K-Means Clustering

Ishwar Suriyaprakash

Problem Statement

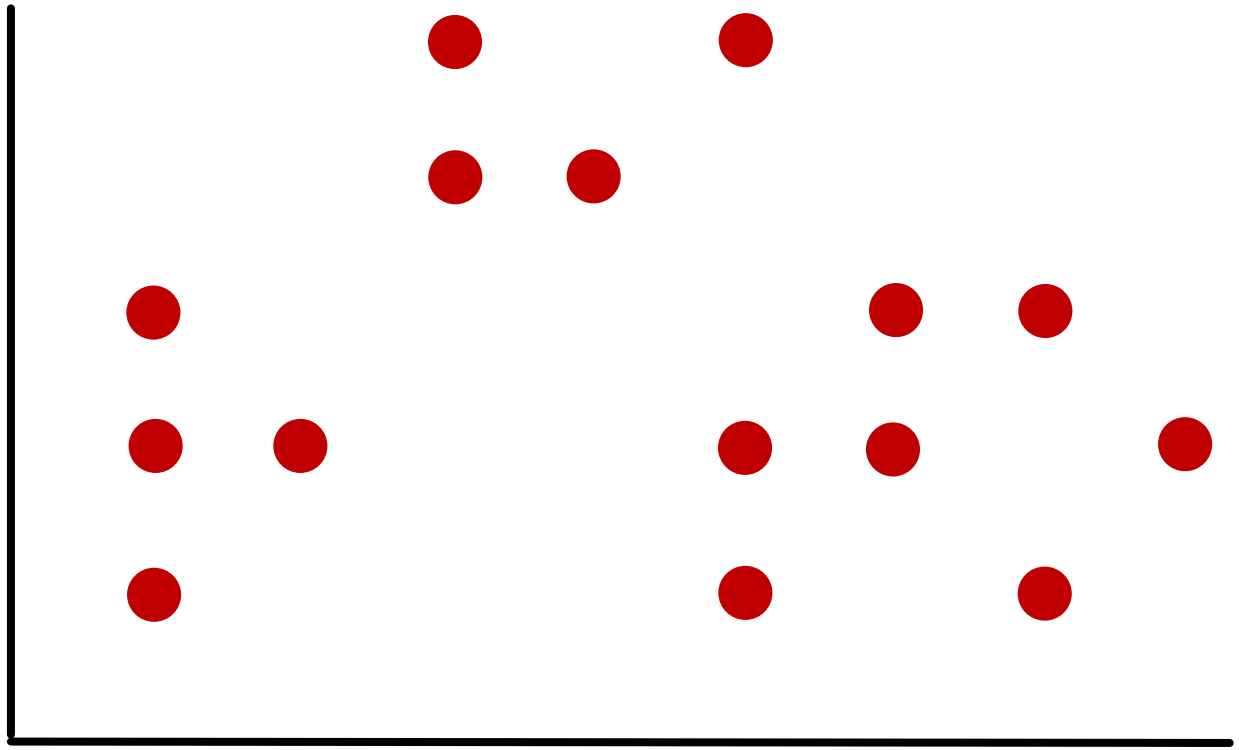
Given

A set S of objects each of which have different degrees of a set of qualities, Q

Goal

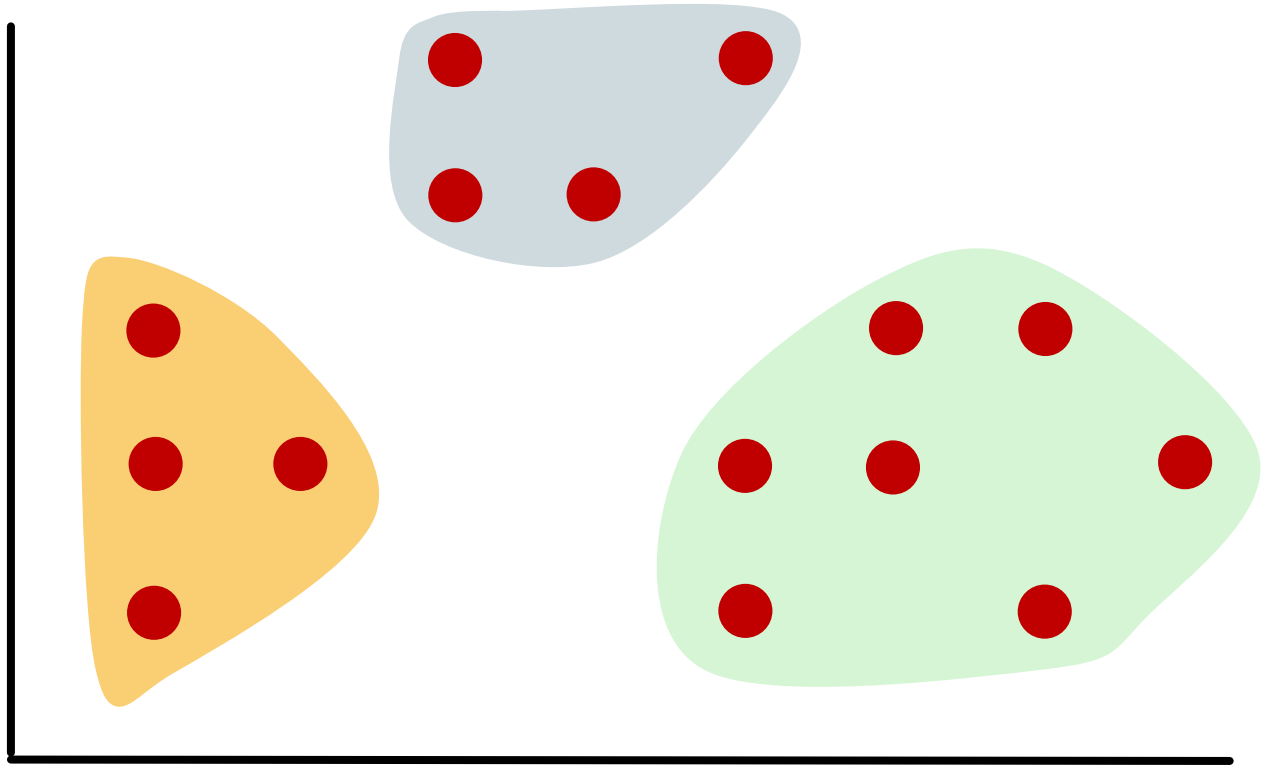
