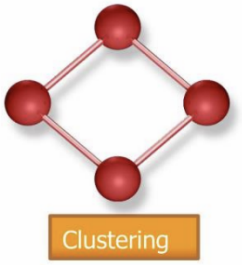
What is Clustering?



Organizing data into *clusters* such that there is:

- ✓ High intra-cluster similarity
- ✓ Low inter-cluster similarity
- ✓ Informally, finding natural groupings among objects.

Why do we want to do it??

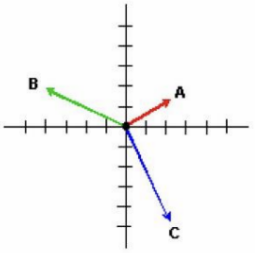
Why Clustering?

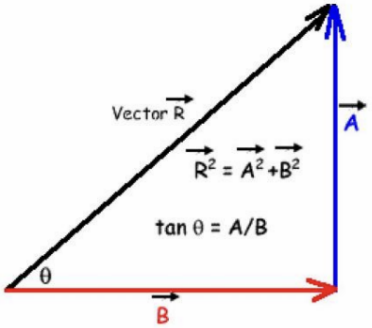
- ✓ Organizing data into clusters shows internal structure of the data
  Ex. Clusty and clustering genes
- ✓ Sometimes the partitioning is the goal
  Ex. Market segmentation
- ✓ Prepare for other AI techniques
  Ex. Summarize news (cluster and then find centroid)
- ✓ Techniques for clustering is useful in knowledge
- ✓ Discovery in data
  Ex. Underlying rules, reoccurring patterns, topics, etc.

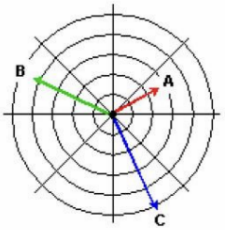
Vector

A **vector** is a quantity or phenomenon that has two independent properties: magnitude and direction.

The term also denotes the mathematical or geometrical representation of such a quantity.

