Cluster Analysis

Rosario Barone*

*Email: rosario.barone@uniroma2.it

Classification

- Usual and natural operation in the organization of knowledge
- Given a set of individuals / objects / elements we want to identify subsets that have somehow common characteristics
- obtain classification criteria to locate within the heterogeneity of a collective / population of aggregations capable of simplifying the perception of this collective or of interpreting the typical elements of differentiation

What does common characteristic mean?

In the simplest case, variables of classification

- disease diagnosis: (healthy / diseased)
- degree of customer satisfaction: (zero / average / total)

proximity to a particular place of reference

- real: topological / geographical classification
- ideal: by mapping individuals through scores based on internal / external variables

- Guessing an already defined classification
  - if already available, it tries to highlight the statistical link with other variables external, assess their limits and potential and perhaps isolate those most related to explicit classification variable
  - if not available yet, supervised classification!

- Explore within a collective phenomenon on which they are detectable numerous variables for each unit and the different possibilities of subdividing into subsets showing some degree of homogeneity: classification unsupervised!
  - understand a complex phenomenon and operate in a targeted manner in the subgroups
  - isolate some particularly interesting subgroups as represent a typology on which you want to deepen the research

Cluster analysis procedure

1. Selection of the variables of interest
2. Raw data manipulation
3. Selection and implementation of one of the clustering methodologies
4. Output analysis:
   - trade off in the choice of the number of groups (Occam's razor)
   - groups interpretation

Preliminary notions

- Partition of a set
- Dissimilarity and distance
- Deviance decomposition

Partition of a set

The partition of a set A is the collection of subset

$$\pi = \{I_1, I_2, \ldots, I_k\}$$

such that each subset has no elements in common with any of the others

$$I_g \cap I_r = \emptyset \ \ \forall g \neq r$$

and the union of all subsets I_g reconstitutes the given set A.

$$A = \bigcup_{g=1}^{k} I_g.$$

There are also methods that do not lead to partitions

Dissimilarity and Euclidean distance

- Distance typically used between two units on which they have been taken p measurements: the Euclidean distance

$$d_{ij} = d(x_i, x_j) = \sqrt{\sum_{r=1}^{p} (x_{ir} - x_{jr})^2}$$

- Dissimlilarity index:

$$d_{ij}^2 = d^2(x_i, x_j) = \sum_{r=1}^{p} (x_{ir} - x_{jr})^2 = ||x_{ir} - x_{jr}||^2$$

Dissimilarity and Euclidean distance

- Distance $d : D \times D \in \mathbb{R}$
  - $d(x_i, x_j) = 0 \Leftrightarrow x_i = x_j$
  - $d(x_i, x_j) = d(x_j, x_i)$ simmetry
  - $d(x_i, x_j) \leq d(x_i, x_k) + d(x_k, x_j)$ triangle inequality
- Dissimilarity $d : D \times D \in \mathbb{R}$
  - $d(x_i, x_j) \geq 0 \ \forall \ x_i, x_j$
  - $d(x_i, x_i) = 0 \ \forall x_i$
  - $d(x_i, x_j) = d(x_j, x_i)$ simmetry

Total deviance decomposition

$$\mathbf{T} = \mathbf{B} + \mathbf{W}$$

where:

- **T**→Total deviance
- **B**→ Between deviance
- **W**→ Within deviance

Suppose we observe one variable n times, the n units partitioned in G groups of size n_g, such that $\sum_{g=1}^{G} n_g = n$

$$\sum_{j=1}^{n}(x_j - \bar{x})^2 = \sum_{g=1}^{G}\sum_{i=1}^{n_g}(x_{ig} - \bar{x})^2 = \sum_{g=1}^{G} n_g(\bar{x}_g - \bar{x})^2 + \sum_{g=1}^{G}\sum_{i=1}^{n_g}(x_{ig} - \bar{x}_g)$$

Clustering algorithms

- Hierarchical
  - Aggregation Algorithms
  - Scissor Algorithms
- Not hierarchical
  - groupings around means (k-means)
  - groupings around representative units (k-medoids)

Not hierarchical algorithm

You get a single partition with number of components (groups) specified a priori. Starting from an initial subdivision into a fixed number of clusters, we proceed sequentially obtaining at each step a new partition that improves the current partition according to a chosen criterion.

K-means algorithm

1. Initialization: we choose the desired number of groups (K), and set as starting points K centers or means (K p-dimensional vectors): $(\mathbf{m}_1, \ldots, \mathbf{m}_K)$
2. Allocate the i-th unit for $i = 1, \ldots, n$ into the group j such that the distnce between the unit x_i and the center $\mathbf{m}_k$ is minimized;
3. Calculate the new centers $(\mathbf{m}_1, \ldots, \mathbf{m}_K)$ basing on the new allocations to the groups (clusters).

Repeat the steps 2 and 3 until convergence, i.e. when the allocation of the observations does not vary between two iterations.

K-means algorithm

This algorithm only works starting with a data matrix in which all variables are quantitative (measurements). Let n_j be the number of units that are assigned in step 2 to the j-th group G_j

$$\mathbf{m}_j = (m_{j1}, \ldots, m_{jr}, \ldots, m_{jp})$$

$$m_{jr} = \frac{1}{n_j} \sum_{i \in G_j} x_{ir}$$

The objective to be achieved is formalized in the following problem of optimization: identify K groups for which the internal variability is minimal and (automatically) the between variability is maximal.

