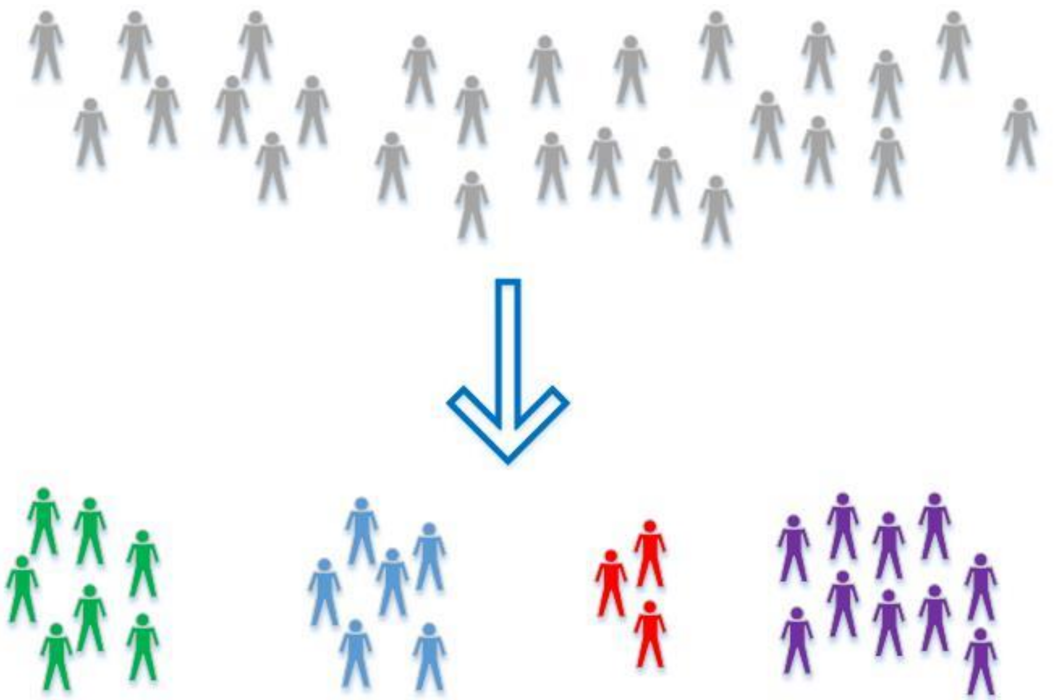


datasciencedojo

Unsupervised learning

Example 3: Target Marketing

Subdivide market into distinct subsets of customers where any subset may conceivably be selected as a segment to be reached with a particular offer.



datasciencedojo

K-means clustering

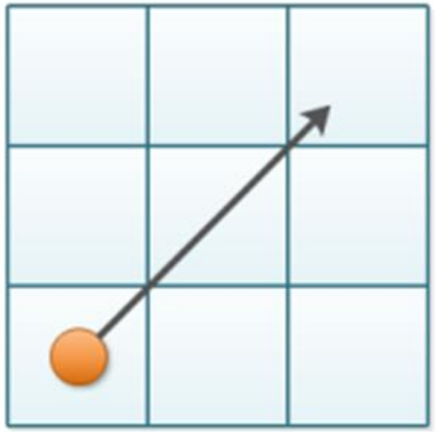
- Process of partitioning data points into similarity clusters

- Unsupervised technique

- Only works for numeric data

Euclidean Distance

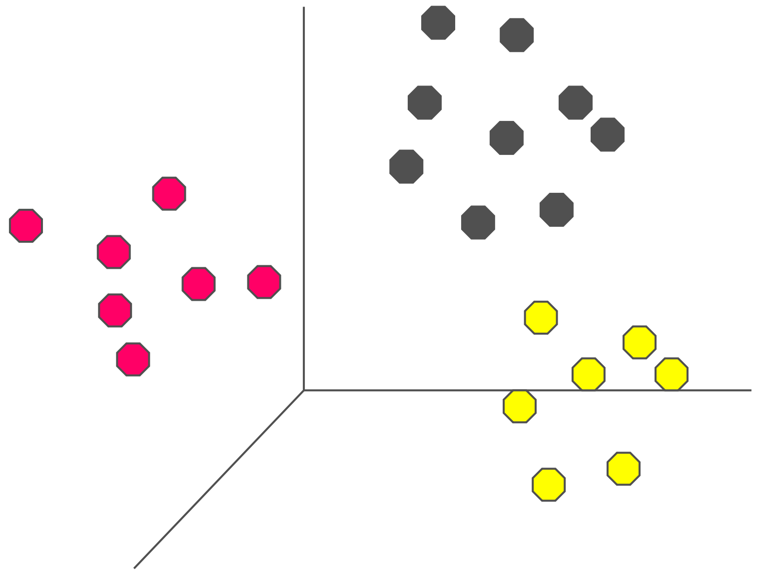
- points in a two-dimensional space to determine intra- and inter-cluster similarity.