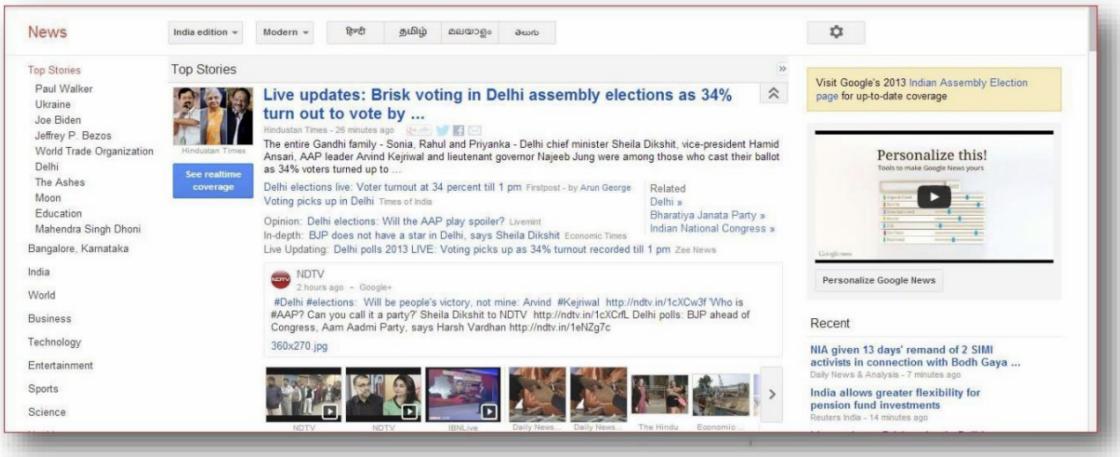




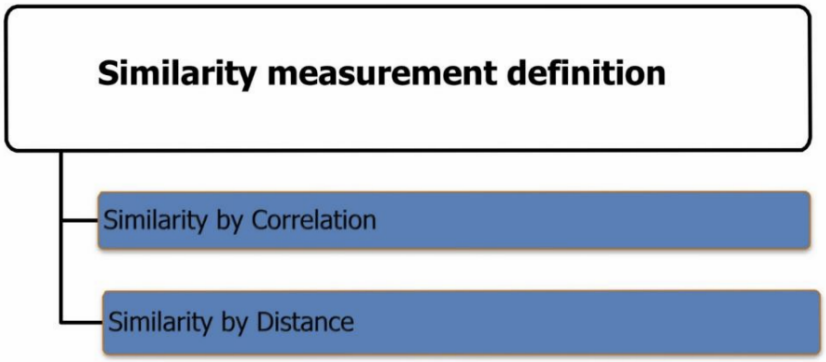


Clustering - Example

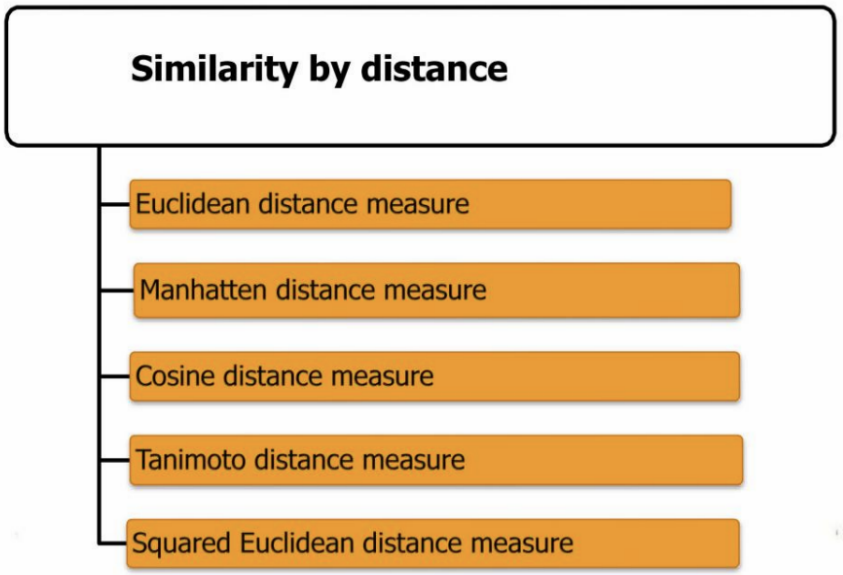
A sample news grouping from Google News:



Similarity Measurement



Distance measures



Euclidean distance measure

Mathematically, Euclidean distance between two n-dimensional vectors

(a1, a2, ... , an) and (b1,b2,...,bn) is:

$$d = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + ... + (a_n - b_n)^2}$$

Manhatten distance measure

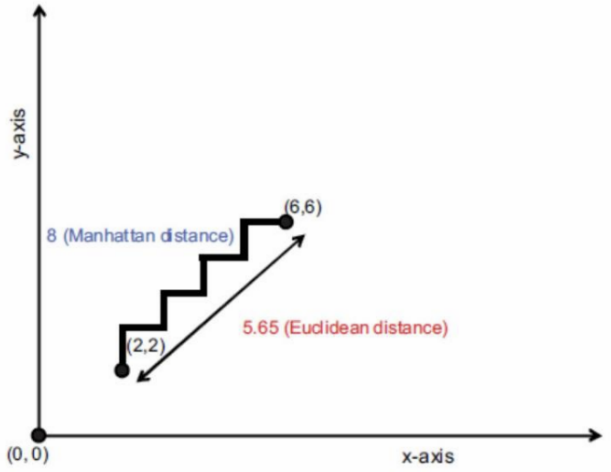
Mathematically, the Manhattan distance between two n-dimensional vectors

(a1, a2, ... , an) and (b1, b2, ... , bn) is

$$d = |a1 - b1| + |a2 - b2| + ... + |an - bn|$$

Difference between Euclidean and Manhattan

From this image we can say that, The Euclidean distance measure gives 5.65 as the distance between (2, 2) and (6, 6) whereas the Manhattan distance is 8.0



Cosine distance measure

The formula for the cosine distance between *n*-dimensional vectors (a1, a2, ... , an) and (b1, b2, ...,bn) is

$$d = 1 - \frac{(a_1 b_1 + a_2 b_2 + \ldots + a_n b_n)}{(\sqrt{(a_1^{\,2} + a_2^{\,2} + \ldots + a_n^{\,2})}\sqrt{(b_1^{\,2} + b_2^{\,2} + \ldots + b_n^{\,2})})}$$

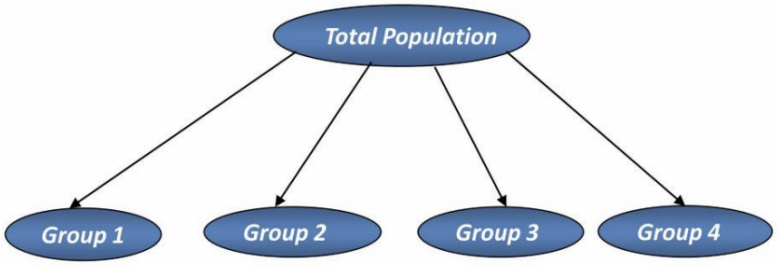
Tanimoto distance measure

The formula for the Tanimoto distance between two *n*-dimensional vectors (a1, a2, ... , an) and (b1, b2, ... , bn) is

$$d = 1 - \frac{(a_1 b_1 + a_2 b_2 + \ldots + a_n b_n)}{\sqrt{(a_1^2 + a_2^2 + \ldots + a_n^2)} + \sqrt{(b_1^2 + b_2^2 + \ldots + b_n^2)} - (a_1 b_1 + a_2 b_2 + \ldots + a_n b_n)}$$

