Unsupervised Learning and K-Means Clustering

datasciencedojo
data science for everyone

Unsupervised Learning

- Trying to find hidden structure in unlabeled data
- No error or reward signal to evaluate a potential solution. *No need to pick a response class*.
- 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

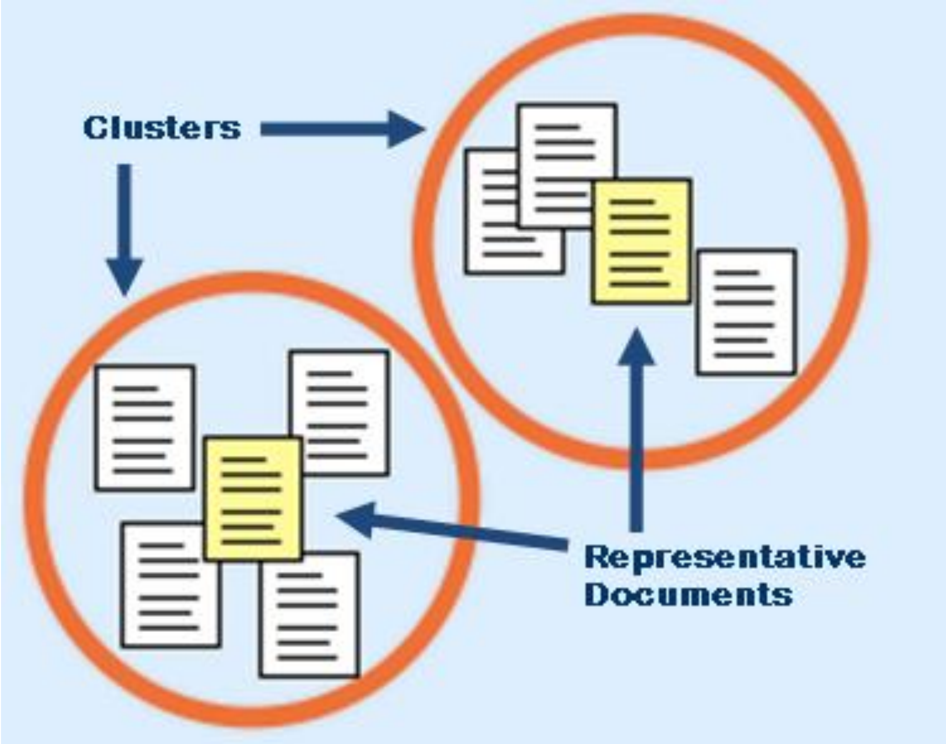
Example 1: Clothing sizes

- 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 tags

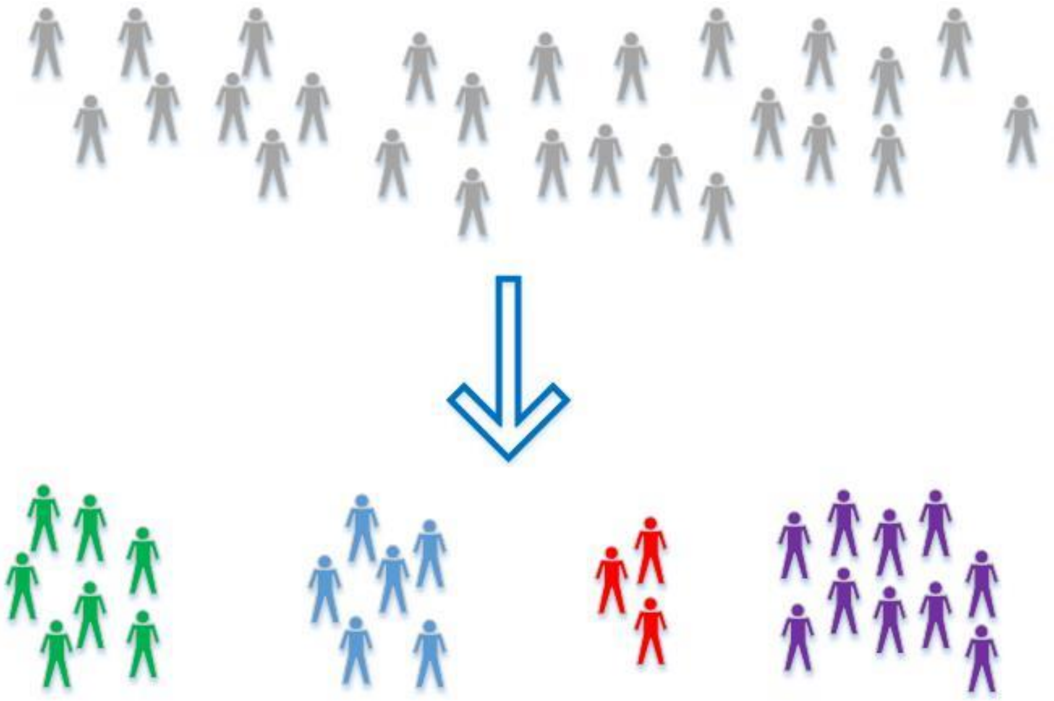
- To find groups of documents that are similar to each other based on the important terms appearing in them



