Clustering Algorithms

R. Srivatsan, IBAB

What is clustering?

“A method by which we can organize objects into groups whose members share certain similarities.”

After clustering, the objects in a cluster are similar between them, and dissimilar with objects of other clusters.

- Clustering is an unsupervised learning problem.

- It involved finding a structure in an unlabelled data.

Supervised and Unsupervised learning

In supervised learning, one set of observations (called inputs) are connected to another set of observations (called outputs) through a relationship (called model).

Example: reaction rate as a function of concentration.

Rate = f(concentration, a,b)

where **a** and **b** are mediating variables.

In unsupervised learning, all observations are assumed to be caused by some latent (hidden) variables, with no assumed relationships between the variables.

Example : blood pressure values measure for 200 persons.

Two basic types of clustering

Look at the following cities as a taxi driver:

- Mysore, Bellary, Bangalore, Tumkur, Delhi, Luknow, Srinagar, Chandigar,

This is a distance based clustering

- Look at the following list as a zoologist:

Tiger, Lion, Cheetah, Cow, Goat, Deer, Crow, Parrot

This is a conceptual clustering, based on similarities between objects.

Goals of clustering

- Finding representatives for homogeneous groups - data reduction
- Finding natural clusters and explaining them – natural data types
- Finding useful groups suitable for a particular problem in hand – useful data classes
- Finding data groups that stand out - outlier detection

Clustering is used in every field - marketing, biology, astronomy, libraries, psychology, ...

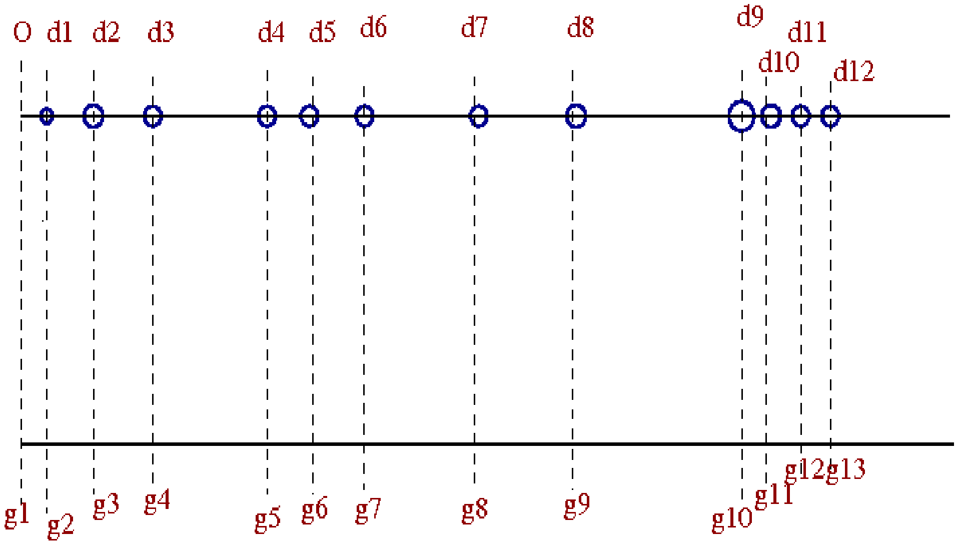
Distance based clustering

Basic types of clustering algorithms:

- Exclusive clustering : Data point assigned to a cluster cannot be assigned to another cluster (e g. K-means clustering)
- Overlapping clustering : has a fuzzy sets of clustered data, where a data point may belong to different clusters. (fuzzy C-means)
- Hierarchical clustering: Data points form a hierarchy of clusters, like a tree branch structure.
- Probabilistic clustering algorithms: (eg. Mixture of Gaussians)

- We will learn K-means and Hierarchical clustering, most popular in bioinformatics.

The concept of 'distance' in clustering



O is the reference point (origin)

- d1, d2, d3,... are distances of objects in meters from O.

