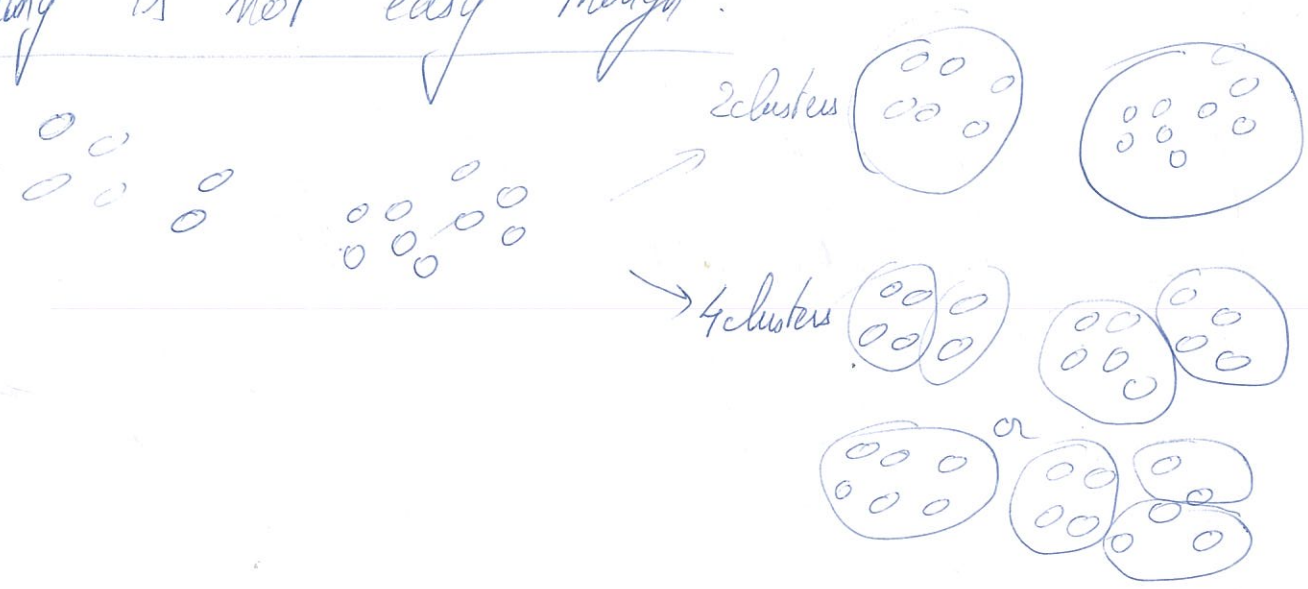
Clustering

I) Clustering is a hard problem.

Classification and clustering are two fundamental tools used in biology.

It is easier to think about a representative than to embrace the information of all individuals.

Clustering is not easy though!



Problems with clustering

→ Problem of noise

→ what is a good /similarity/distance/ measure?

→ How to handle non convexity?

→ How to handle very large datasets?

II Distance measures

Mathematically, a distance measure needs to satisfy:

1. $d(x,x) = 0$
2. $d(x,y) = d(y,x)$
3. $d(x,z) \leq d(x,y) + d(y,z)$
   (triangular inequality)

In addition, $d(x,y) \geq 0$ and $d(x,y) = 0$ if and only if $x = y$

Let us consider two points P_1 and P_2, each represented by a set of "features":

$P_1: (f_{11}, f_{21}, \ldots f_{P1})$

$P_2 (f_{12}, f_{22}, \ldots f_{P2})$

There are usually two types of distances considered to define the similarity of P_1 and P_2:

- Euclidean distances that consider the features as "coordinates"
- Non Euclidean distances that compare properties of the features.

II.1 Examples of Euclidean distances:

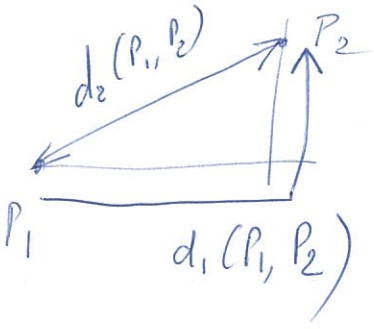
a) Based on L_2 norm:

$$d(P_1, P_2) = \sqrt{(f_{11} - f_{12})^2 + \ldots (f_{P1} - f_{P2})^2}$$

b) Based on L_1 norm:

$$d_1(P_1, P_2) = \sum_{i=1}^{p} |f_{i1} - f_{i2}| \quad \text{(Manhattan distance)}$$

Note.



c) In general:

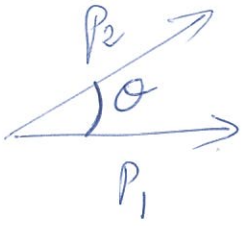
$$d_p(P_1, P_2) = \left(\sum_{i=1}^{p} |f_{i1} - f_{i2}|^p \right)^{1/p}$$

In practice however, limit to $p=1$ or $p=2$.