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
- Unsupervised technique: there is no partitioning into a learning or a test set in unsupervised learning
- Useful in grouping observations
- Only works for numeric data

K-Means Clustering

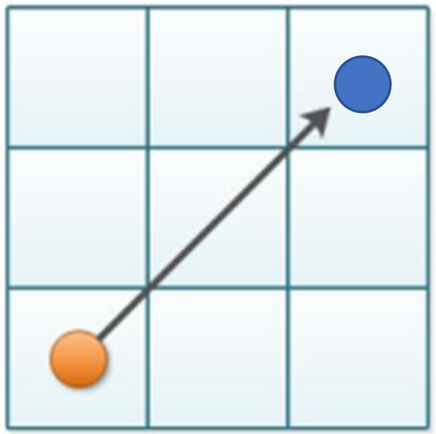
- Transform categorical variables into numeric
- Datasets will become wide quickly
- Needed to compute similarity

Often called "dummy variables" or "one-hot encoding"

Age	Pclass.1	Pclass.2	Pclass.3	Sex.female	Sex.male
19	0	1	0	0	1
28	1	0	0	1	0
64	0	0	1	0	1

Euclidean Distance

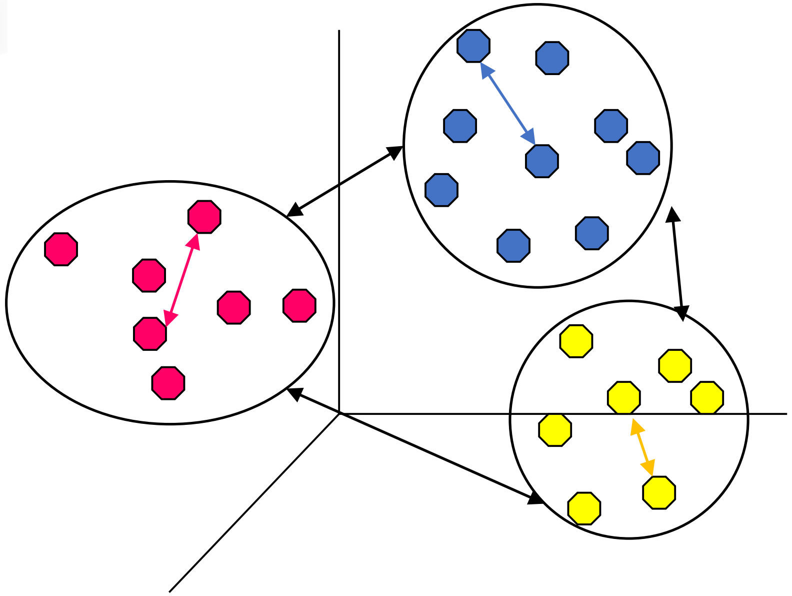
Points in a two-dimensional space to determine intra- and inter-cluster similarity



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

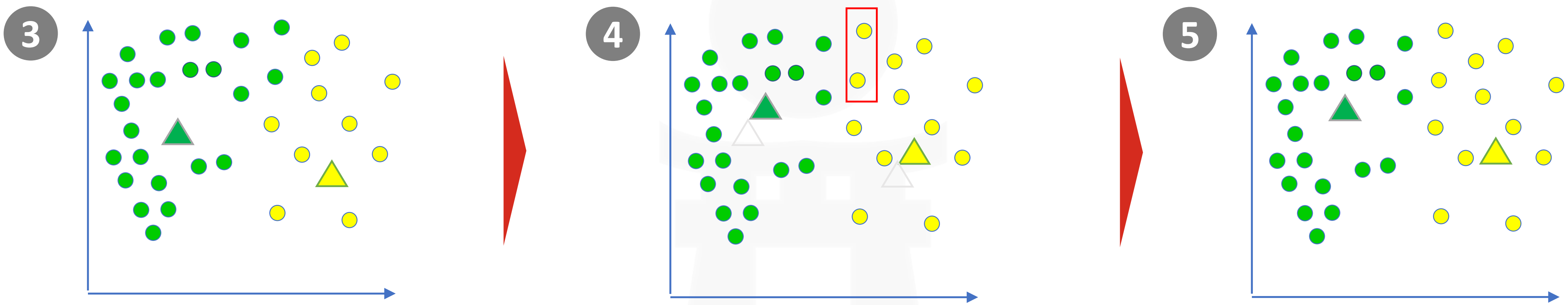
Intra-cluster distances are minimized

Inter-cluster distances are maximized



K-Means Clustering (1/2)