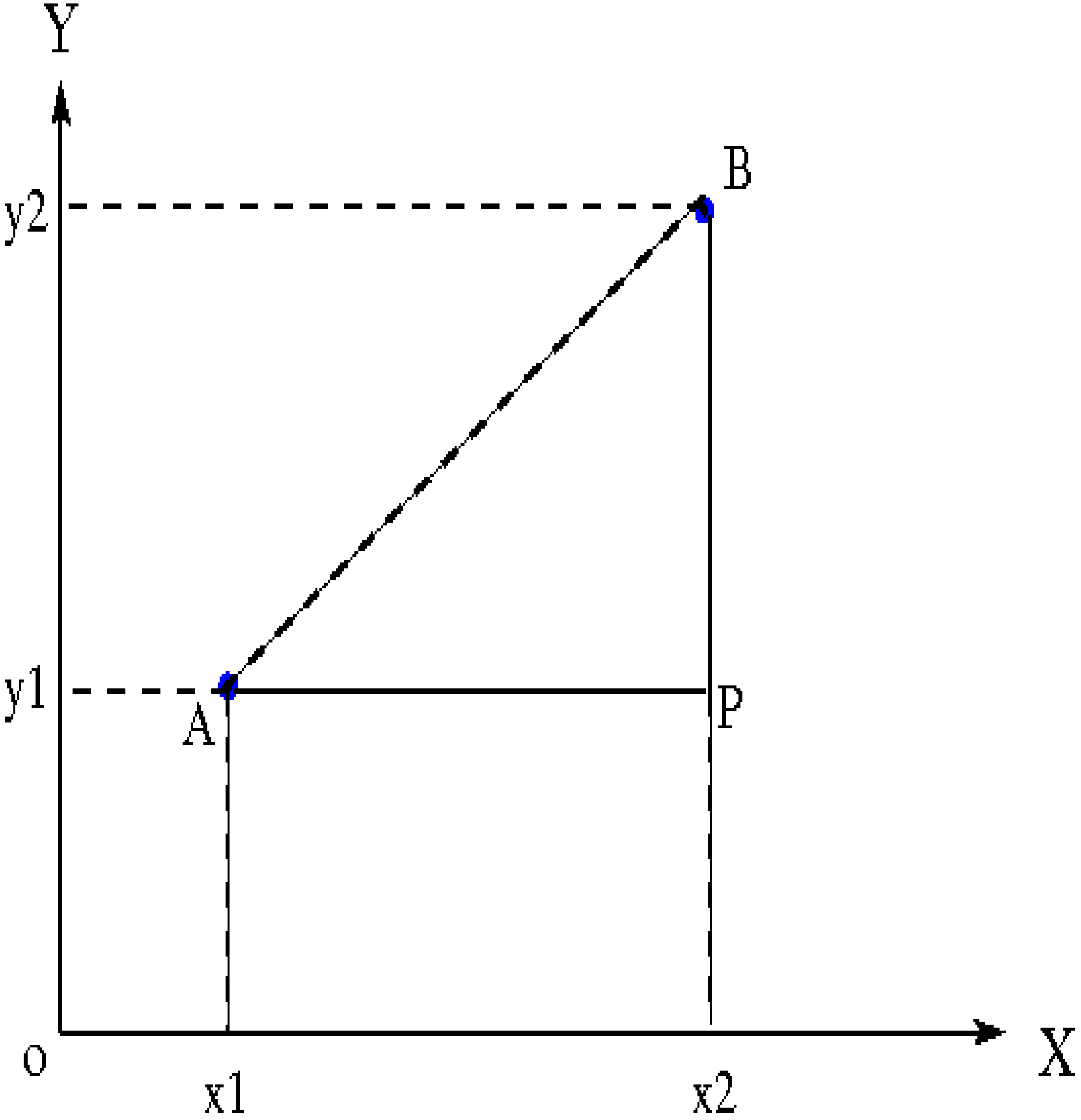
d1=3m, d2=5m, d3=7m, d4=13m etc.

- g1, g2,... can be gene expression levels in a microarray experiment.

g1=122.5, g2=145.5, g3=162.3, g4=220.5, g5=234.6, etc.

- Any quantity is “mapped” as a distance from an origin.
- Clustering algorithms look for “nearness” in this distance space.