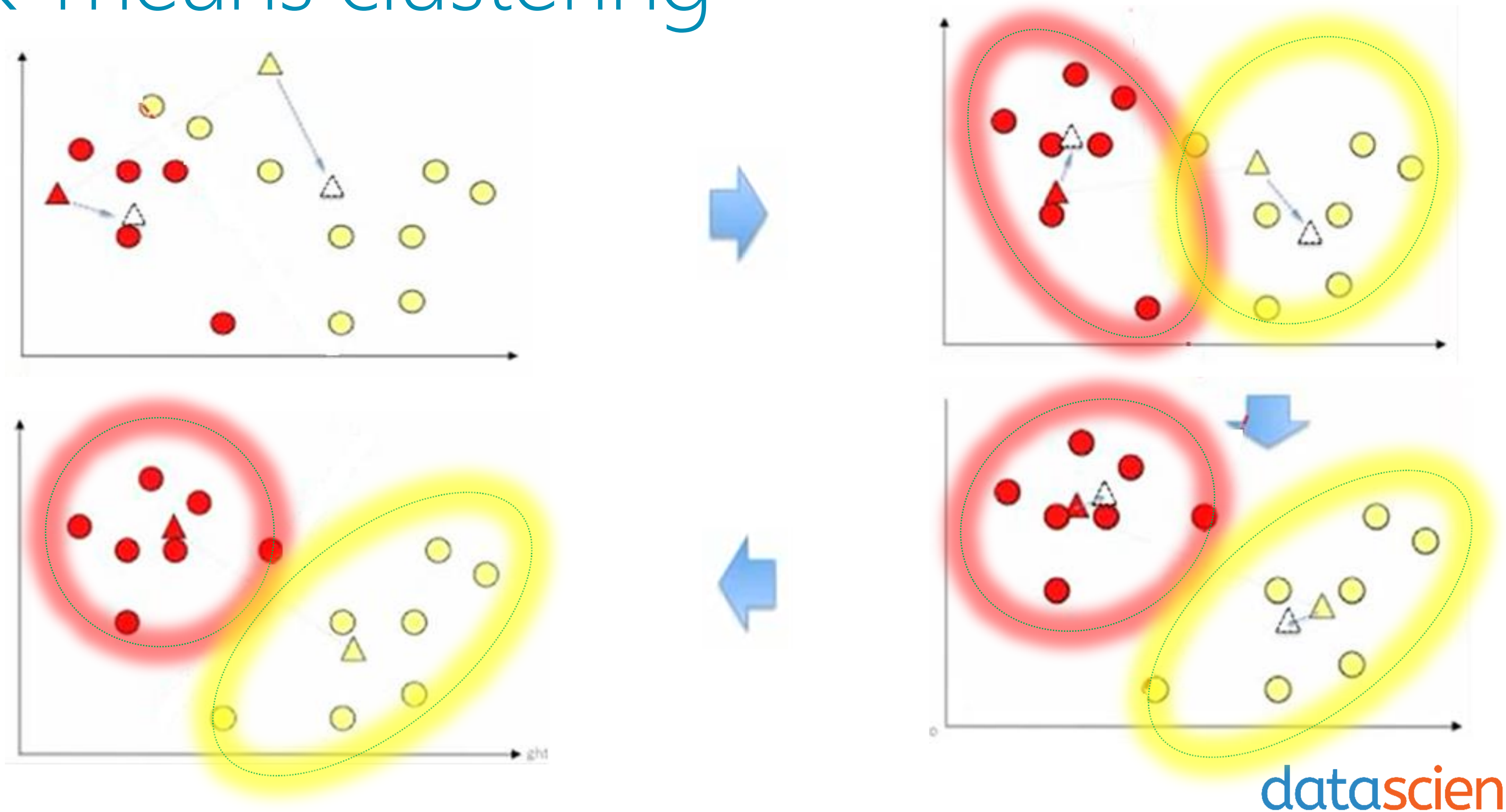
$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Intra-cluster distances are minimized