II.2 Examples of Non Euclidean distances

a) Cosine

$$d_k(P_1, P_2) = 1 - \frac{P_1 \cdot P_2}{\|P_1\| \, \|P_2\|}$$



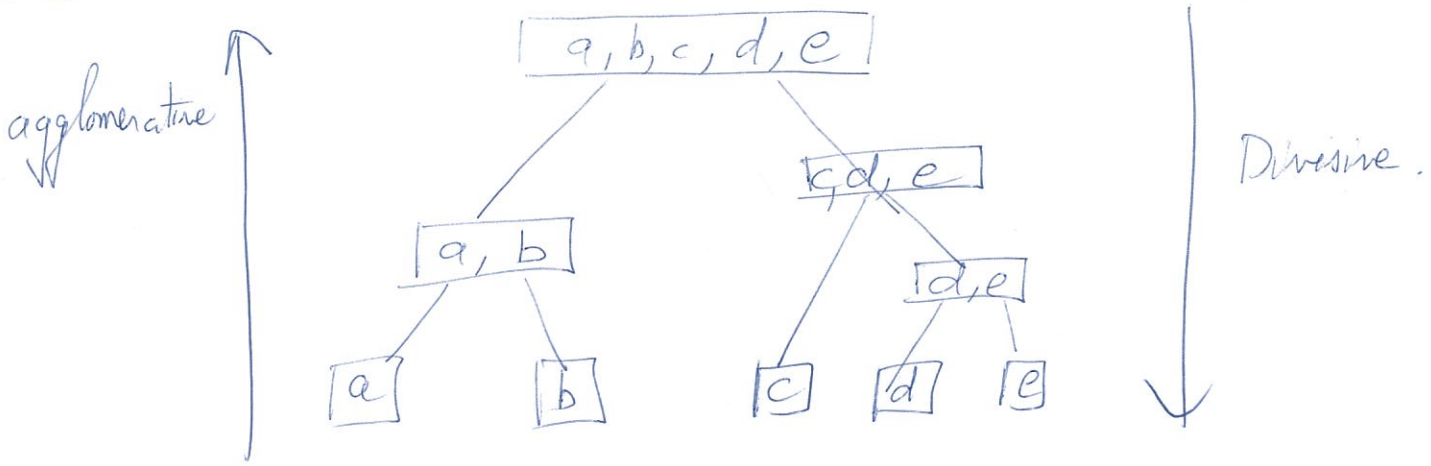
b) Pearson's correlation coefficient.

$$d(P_1, P_2) = \frac{\sum_{i=1}^{p} (f_{i1} - \bar{f}_1)(f_{i2} - \bar{f}_2)}{\sqrt{\sum_{i=1}^{p} (f_{i1} - \bar{f}_1)^2} \sqrt{\sum_{i=1}^{p} (f_{i2} - \bar{f}_2)^2}}$$

The choice of the distance measure is usually guided by the data.

III Hierarchical clustering

III.1 Method



The most common form is agglomerative.

The procedure is:

(1) Start with N independent clusters $\{P_1\}, \dots \{P_N\}$

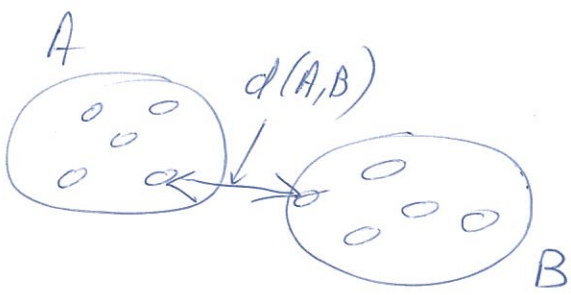
(2) Find the two closest (most similar) clusters and join them

(3) Repeat step (2) until all points belong to the same cluster.

but... we know how to compute the distance between two points... we need a distance between two clusters!