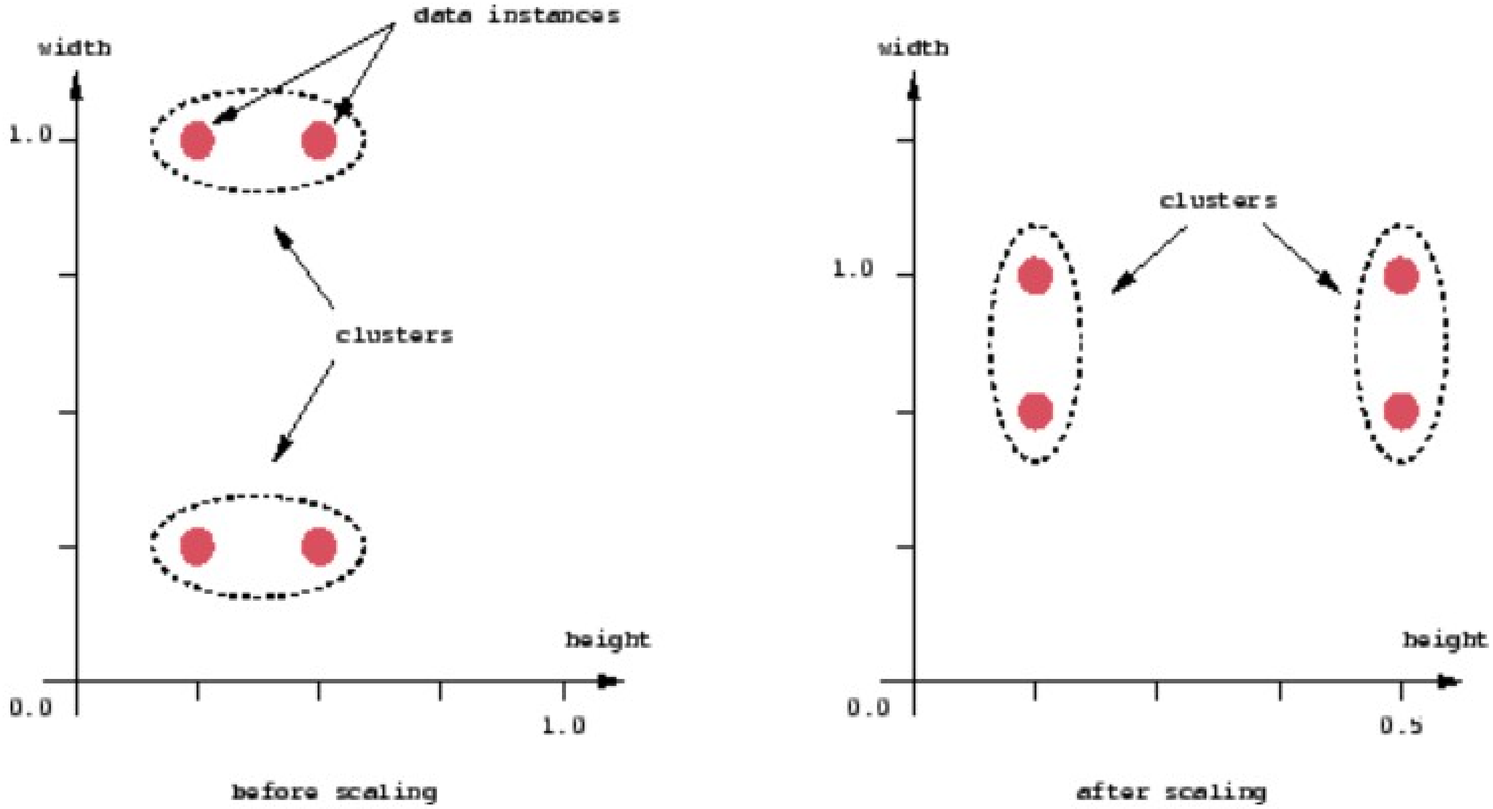
Distance measure - Euclidean distance



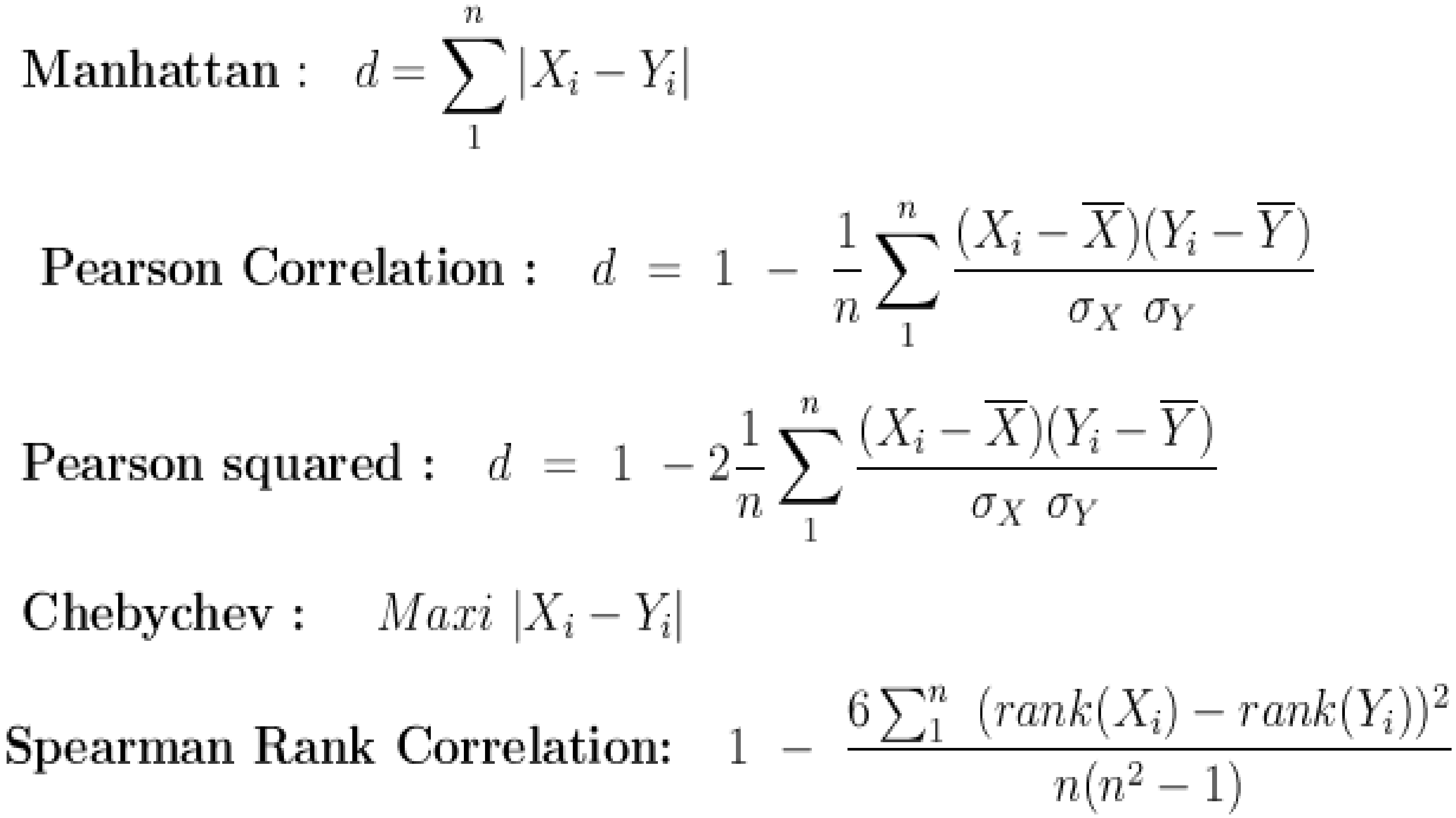
$$
\begin{aligned}
d &= \sqrt{AP^2 + PB^2} \\
&= \sqrt{(x2 - x1)^2 + (y2 - y1)^2}
\end{aligned}
$$

The Euclidean distance works fine if both (X,Y) axes are in the same units. Even then, scaling can affect the clustering. See below:

(http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/)



Important distance measures used in clustering



Manhattan : $d = \sum_{1}^{n} |X_i - Y_i|$

Pearson Correlation : $d \; = \; 1 \; - \; \frac{1}{n} \sum_{1}^{n} \frac{(X_i - \overline{X})(Y_i - \overline{Y})}{\sigma_X \; \sigma_Y}$