Partition S into groups such that objects with ‘similar’ qualities belong to the same group

Example: Points in space



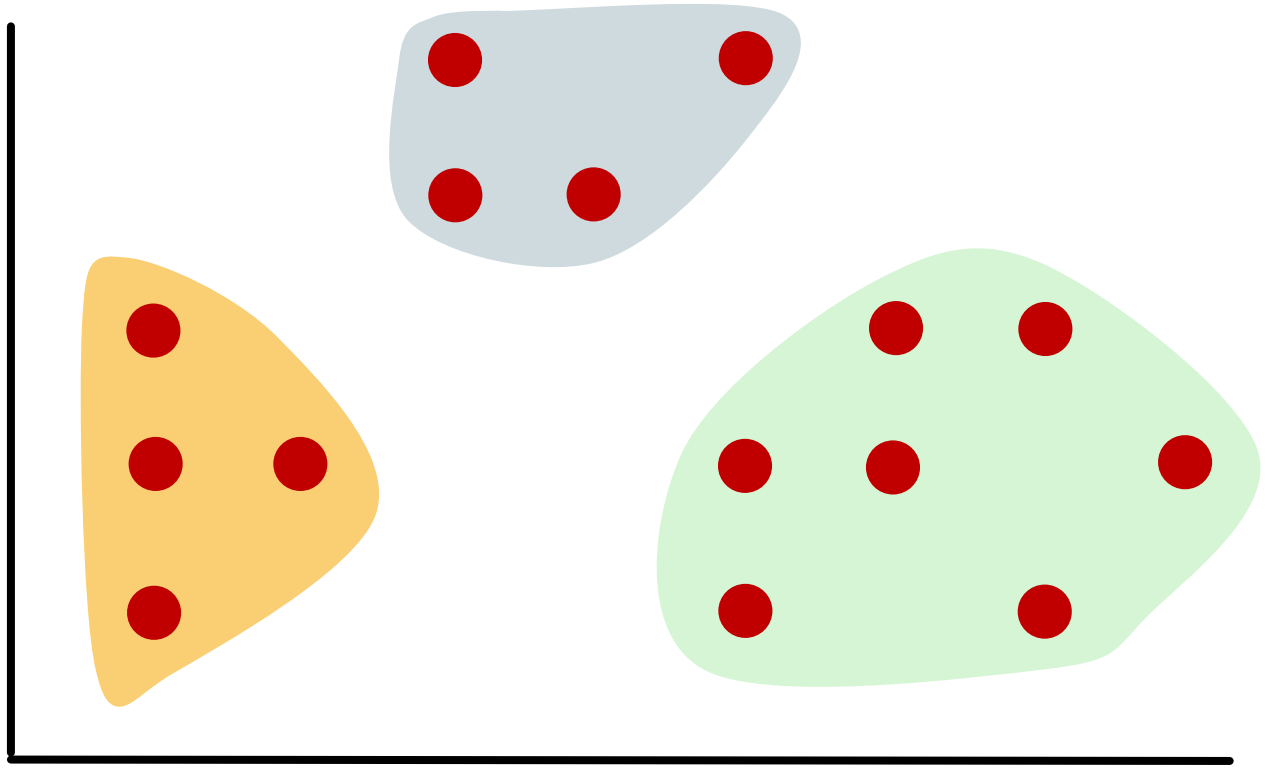
How can we group, or cluster, the points above?

