Clustering Large Datasets into Meaningful Groups

Swetha Kolalapudi
CO-FOUNDER, LOONYCORN

www.loonycorn.com

Overview

Spot applications of clustering

Recognize the difference between Classification and Clustering

Understand how the K-Means Clustering algorithm works

Clustering

is a way to group items together based on some measure of similarity

Clustering

Let’s say we want to understand user behavior at a Social Network

Clustering

The objective is to divide all users into groups i.e. clusters

Clustering

Users in a group must be **“similar”** to one another

Maximize intracluster similarity

Users in different groups must be **“dissimilar”** to one another

Minimize intercluster similarity

Clustering

All users can be represented using some features

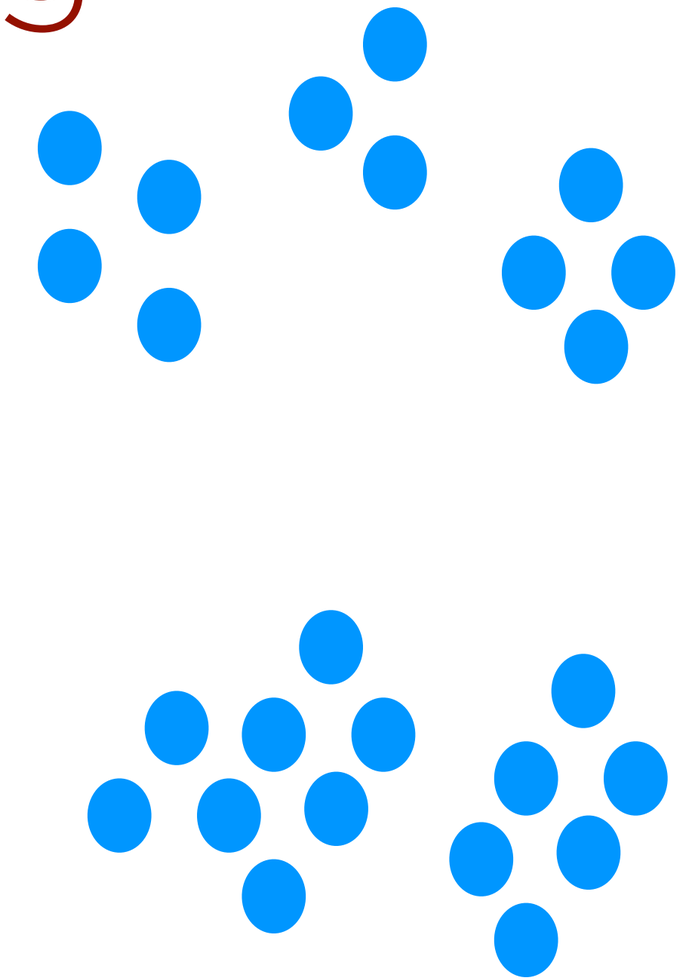
Age

Location

Frequency of usage for each topic

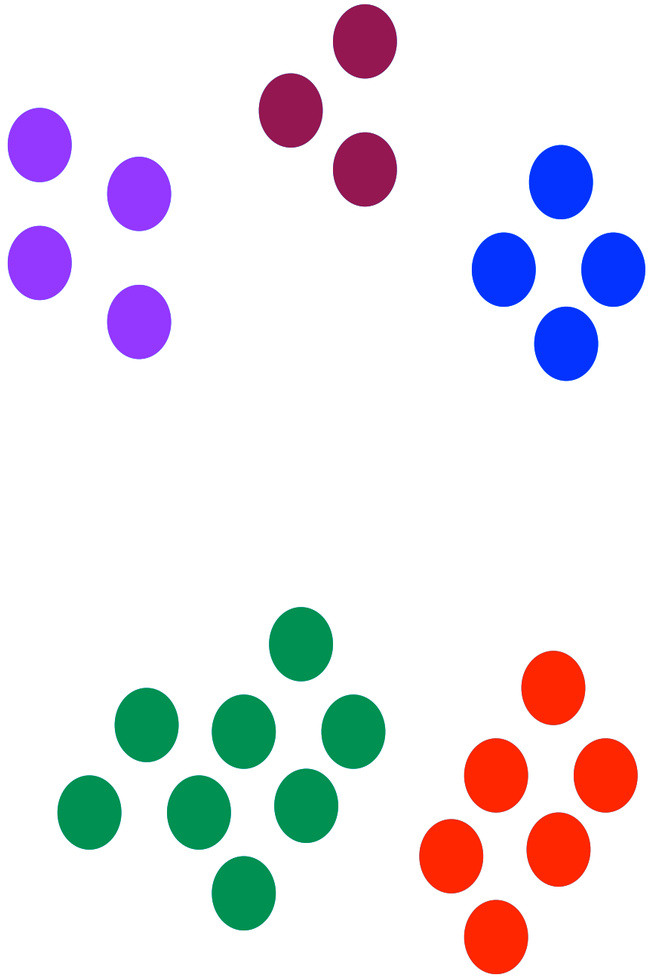
Clustering

Users represented using features can be seen as points in an N-Dimensional space



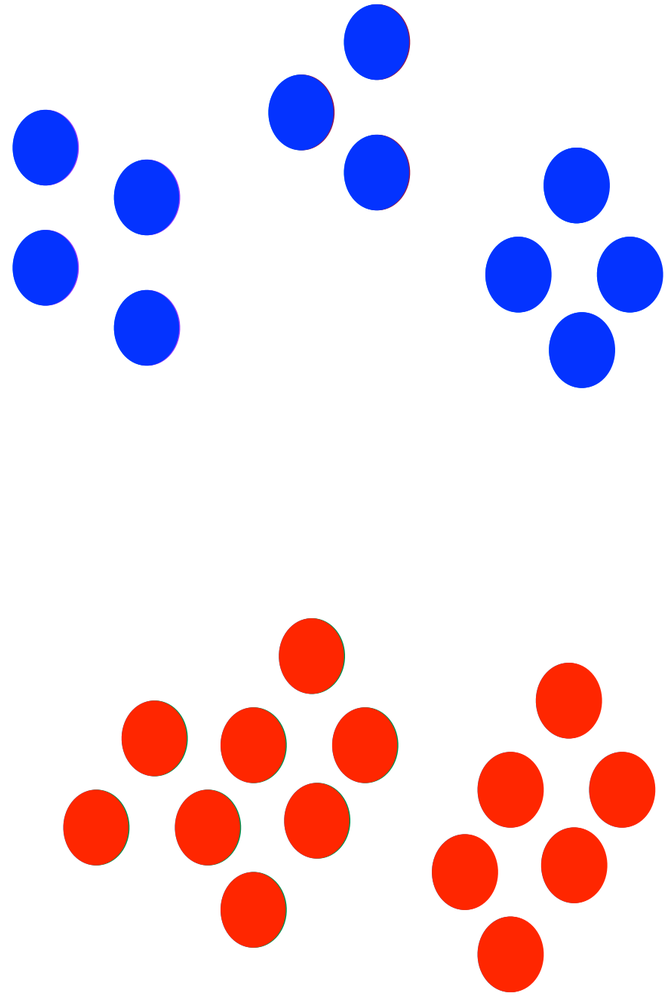
Clustering

Here is 1 way to divide the groups



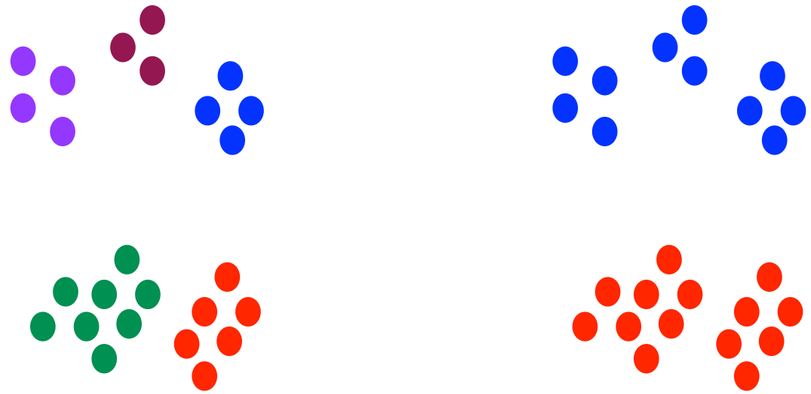
Clustering

Here's another



Clustering

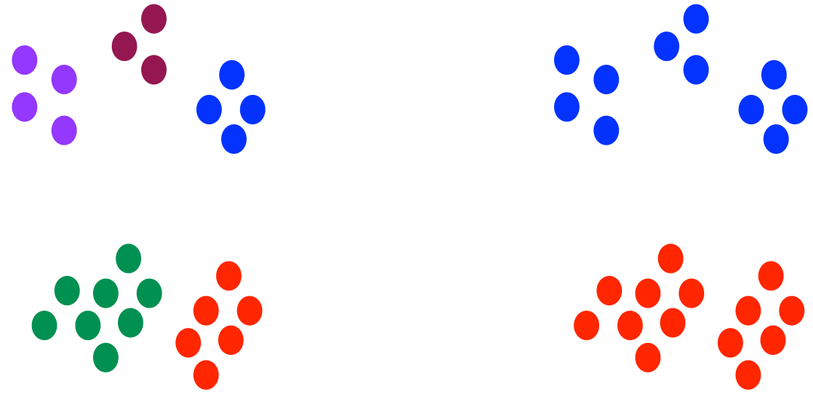
In both cases, **“Similarity”** is being measured based on the distance between users



Clustering

In real life, the “nearness” might translate to

1. Liking/following the same topics
2. Being in the same state