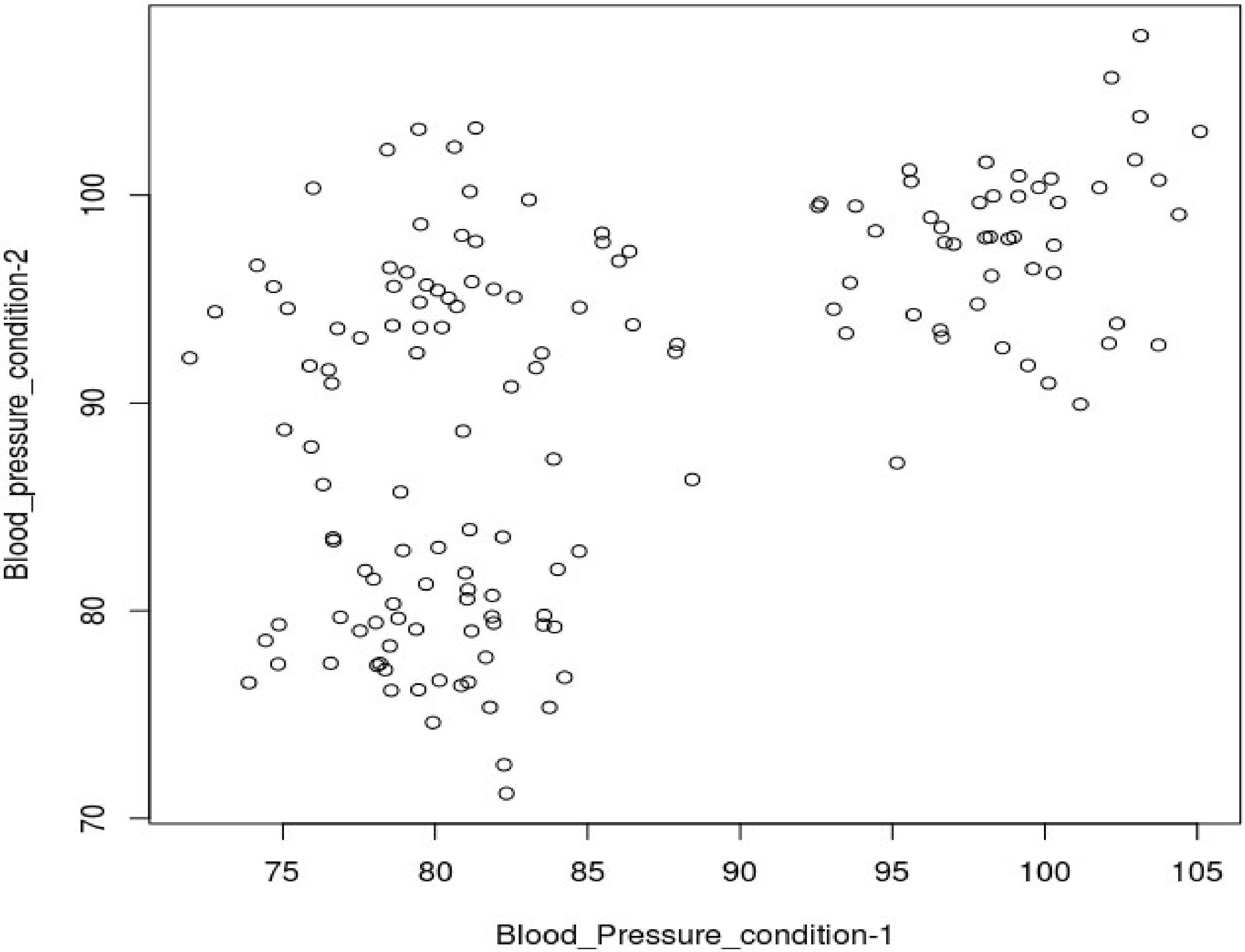
Pearson squared : $d \; = \; 1 \; - 2\frac{1}{n} \sum_{1}^{n} \frac{(X_i - \overline{X})(Y_i - \overline{Y})}{\sigma_X \; \sigma_Y}$

Chebychev : $Maxi \; |X_i - Y_i|$

Spearman Rank Correlation: $1 \; - \; \frac{6 \sum_{1}^{n} \; (rank(X_i) - rank(Y_i))^2}{n(n^2 - 1)}$

K-means clustering

Data : Blood pressures measured on each patient under two conditions



K-means clustering Algorithm
(MacQueen, 1967)