K-Means Clustering (2/2)



The positions of the cluster centers are determined by the mean of all the points in the cluster.

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 $x_i \rightarrow$ find nearest centroid c_j (eg. Euclidean distance)

$\rightarrow$ assign the point x_i to cluster j

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

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

}

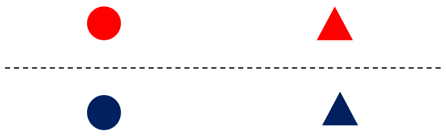
- Step 4: Stop when none of the cluster assignments change

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 → very different results
    - Run many times with random starting points
- Nearby points may not be assigned to the same cluster



K-Means Clustering

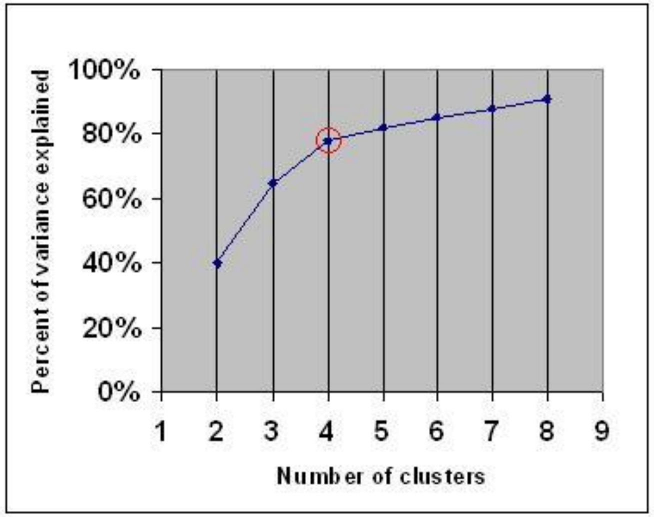
Strengths

- Simple: easy to understand and to implement
- Efficient: linear time, minimal storage

Weaknesses

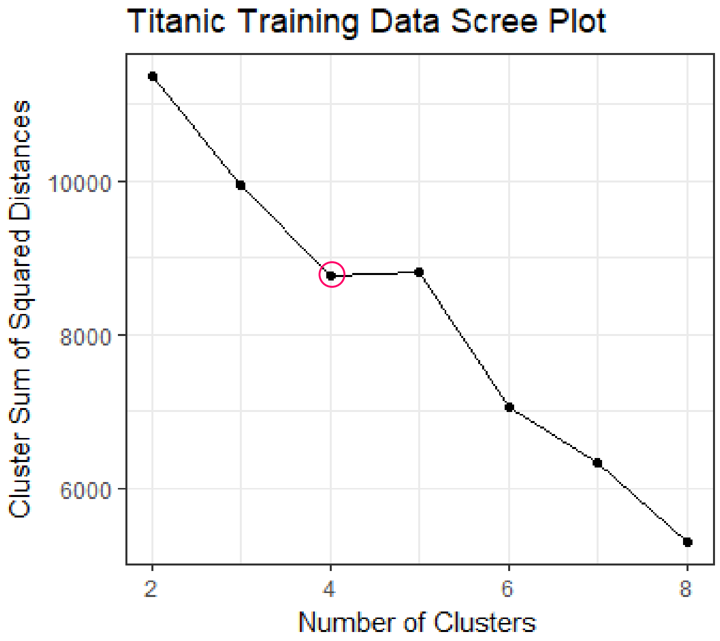
- Mean must be well defined
- The user needs to specify k
- Algorithm is sensitive to outliers

Finding K with Elbow Method



Option 1 - Percentage of variance explained as a function of the number of clusters.

Option 2 -Total of the squared distances of cluster point to center.



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

Other Clustering Techniques

- Silhouette
- Calinski-Harabasz Criterion
- Bayesian Information Criterion
- Affinity propagation (AP) Clustering
- Gap Statistic

QUESTIONS