Unsupervised Learning and K-Means Clustering

Data Science Dojo

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

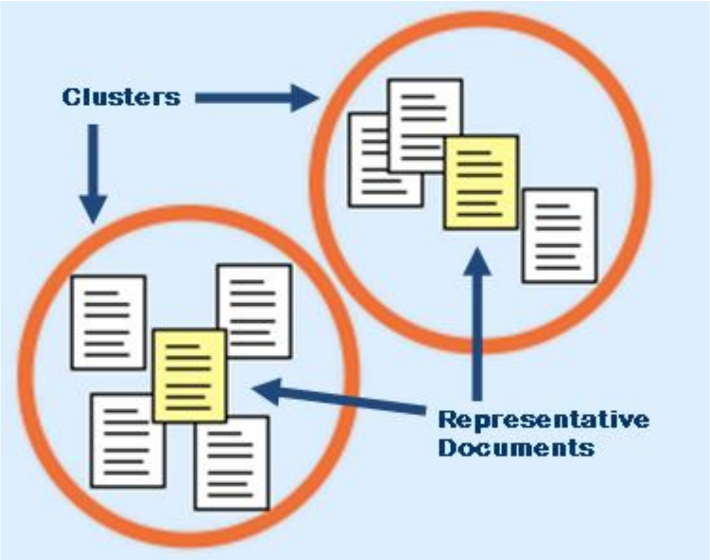
Example 1: Clothing size

- Tailor-made for each person is too expensive
- One-size-fits-all: does not work!
- Groups people of similar sizes together to make “small”, “medium”, and “large” t-shirts

Unsupervised Learning

Example 2: Text document organization

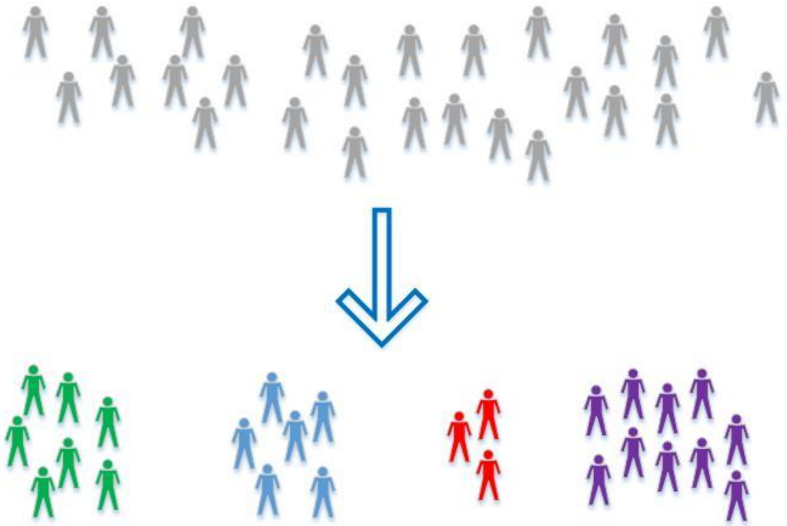
- To find groups of documents that are similar to each other based on the important terms appearing in them



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

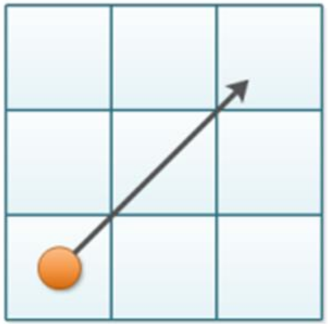


K-Means Clustering

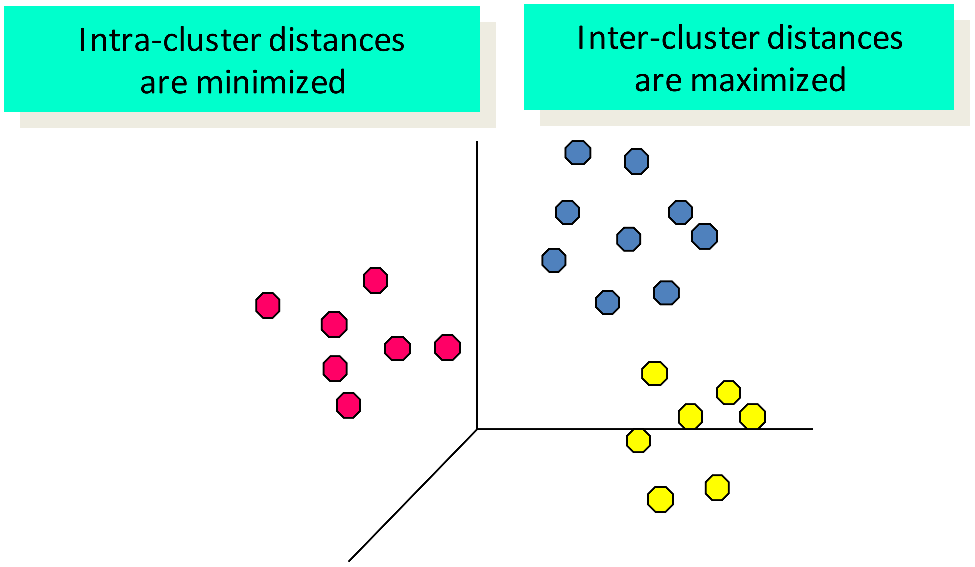
- Partitions data points into similarity clusters
- Only works for numeric data