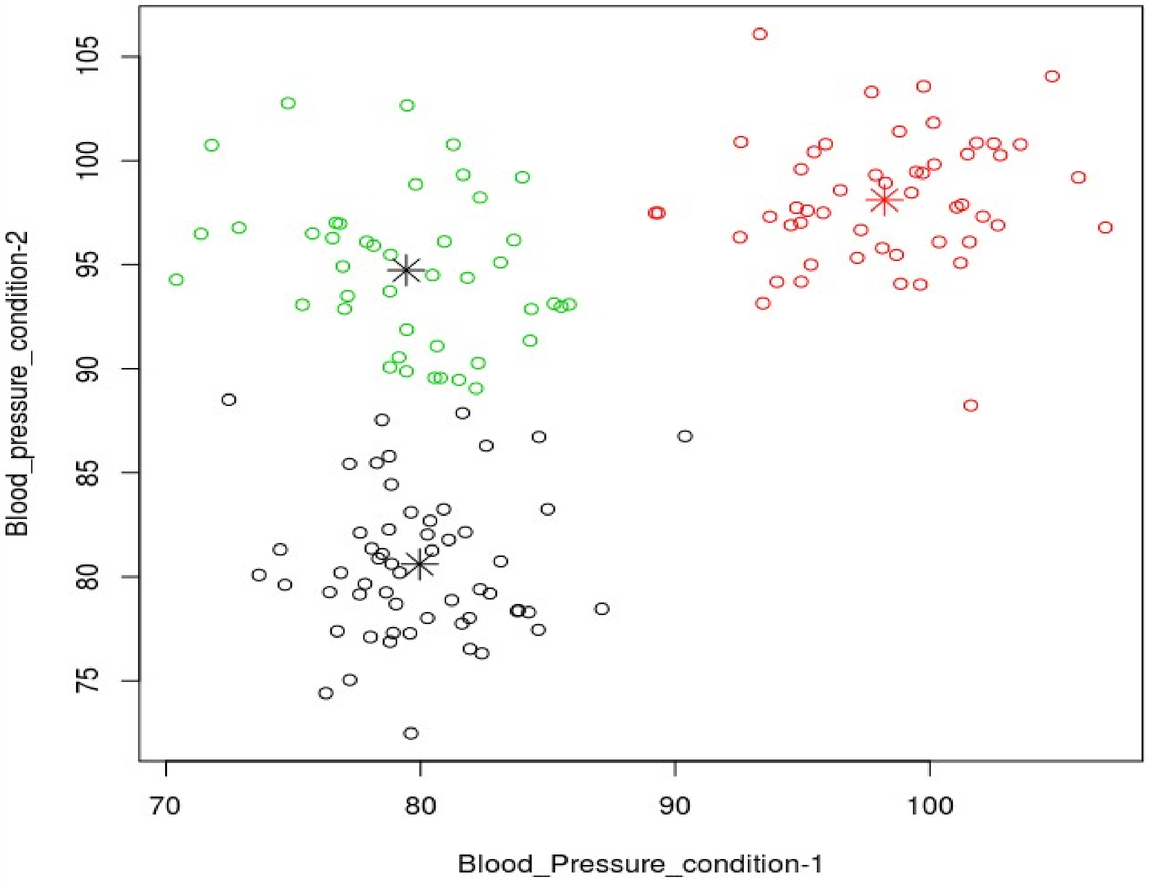
1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.

2. Assign each object to the group that has the closest centroid.

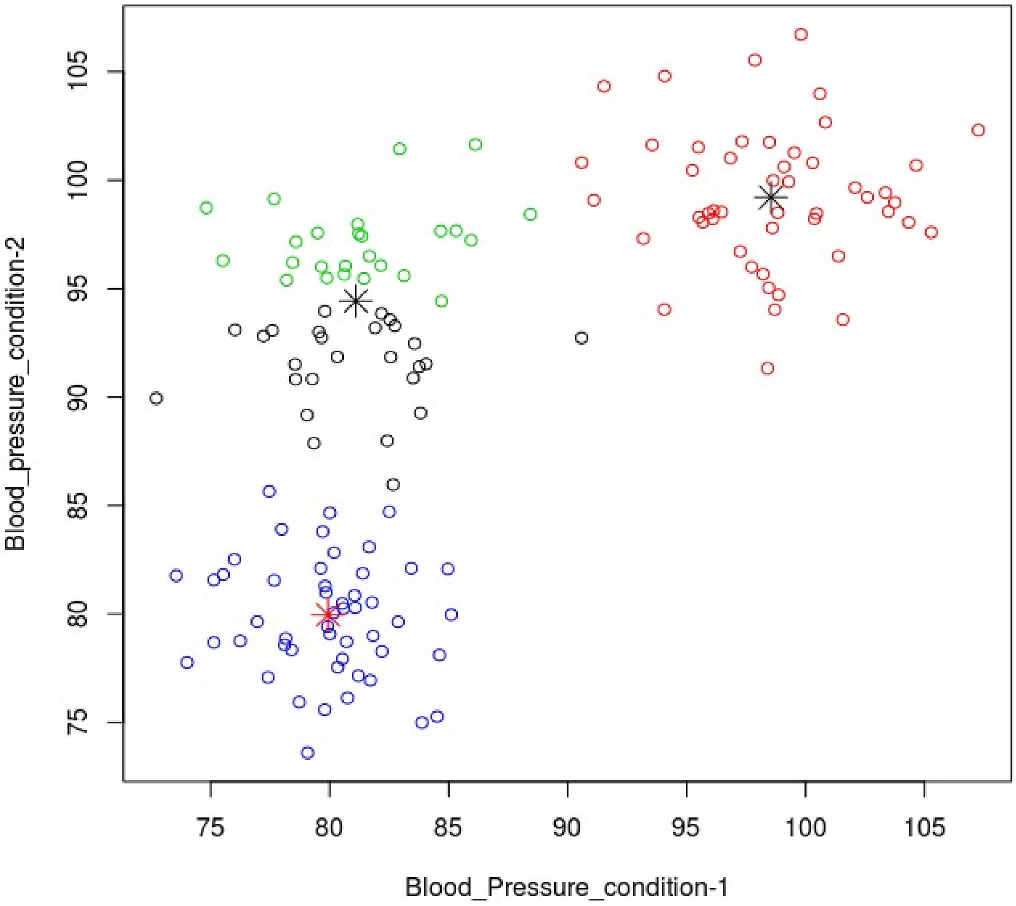
3. When all objects have been assigned, recalculate the positions of the K centroids.