3. Being in the same age group
4. Or all of the above



Typical Clustering Setup

Dataset

The entire set of items which will be grouped

Features

Represent each datapoint using numeric attributes

Clustering

Use an algorithm to group the items

Features

Choose attributes relevant to the groups you are seeking

Features

To group users based on the similarity of usage patterns

Frequency of Morning Log in

Frequency of Evening Log in

Time spent per session

Features

To group users based on the similarity of likes/dislikes

Likes for each topic

Shares for each topic

Typical Clustering Setup

Dataset

The entire set of items which will be grouped

Features

Represent each datapoint using numeric attributes

Clustering

Use an algorithm to group the items

Typical Clustering Setup

Dataset

The entire set of items which will be grouped

Features

Represent each datapoint using numeric attributes

Clustering

Use an algorithm to group the items

Clustering

A few different clustering techniques

K-Means Clustering

Hierarchical Clustering

Density based Clustering

Distribution based Clustering

Classification vs Clustering

What's the Difference?

Classification

Classifying data into pre-defined categories

Clustering

Grouping data into a a set of categories

Classification

- Take one instance
- Classify it into a pre-defined category (labels)
- Do this based on training data which has already been classified

Classification

Is this e-mail **Spam** or **Ham**?

Is this tweet **positive** or **negative**?

Is this trading day an **up-day** or a **down-day**?

Clustering

- Take a large number of instances
- Divide them into groups
- The groups are unknown beforehand

Clustering

What kind of groups can these user be divided into?

What kind of themes are present in this set of articles?

Typical Classification Setup

Problem Statement

Define the problem statement

Features

Represent the training data and test data using numerical attributes

Training

“Train a model” using the training data

Test

“Test the model” using test data

Typical Clustering Setup

Dataset

The entire set of items which will be grouped

Features

Represent each datapoint using numeric attributes

Clustering

Use an algorithm to group the items

Comparison with Classification

Classification

Assigns a category to 1 new item, based on already labelled items

The categories/groups to be divided into are known beforehand

There are 2 phases, an explicit training phase, and then a test phase

Supervised Learning

Clustering

Takes a bunch of unlabelled items and divides them into categories

The categories/groups are unknown before hand

There is only 1 phase i.e. dividing of training data into Clusters

Unsupervised Learning

Supervised Learning

An explicit training phase

Requires a set of training data for which the output of the ML algorithm is known