Euclidean Distance

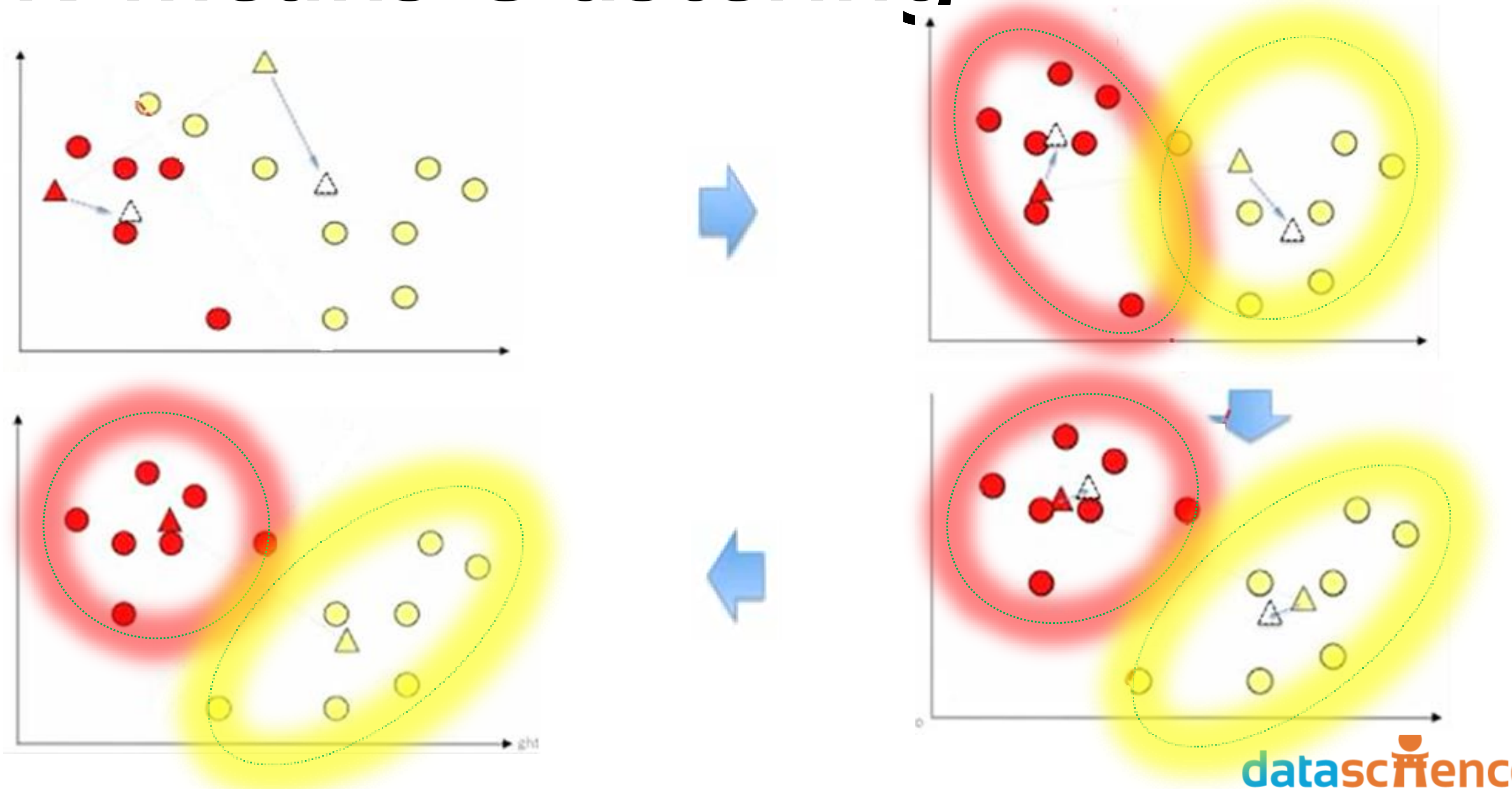
points in a two-dimensional space to determine intra- and inter-cluster similarity



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



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 ⟶ find nearest centroid c_j (eg. Euclidean distance)

⟶ assign the point x_i to cluster j

for each cluster j = 1...k ⟶ 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

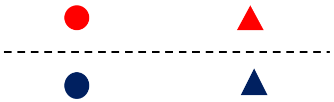
unleash the data scientist in you

K-Means Clustering

- Minimizes aggregate intra-cluster distance
  - Measure squared distance from point to center of its cluster.

$$\sum_{j=1}^{K} \sum_{x \in g_j} D(c_j, x)^2$$

- 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



K-means Clustering

- Strengths
  - Simple: easy to understand and to implement
  - Efficient: linear time, minimal storage
- Weaknesses
  - Mean must be well defined
  - The user needs to specify K
  - Sensitive to outliers

How many clusters?