Inter-cluster distances are maximized



K-means clustering



K-means clustering algorithm

Suppose set of data points: $\{ x_1, x_2, x_3 \ldots\ldots\ldots x_n \}$

- Step 1: Decide the number of clusters, K=1,2,…k.
- Step 2: Place centroids at random locations
  - $c_1, c_2, \ldots. c_k$
- Step 3: Repeat until convergence:

{ for each point $x_i \longrightarrow$ find nearest centroid c_j (eg. Euclidean distance)

$\longrightarrow$ assign the point x_i to cluster j

for each cluster j = 1…k $\longrightarrow$ calculate new centroid c_j

c_j=mean of all points x_i assigned to cluster j in previous step

}

- Step 4: Stop when none of the cluster assignments change

datasciencedojo

K-means clustering

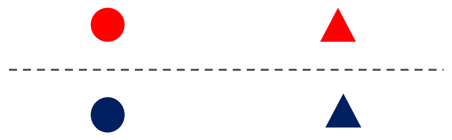
Minimizes aggregate intra-cluster distance

- Measure squared distance from point to center of its cluster. $\sum_j \sum_{x_j - ci} D(c_j x_i)^2$

Could converge to local minimum

- Different starting points $\longrightarrow$ very different results
- Run many times with random starting points

Nearby points may not be assigned to the same cluster



datasciencedojo

K-means clustering

Strengths:

Simple: easy to understand and to implement

Efficient: Complexity: $O(t \ x \ k \ x \ n)$

n = number of data points,
k =number of clusters, and
t = number of iterations.

K-means clustering

Weaknesses:

- The algorithm is only applicable if the mean is defined.
- The user needs to specify *k*.
- The algorithm is sensitive to **outliers**

K-Means clustering

Rule of thumb

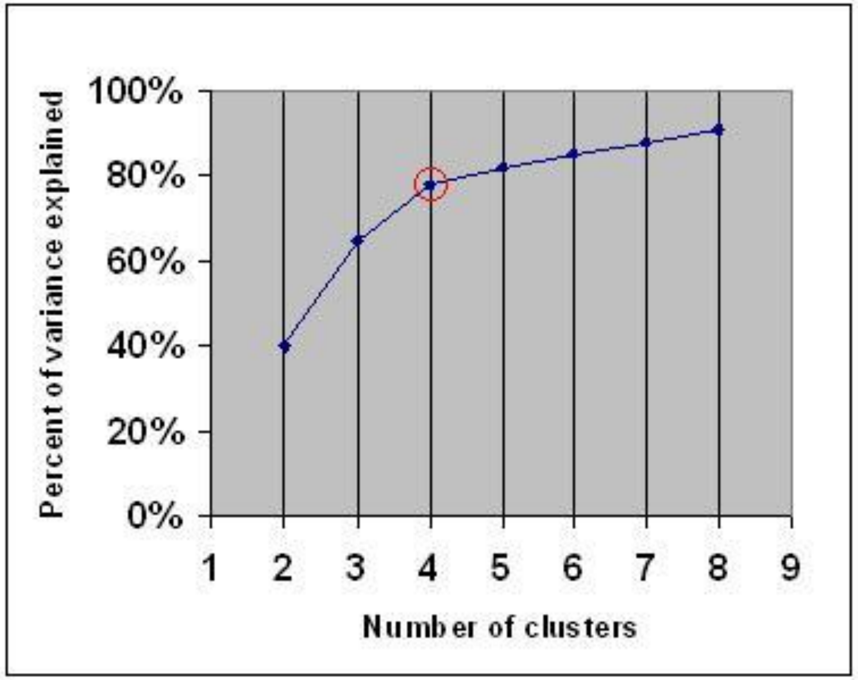
$$k \approx \frac{\sqrt{n}}{2}$$

n = number of data points

Elbow method

- percentage of variance explained as a function of the number of clusters

- choose a number of clusters so that adding another cluster doesn't give much better modeling of the data.



K-means clustering

Others k optimization techniques

- Silhouette
- Calinsky criterion
- Bayesian Information Criterion
- Affinity propagation (AP) clustering
- Gap statistic

Questions?

datasciencedojo
unleash the data scientist in you