Use when you are looking for a specific output

Unsupervised Learning

No training phase

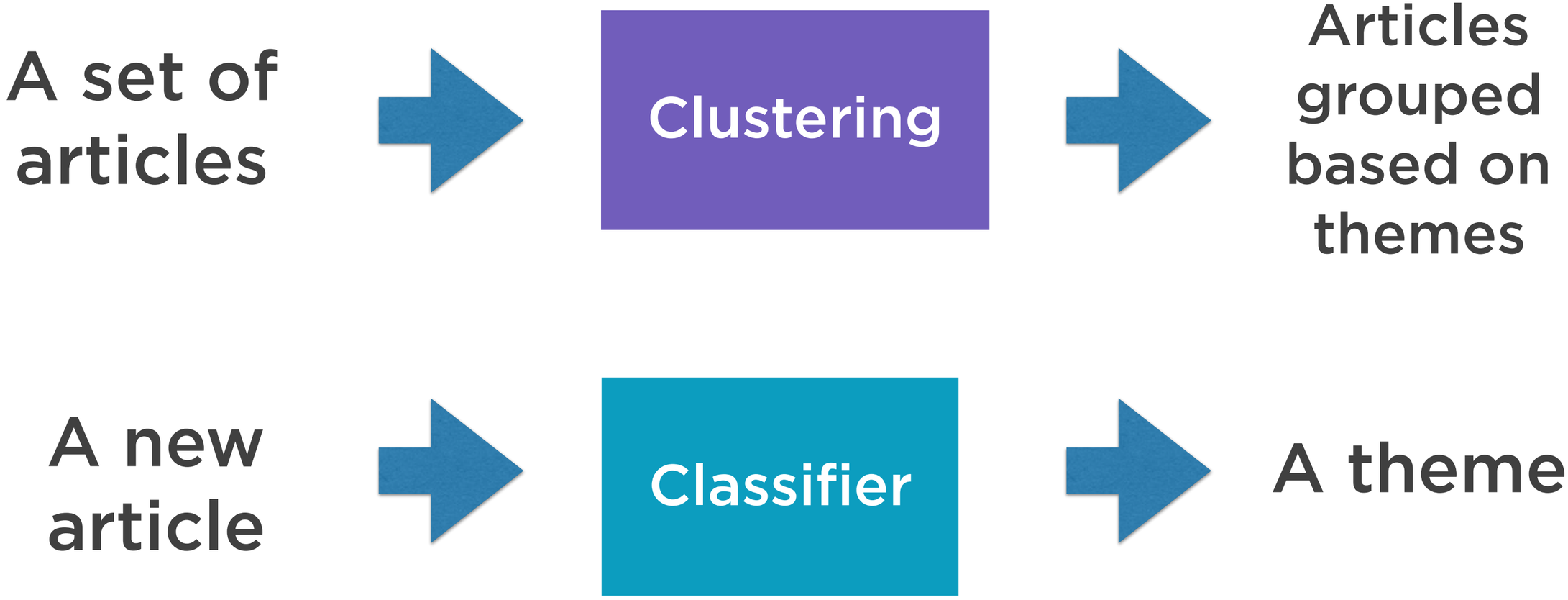
Requires a large set of data but the output you are seeking is unknown

Use when you are looking for interesting patterns that you didn't know existed

Clustering + Classification

Sometimes these techniques go hand in hand

Clustering + Classification



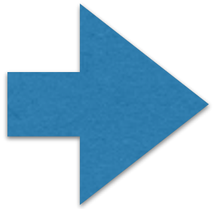
Clustering + Classification

Training data for the classifier

Articles grouped based on themes

A new article

Classifier



A theme

A few applications of clustering

A document is a piece of text

An article

A comment

A book

A webpage

A review

A tweet

Document Clustering

Given any set of documents

Group them based on the similarity of content

Document Clustering

Study the identified clusters

To find interesting themes

User Segmentation

Consider users of an online service

Group users based on the similarity of their behavior

Grouping Trading Days

Consider a quant trading scenario

Group days based on the similarity of stock behavior on that day

Applications of Clustering

User Segmentation

News article clustering based on topics

Grouping trading days based on similarity of stock movements

Document Clustering

Group documents together to see if any interesting themes emerge

K-Means Clustering

Typical Clustering Setup

Dataset

The entire set of items which will be grouped

Features

Represent each datapoint using numeric attributes

Clustering

Use an algorithm to group the items

Typical Clustering Setup

Dataset

The entire set of items which will be grouped

Features

Represent each datapoint using numeric attributes

Clustering

Use an algorithm to group the items

Representing Text Using Features

Term Frequency Representation

Term Frequency Representation

Hello, this is a test

(hello, this, is, the, universe, of, all, words, in, any, text, a, an, test, goodbye)

(1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0)