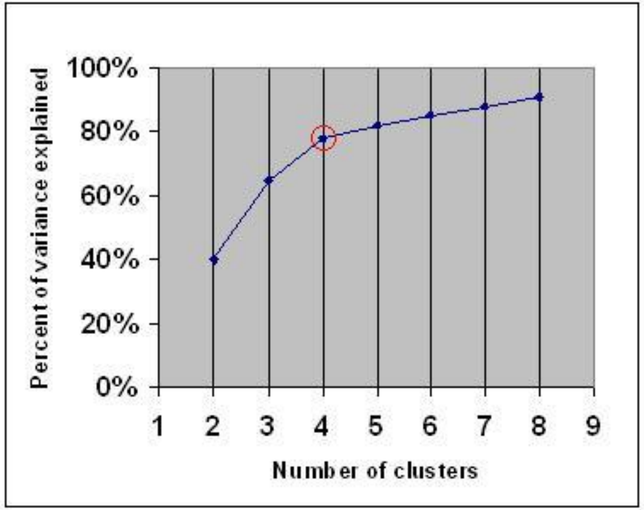
Rule of thumb

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

n = number of data points

Elbow method

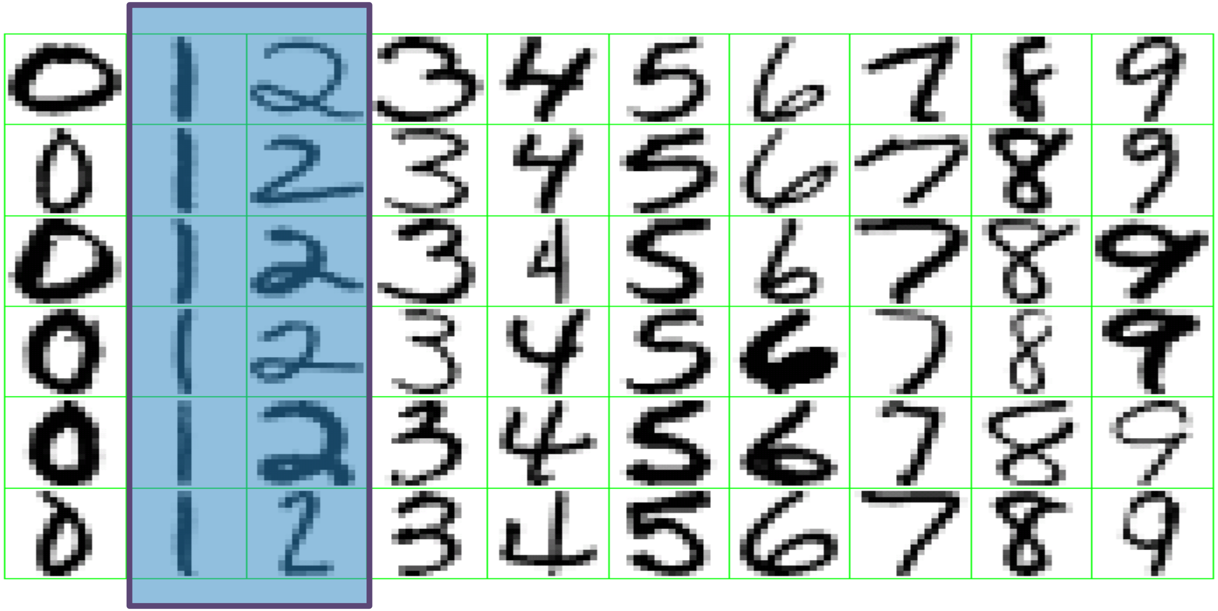
- percentage of variance explained as a function of the number of clusters
- choose a number of clusters so that adding another cluster doesn't give much better modeling of the data.



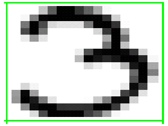
Other K Optimization Techniques

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

Example: Handwritten Digit Recognition



Extracting Features For Learning

 $\{x_1, x_2, x_3 \ldots\ldots\ldots\, X_{256}, y = \text{'three'}\}$

- Each xi corresponds to a feature value in the image
- y is a label of the training data; can be numeric or categorical, ‘3’ or ‘three’
- Each image is converted to row vectors and the appropriate learning algorithm is used
- Convention
  - x_i represents the ith feature in a training sample
  - y represents the label for the training sample

QUESTIONS