III.2 Intercluster distances

1) Single linkage clustering



Distance between closest pair of points:

$$d(A,B) = \min \{ d(P_1, P_2),\ P_1 \in A,\ P_2 \in B \}$$

2) Complete linkage

Distance between farthest pairs of points.

$$d(A,B) = \max \{ d(P_1, P_2) \; P_1 \in A, \; P_2 \in B \}$$

3) Average linkage:

Mean distance over all mixed pairs of points:

$$d(A,B) = \frac{\sum_{i=1}^{N_A} \sum_{j=1}^{N_B} d(P_i, P_j)}{N_A \; N_B}$$

4) Average group linkage

Mean distance over all pairs of points (i.e. all points considered part of the same cluster

$$d(A,B) = \frac{\sum_{i=1}^{N_A+N_B} \sum_{j=1}^{N_A+N_B} d(P_i; P_j)}{2 (N_A+N_B)^2}$$

Hierarchical clustering is often the method of choice:

- builds a tree.
- Once the tree is built, decide on number of clusters
- does not perform well on non convex data!

IV K-means clustering

The k-means algorithm partitions the data into K mutually exclusive clusters, where K is predefined.

IV.1 Center, a centroids?

For a given cluster $\{P_1, \ldots P_N\}$,

- If the features are coordinates, it makes sense to consider the center of the cluster:

$$C_i = \frac{\sum_{i=1}^{N} P_i}{N}$$

- If the features are not coordinates, define the "centroid" of the cluster:
The centroid is the point P_j in the cluster such that $E(j) = \sum_{i=1}^{N} d(P_j, P_i)$ is minimum.

IV.2 The method

(1) Choose the number of clusters K.
(2) Randomly choose the initial positions of the "centers" of the K clusters.
(3) Assign each point to the nearest "center"
(4) Recompute centroid positions.
(5) Repeat steps 3 and 4 until convergence.

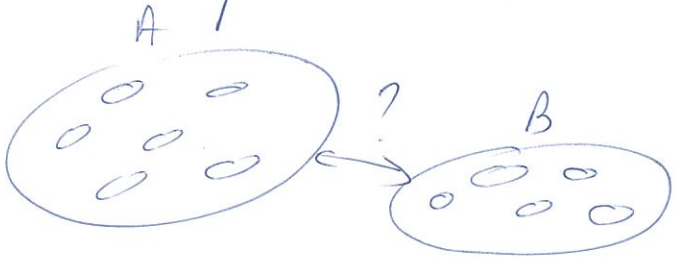
V How many clusters?

Clustering is hard: it is an unsupervised machine learning technique. Once a clustering has been obtained, it is important to assess its validity.

The questions to answer:

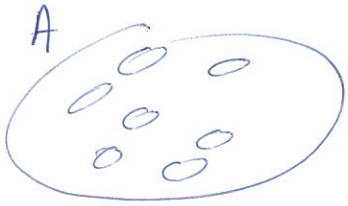
- Did we choose the right number of clusters?
- Are the clusters compact?
- Are the clusters well separated?

V.1 Cluster separation.



single linkage
complete linkage
average linkage
average group linkage.

V.2 Cluster compactness



Similarly, there are several possible definitions for the size of a cluster:

- Complete diameter.

$$\Delta(A) = \max_{P_i, P_j \in A^2} \left(d(P_i, P_j) \right)$$

- Average diameter

$$\Delta(A) = \frac{1}{N_A (N_A - 1)} \sum_{\substack{P_i, P_j \in A^2 \\ P_i \neq P_j}} d(P_i, P_j)$$

Centroid diameter: $\Delta(A) = \frac{2}{N} \sum_{P_i \in A} d(P_i, C)$

where C is the center / centroid of the cluster

V.3 Clustering quality

1) Dunn's index:

$$D = \min_{1 \le i \le K} \left(\min_{\substack{1 \le j \le K \\ j \neq i}} \left[\frac{\delta(S_i, S_j)}{\max_{1 \le k \le K} (\Delta S_k)} \right] \right)$$

Large values of D correspond to good clusters.

2) Davies-Bouldin's index:

$$DB = \frac{1}{K} \max_{i \neq j} \left(\frac{\Delta(S_i) + \Delta(S_j)}{\delta(S_i, S_j)} \right)$$

Low values of DB correspond to good clusters

V.4 Clustering quality: silhouette index

The silhouette index defines a quality index for each point in the original dataset.