K-Means clustering

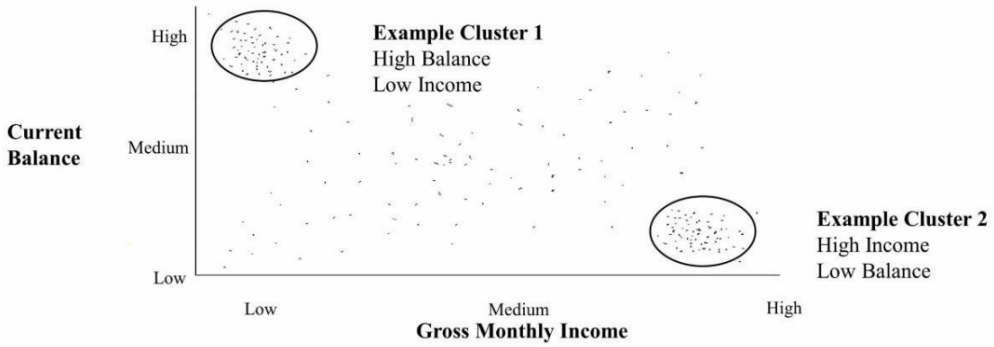
- ✓ The process by which objects are classified into a number of groups so **that they are as much dissimilar as possible from one group to another group, but as much similar as possible within each group.**
- ✓ In other words Cluster analysis means dividing the whole population into groups which are distinct between themselves but internally similar.



- ✓ The objects in group 1 should be as similar as possible.
- ✓ But there should be much difference between an object in group 1 and group 2.
- ✓ The attributes of the objects are allowed to determine which objects should be grouped together.

K-Means clustering

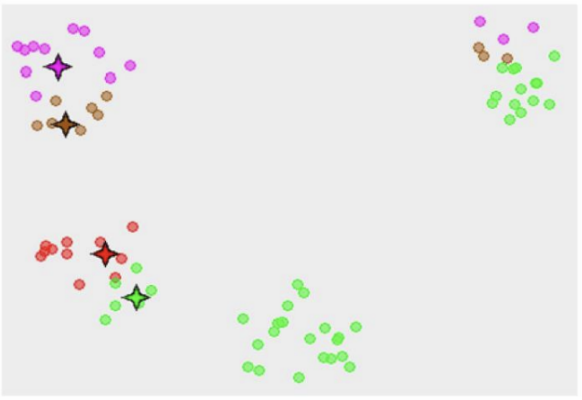
Basic Concepts of Cluster Analysis using Two variables



- ✓ Cluster 1 and Cluster 2 are being differentiated by Income and Current Balance.
- ✓ The objects in Cluster 1 have similar characteristics (High Income and Low balance), on the other hand the objects in Cluster 2 have the same characteristic (High Balance and Low Income).
- ✓ But there are much differences between an object in Cluster 1 and an object in Cluster 2

K-Means clustering steps

1. k initial "means" (in this case k=3) are randomly generated within the data domain.
2. k clusters are created by associating every observation with the nearest mean.
3. The centroid of each of the k clusters becomes the new mean.
4. Steps 2 and 3 are repeated until convergence has been reached.



Step by Step pictorial representation of K-Means clustering

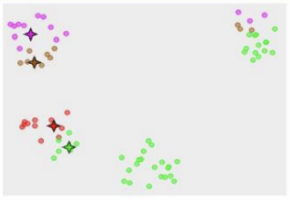


figure-1

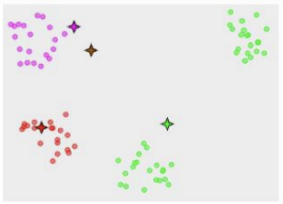


figure-2

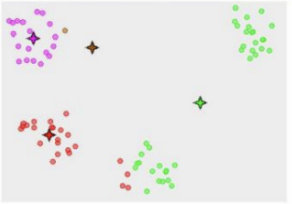


figure-3

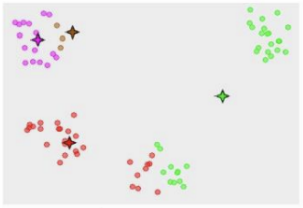


figure-4

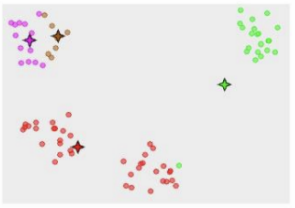


figure-5

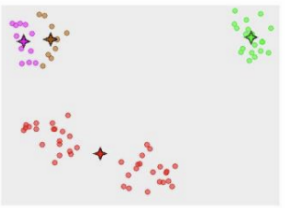


figure-6

- ✓ The small circles are the data points, the four ray stars are the centroids (means).
- ✓ The initial configuration is on the figure-1.
- ✓ The algorithm converges after five iterations presented on the figures, from figure-2 to figure-6.