In this method, we represent each text using the frequencies of words (or) terms

Term Frequency

Some words characterize a document more than others

The house was in New York

Term Frequency

The **house** was in **New York**

Words which occur more rarely, clearly differentiate a document from other documents

Term Frequency

The house was in New York

Words which are very common don’t do much to differentiate a document

Term Frequency - Inverse Document Frequency

Weight the term frequencies to take the rarity of a word into account

(hello, this, is, the, universe, of, all, words, in, any, text, a, an, test, goodbye)

$$\text{Weight} = \frac{1}{\#\text{ documents the word appears in}}$$

Term Frequency - Inverse Document Frequency

$$\text{Weight} = \frac{1}{\#\text{ documents the word appears in}}$$

(hello, this, is, the, universe, of, all, words, in, any, text, a, an, test, goodbye)

TF-IDF

Typical Clustering Setup

Dataset

The entire set of items which will be grouped

Features

Represent each datapoint using numeric attributes

Clustering

Use an algorithm to group the items

Typical Clustering Setup

Dataset

The entire set of items which will be grouped

Features

Represent each datapoint using numeric attributes

Clustering

Use an algorithm to group the items

K-Means Clustering

K-Means Clustering

Documents are represented using TF-IDF

Each document is a tuple of N Numbers

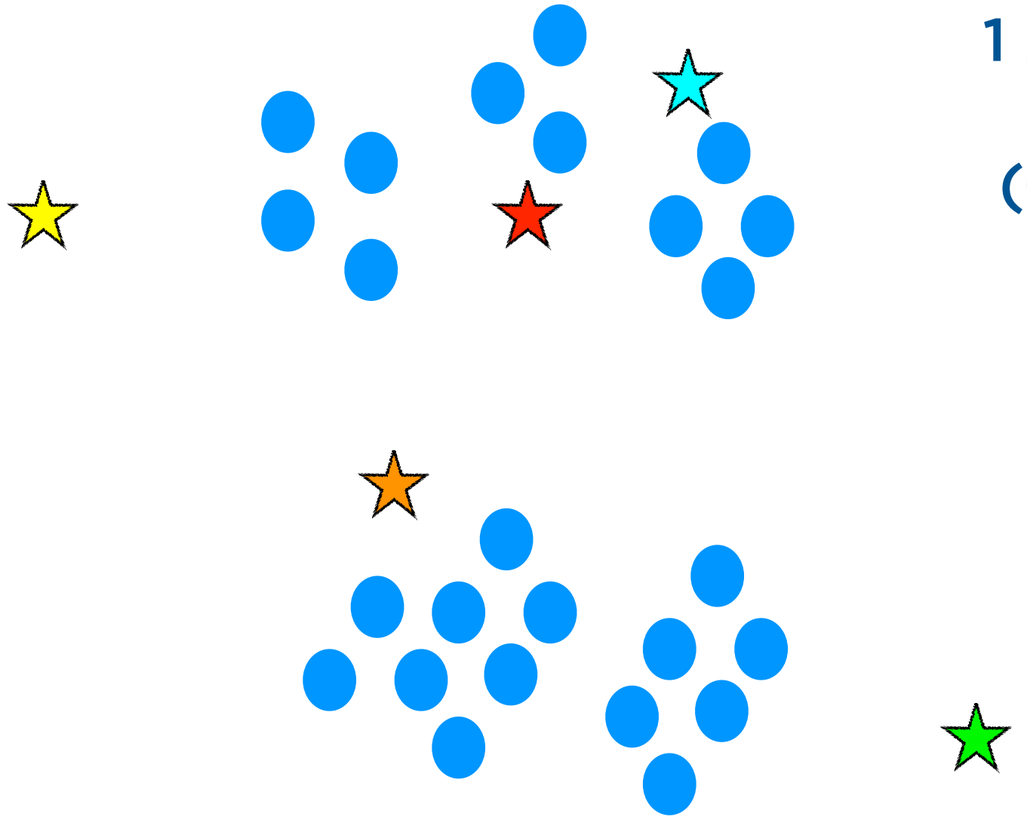
N is the total number of distinct words in all documents

