Unsupervised Learning and K-Means Clustering

Data Science Dojo

datasciencedojo
data science for everyone

Unsupervised Learning (1/5)

- 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.

datasciencedojo
data science for everyone

Unsupervised Learning (2/5)

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

datasciencedojo

data science for everyone

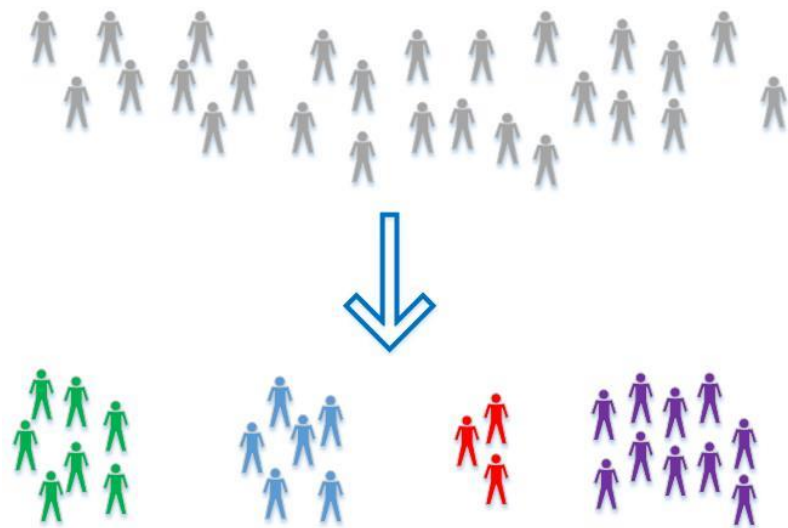# Unsupervised Learning (3/5)

Example 2: Text document organization

- To find groups of documents that are similar to each other based on the important terms appearing in them
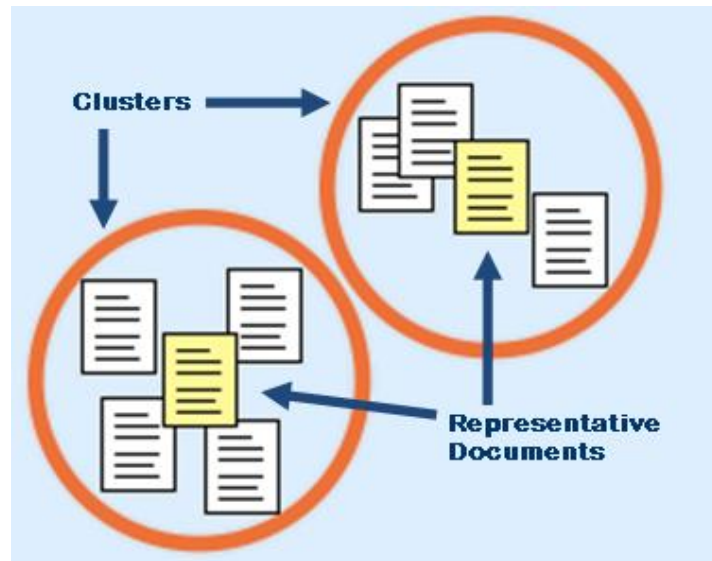
Unsupervised Learning (4/5)

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

Unsupervised Learning (5/5)

Example 4: Social network graphs

- Subdivide social network into distinct subsets of user groups (Facebook friends, LinkedIn contacts...)
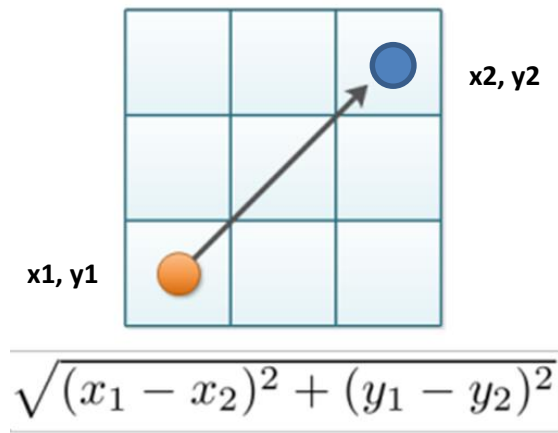
K-Means Clustering