4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

K-means with 3 clusters (n = 3)



K-means with 4 clusters (n=4)



Data : Expression levels of genes g1,g2,... across tissues t1,t2,...

We can perform hierarchical clustering (across row, column)

	t1	t2	t3	t4	t5
g1	419.5	356.2	452.8	308.1	386.1
g2	431.9	323.1	250.8	470.2	531.1
g3	441.4	397.8	466.4	683.9	489.8
g4	260.9	478.8	340.8	459.1	399.1
g5	383.3	351.4	454.9	444.5	229.6
g6	386.9	337.7	371.4	445.0	402.2
g7	436.1	465.6	401.2	525.8	421.5
g8	461.7	324.0	462.1	290.8	372.6
g9	313.2	310.0	461.8	436.6	236.7
g10	392.4	559.4	514.0	402.9	407.2

Cluster along tissues

	t1	**t2**	**t3**	**t4**	**t5**
Mean of column	392.7	390.3	417.6	446.7	387.6

Compute the mean (or, median ...) of expression across genes for each tissue.

Cluster along the tissue.

We get relationship between tissues.

Cluster along genes

Gives relationship between genes

	Mean expression across tissues
G1	384.5
G2	401.42
G3	495.8
G4	387.7
G5	372.7
G6	388.6
G7	450.0
G8	382.2
G9	351.7
G10	455.2

Hierarchical clustering algorithm (S.C. Jhonson, 1967)

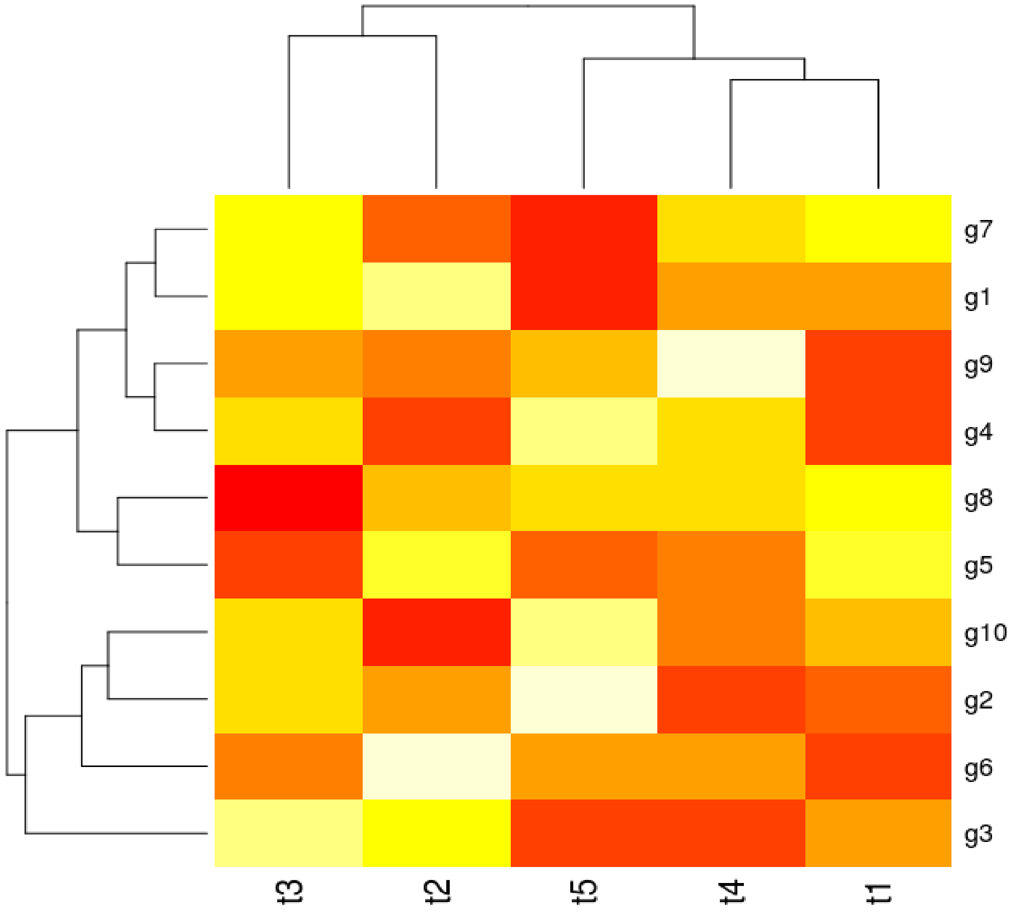
1. Start by assigning each item to a cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters the same as the distances (similarities) between the items they contain.

2. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one cluster less.

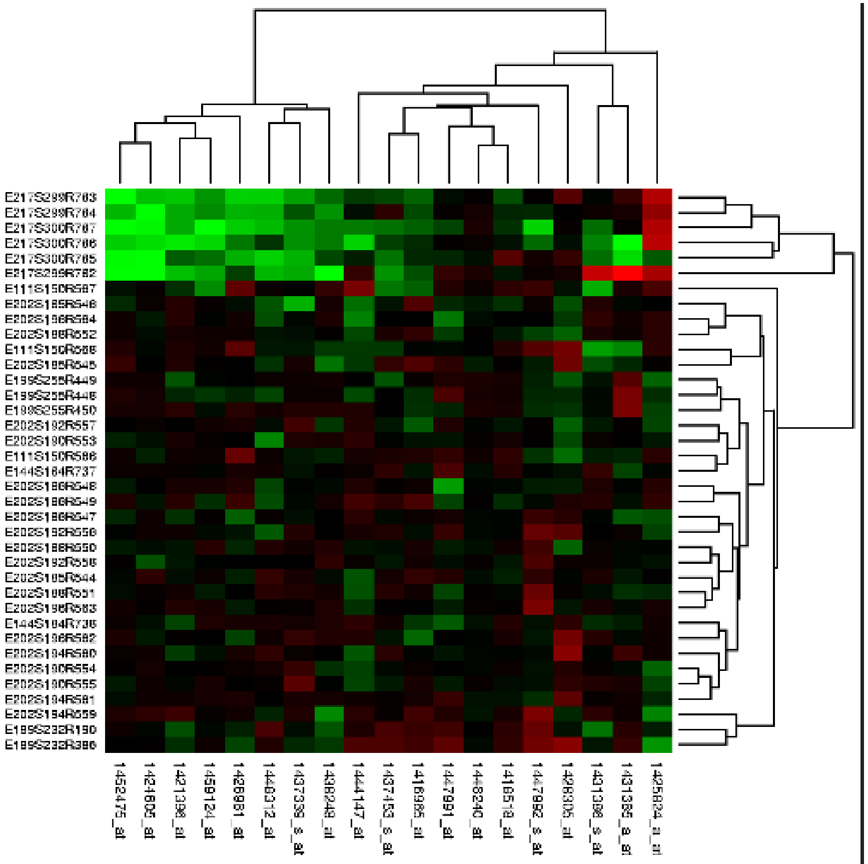
3. Compute distances (similarities) between the new cluster and each of the old clusters.

4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N. (*)

Result of hierarchical clustering in R (dendogram + heatmap)



(Heat map + dendogram) of Real experiment (wikipedia pictures)



For a beautiful and simple tutorial on clustering algorithms for the beginners, visit:

https://matteucci.faculty.polimi.it/Clustering/tutorial_html/index.html