Example: Points in space



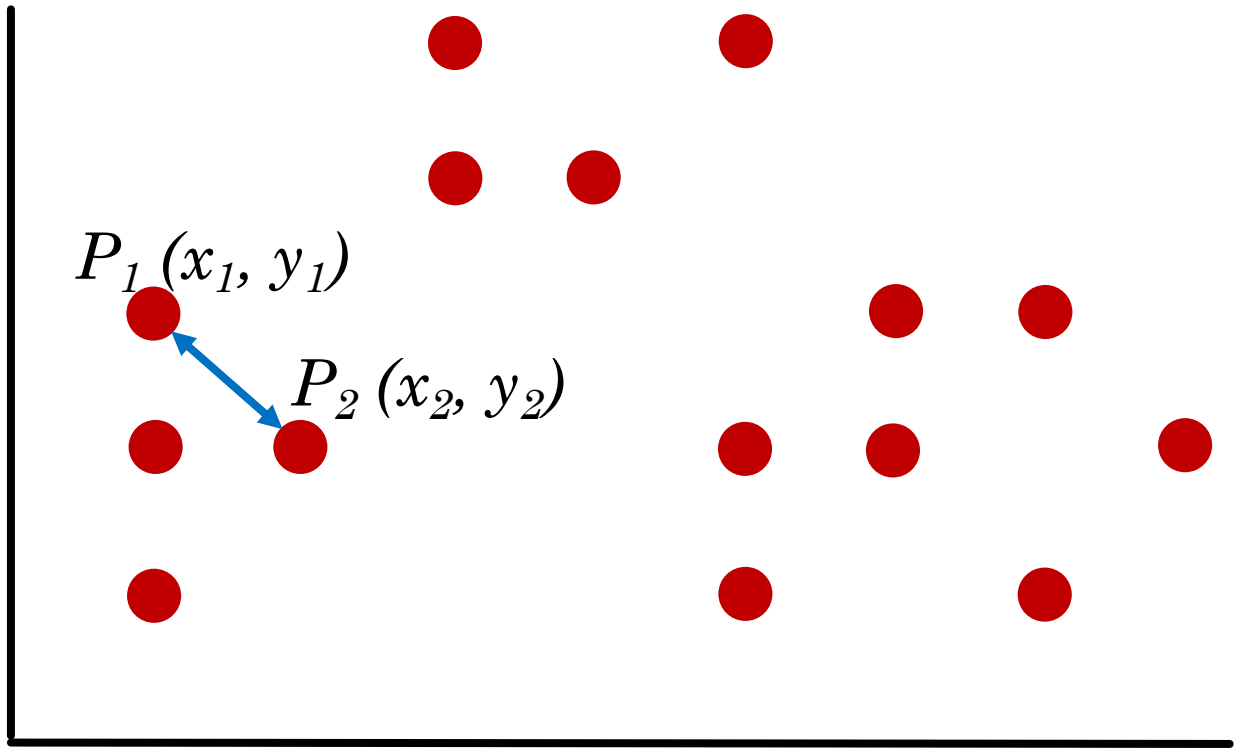
One possible way is shown above.