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

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

datasciencedojo
data science for everyone

Euclidean Distance

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

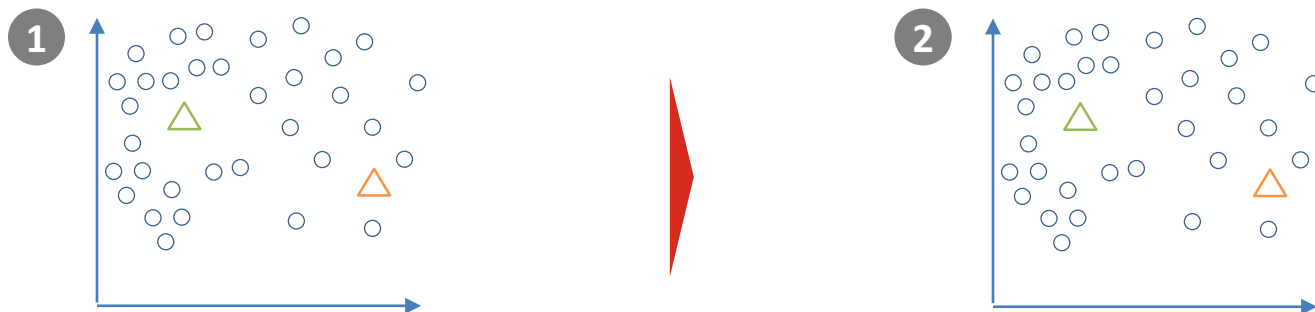$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$
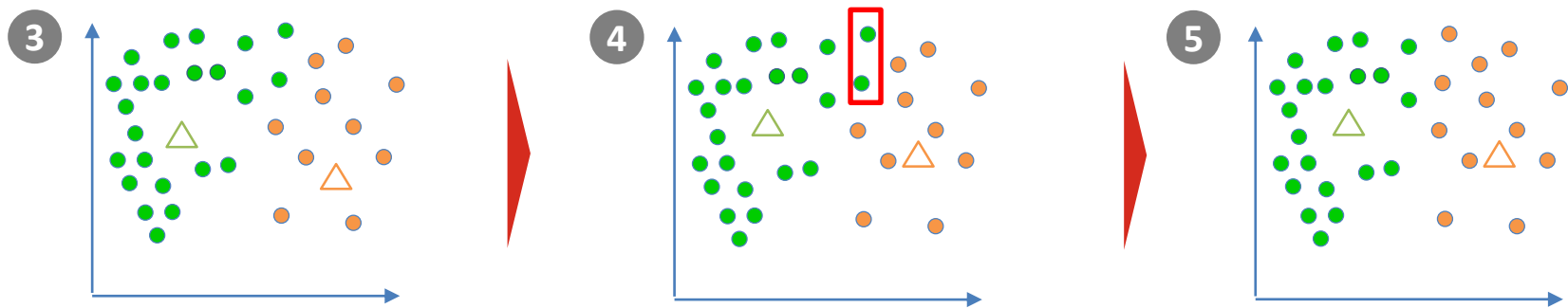
Intra-cluster distances are minimized

Inter-cluster distances are maximized
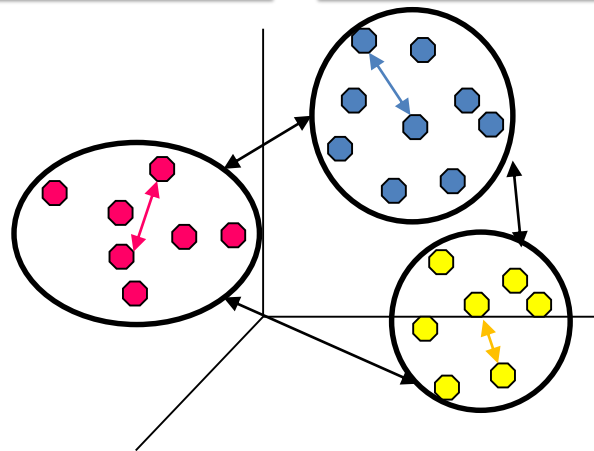
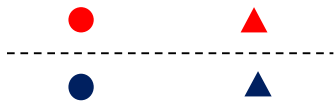# K-means Clustering (1/2)