- For object P_i in cluster S: calculate the average distance to all other points in $S \rightarrow$ call this value a_i.
- For the same object P_i, and any cluster S' that do not contain P_i, compute the average distance $d(P_i, S')$ and find the minimum of these distances

$$b_i = \min_{S'} \frac{1}{size(S')} \sum_{P_k \in S'} d(P_i, P_k)$$

The silhouette coefficient of P_i is then defined as:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Note that $-1 \leq s(i) \leq 1$

if $s(i) = 1 \rightarrow$ P_i is likely to be well classified

$s(i) = -1 \rightarrow$ P_i is likely to be misclassified

$s(i) = 0 \rightarrow$ indifferent.

From the silhouette of each point we can compute the quality index for the clusters.

Silhouette index for a cluster:

$$Si(S) = \frac{1}{N} \sum_{P \in S} s(P))$$

Global silhouette index:

$$GS = \frac{1}{K} \sum_{i=1}^{K} s(S_i)$$

where K is the total number of clusters.

The larger GS, the better the clustering.

VI Comparing two clustering of the same data.

VI.1 Rand index

Given a set $P = \{p_1, \dots p_N\}$ of N data points, and two partitions of P to compare, $C = \{c_1, \dots c_p\}$ and $D = \{p_1, \dots p_M\}$, we define:

- a : the number of pairs of elements in P that are in the same cluster in C and in the same cluster in D
- b the number of pairs of elements in P that are in different clusters in C and in different clusters in D
- c the number of pairs of elements in P that are in the same cluster in C but in different clusters in D
- d the number of pairs of elements in P that are in different clusters in C but in the same clusters in D

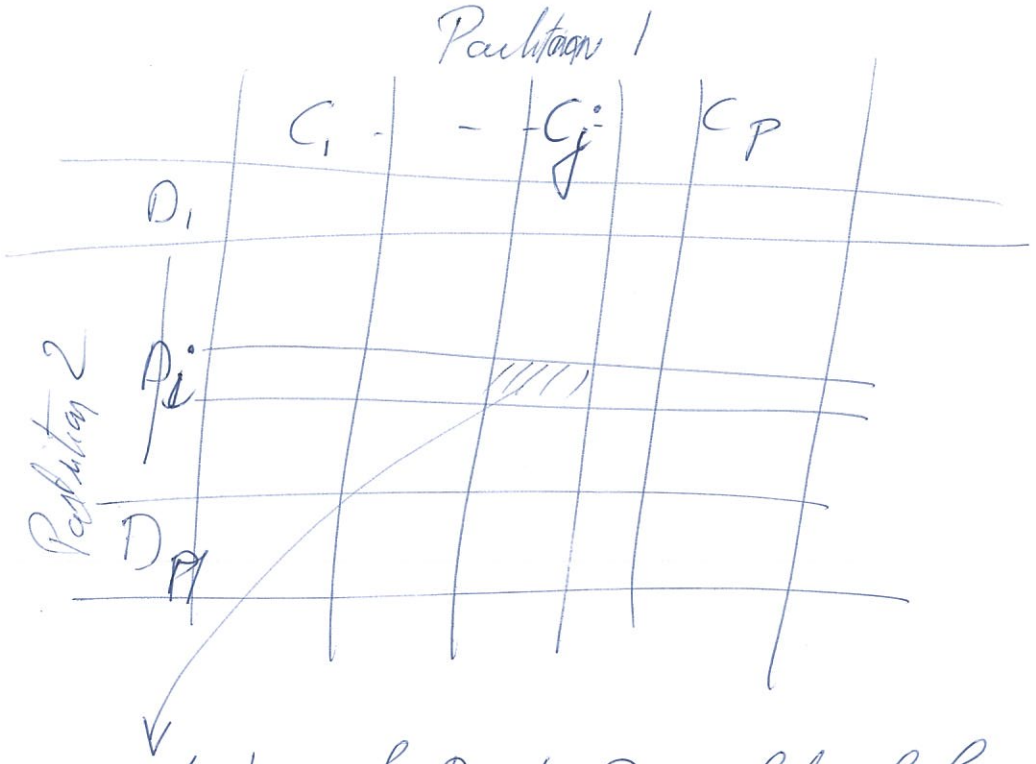
The Rand index, R, is:

$$R = \frac{a+b}{a+b+c+d} = \frac{a+b}{\frac{n(n-1)}{2}}$$

Intuitively, $a+b$ is the number of agreements between C and D, while $c+d$ is the number of disagreements.

VI.2 Confusion matrix

A confusion matrix, also called contingency table or error matrix, is a table that allows visualization of the difference between the two clustering:



Numbers of Point P_i that belong to D_i and to C_j

Ideally, $P = M$ and the table is fully diagonal.