Example: Points in space



Points are ‘closer’ to their own group’s ‘centroid’ than to the centroid of another group

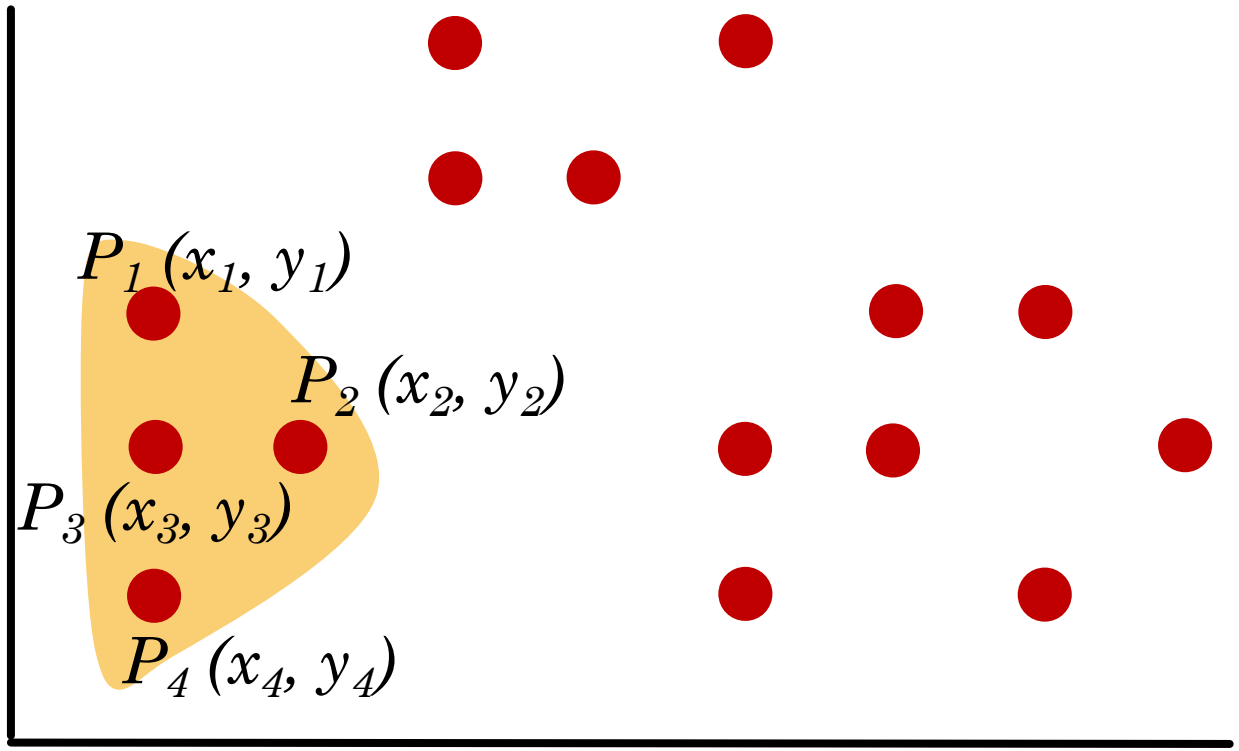
Distance & Centroid



Distance measures closeness of points

Example: $D(P_1, P_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$

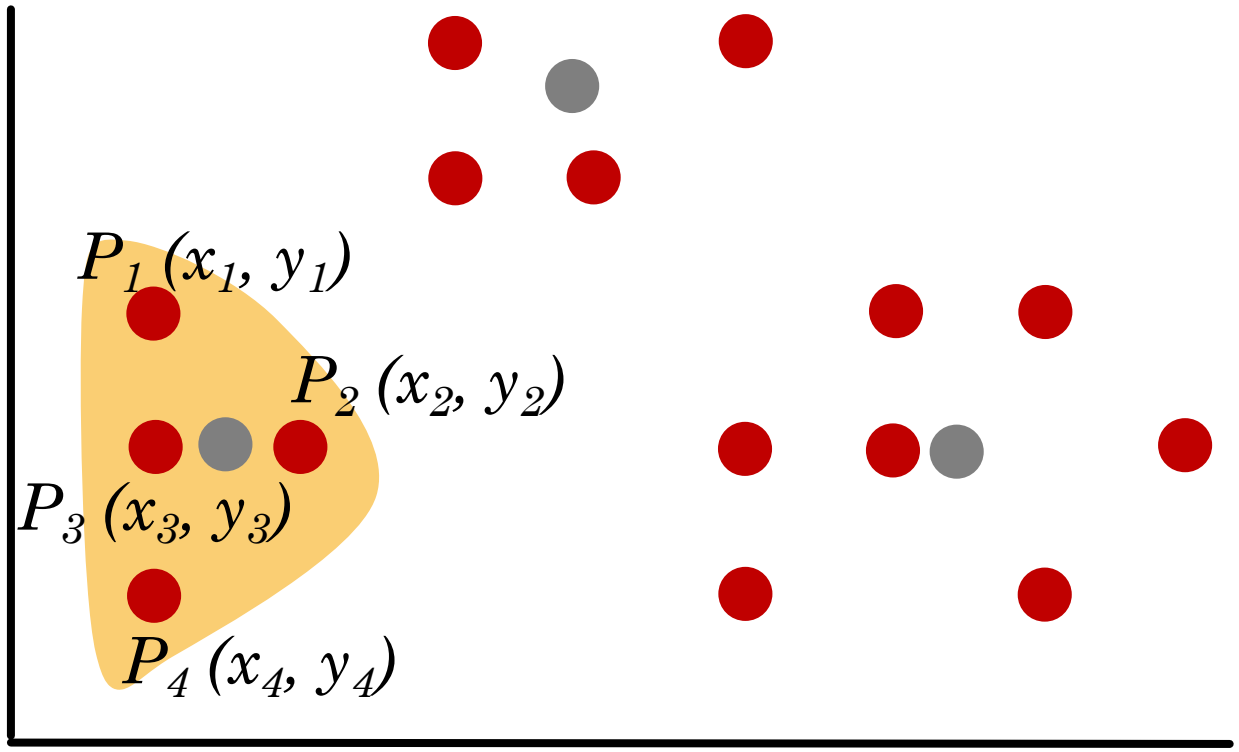
Distance & Centroid



Centroid is a location indicative of center of mass

$$C(P_1, P_2, P_3, P_4) = \left(\frac{(x_1+x_2+x_3+x_4)}{4}, \frac{(y_1+y_2+y_3+y_4)}{4}\right)$$

K-Means Clustering



K-Means clustering is an *iterative* algorithm

Minimizes distance of points to cluster centroids

K-Means Clustering

Inputs

1. Set of points (S) and their coordinates
2. Number of clusters (K)

Algorithm

1. Select K random points to be initial cluster centroids
2. Iterate steps below until centroids don't change
    1. Compute distance of each point to each centroid
    2. Assign each point to the closest cluster
    3. Compute new centroids for each cluster

Output: Partition of S into K sets

K-Means Clustering

Demo *(courtesy of Naftali Harris)*

K-Means Clustering

Can be extended to m points in n-dimension space

$$S = \{P_1(x_{11}, x_{12}, \dots, x_{1n}), P_2(x_{21}, x_{22}, \dots, x_{2n}), \dots, P_m(x_{m1}, x_{m2}, \dots, x_{mn})\}$$

$$Distance\ D(P_i, P_j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{in} - x_{jn})^2}$$

$$Centroid\ C(P_1, P_2, \dots, P_r)$$