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

- 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
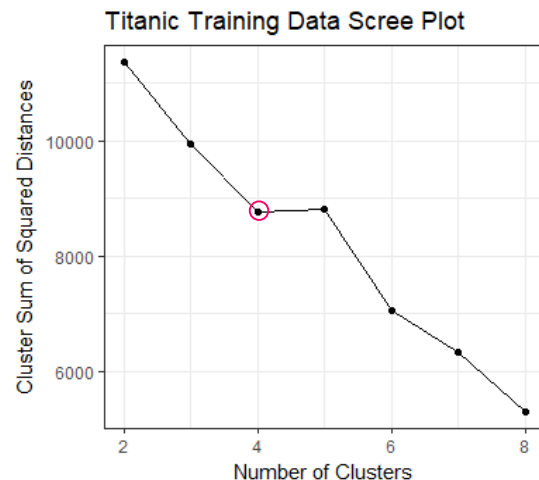 - Algorithm is sensitive to outliers

Finding K with Elbow Method
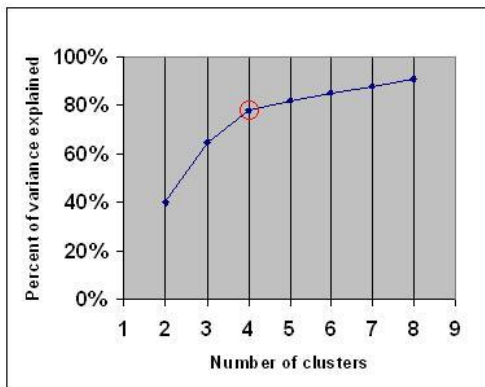
Option 1 - Percentage of variance explained as a function of the number of clusters.

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

Titanic Training Data Scree Plot

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

14

Other K Optimization Techniques

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

Comparing Clustering Algorithms
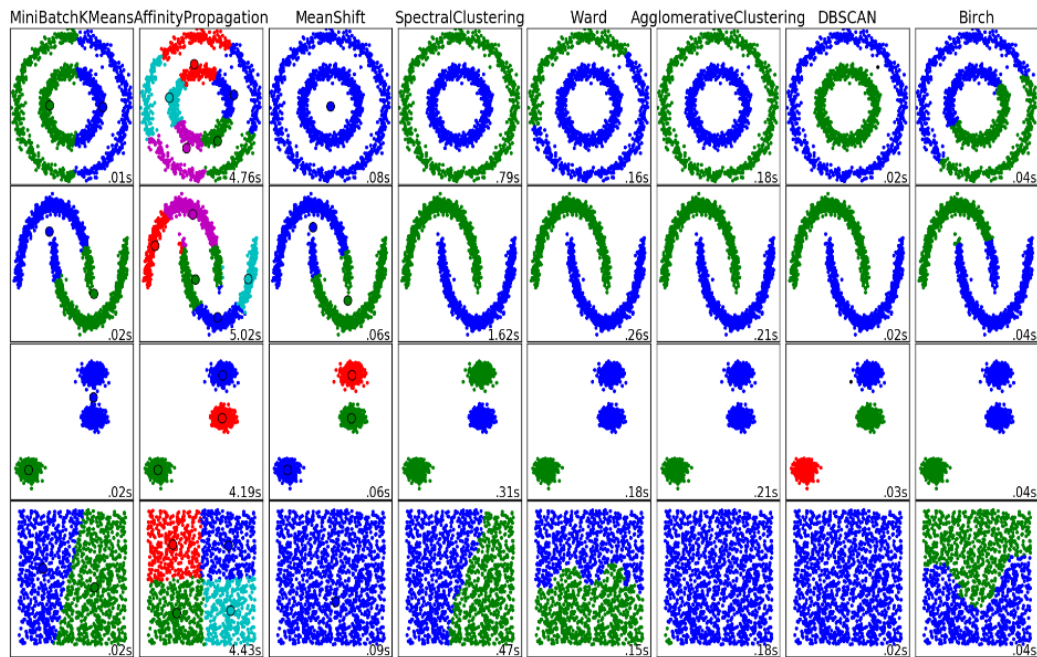
A comparison of the clustering algorithms in scikit-learn

QUESTIONS