K-Means Clustering

A tuple of N Numbers

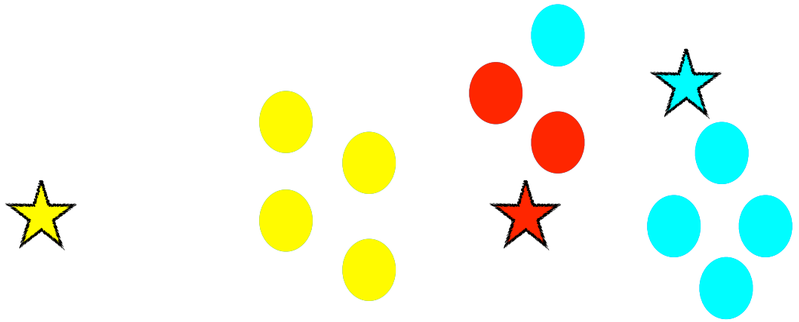
A point in an N-Dimensional Hypercube

K-Means Clustering

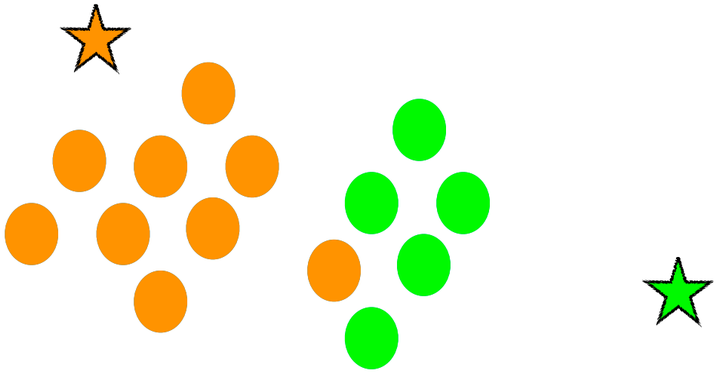


1 . Initialize a set of points as the “K” Means (Centroids of the clusters you want to find)

K-Means Clustering

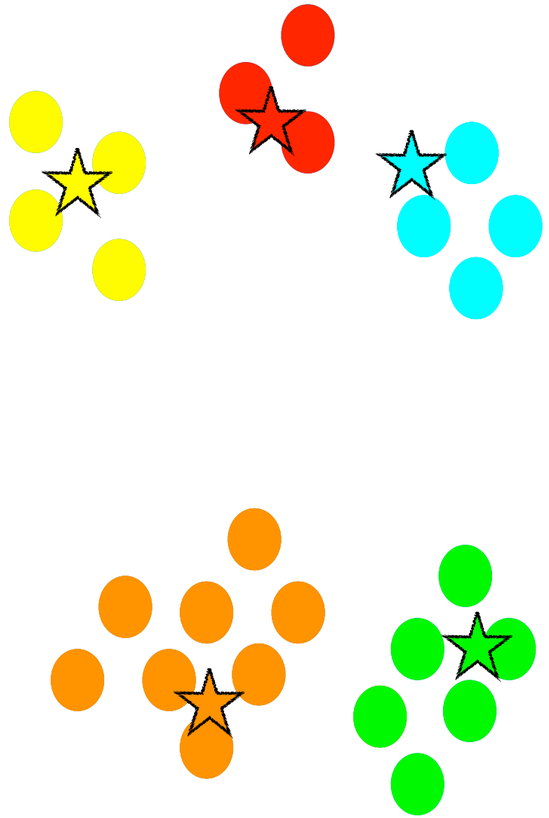


2. Assign each point to the cluster belonging to the nearest mean



3. Find the new means/ centroids of the clusters

Convergence



2. Assign each point to the cluster belonging to the nearest mean

3. Find the new means/ centroids of the clusters

Rinse and repeat steps 2,3 until the means don't change anymore

The dataset for this demo is from the UCI Machine Learning Repository

Sentiment Labelled Sentences Data Set

Download: Data Folder, Data Set Description

Abstract: The dataset contains sentences labelled with positive or negative sentiment.

https://archive.ics.uci.edu/ml/datasets/Sentiment+Labelled+Sentences

Each line in the data set

review	label
Wasted two hours.	0

Demo

Implement K-Means Clustering on IMDB reviews

Summary

Spot applications of clustering

Recognize the difference between Classification and Clustering

Understand how the K-Means Clustering algorithm works