$$= \left(\frac{(x_{11} + x_{21} + \dots + x_{r1})}{r}, \frac{(x_{12} + x_{22} + \dots + x_{r2})}{r}, \dots, \frac{(x_{1n} + x_{2n} + \dots + x_{rn})}{r}\right)$$

K-Means Clustering

Can be extended to m points in n-dimension space

Can be extended to m objects with n features

K-Means Clustering

Can be extended to m points in n-dimension space

Can be extended to m objects with n features

- Objects are represented by points
- Features are quantified using coordinates
- Distance between two points is an attempt to measure the similarity between corresponding objects

K-Means Clustering

Can be extended to m points in n-dimension space

Can be extended to m objects with n features

- Objects are represented by points
- Features are represented by coordinates
- Distance between two points is an attempt to measure the similarity between corresponding objects

- m movies with n features (e.g. genre, age limit)

K-Means Clustering

Can be extended to m points in n-dimension space

Can be extended to m objects with n features

- Objects are represented by points
- Features are represented by coordinates
- Distance between two points is an attempt to measure the similarity between corresponding objects

- m movies with n features (e.g. genre, age limit)
- m people with n features

K-Means Clustering

Fun Activity

Club members: Objects
Likes & dislikes: Features

	Sports	Humanities	STEM	Social Media
Member 1	-5	4	0	1
Member 2	0	1	-2	3
Member 3	3	-1	5	2

K-Means Clustering

Thank you! Questions?