Generalization fo the K-means method

To allow to use only one dissimilarity matrix between units you can modify it the identification of centers through averages with corresponding actual centers to the coordinates of central units (Medoids) which minimize the distance between the unit candidate to play the role of center and the other units of the group. There modification of step 3 becomes much more expensive from the point of view computational.

The modification of step 2 is immediate and does not involve additional difficulties. Such modifications can lead to more robust algorithms than the presence of anomalous distances between single pairs of units

K-means vs K-medoids

PAM (Partitioning Around Medoids) algorithm

1. Initialization:
   - calculation of the dissimilarity matrix
   - choice of K points as initial candidate medoids
2. (build phase) assign each unit to the nearest medoid to form the K matching clusters
3. (swap phase) for each cluster formed, check if each unit of the cluster it is able to decrease the overall dissimilarity of the same cluster ed eventually proceed with the replacement of the medoid with that unit it does decrease the dissimilarity of the cluster more. If for at least one cluster it was replaced the previous medoid go back to step 2.

Hierarchical algorithms

Do not specify the number of groups, which ranges from the extreme minimum (1 group only) to the extreme maximum (many groups how many units are considered). This range of possibilities is typically represented by the dendrogram.

- *Aggregatives algorithms*: start from the consideration of the maximum number of groups n obtainable considering each group formed by a single unit; proceed sequentially to obtain a smaller number of groups ($n \to n-1 \to n-2 \dots$).

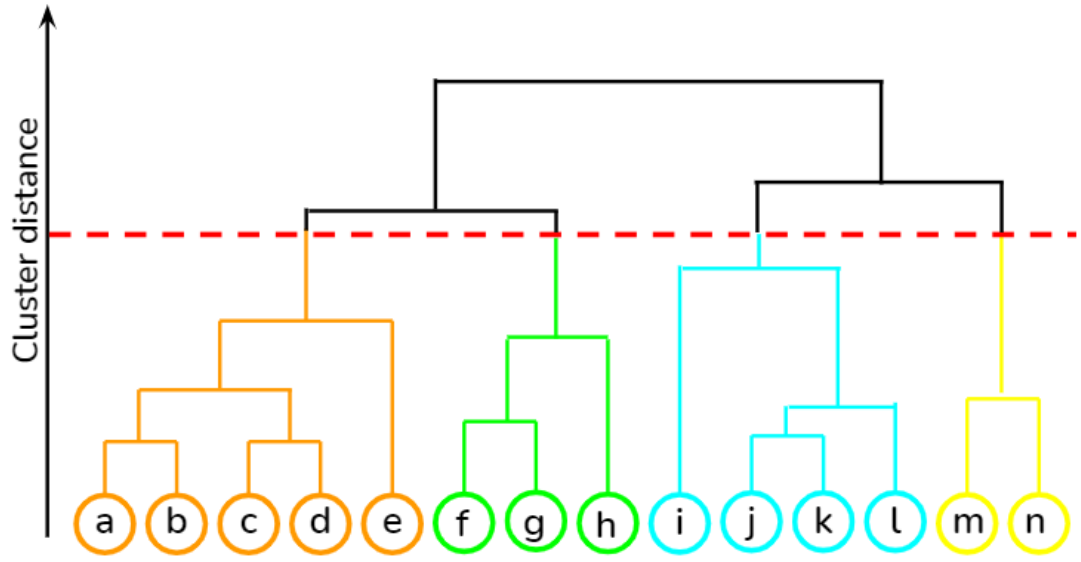
- *Scissors algorithms*: start from the consideration of all the units belonging to only one group and proceed by subdivision according to suitable criteria of optimality.

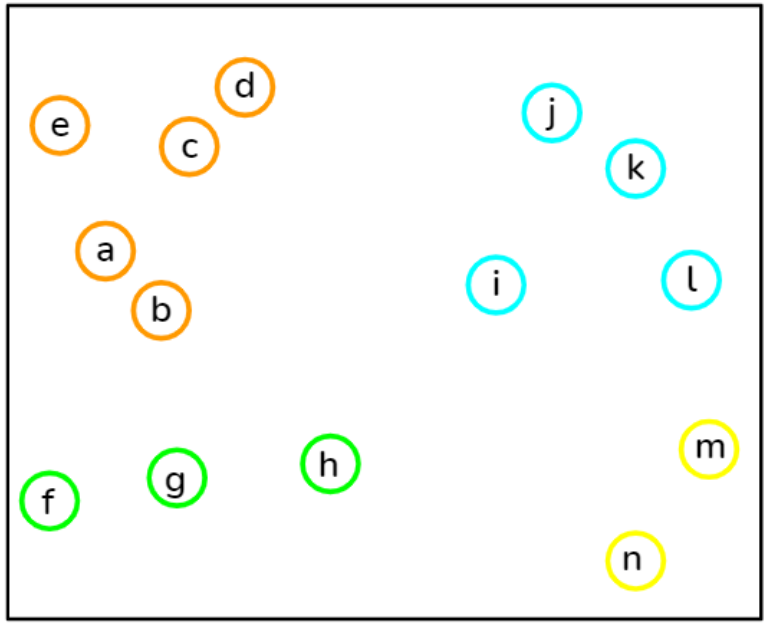
Dendrogram

It is a visualization of the subsequent aggregation process (aggregation methods) of the n minimal subgroups (each consisting of a single unit) in groupings composed of the aggregation of pairs of subgroups according to a pre-established criterion; or the subsequent subdivision process (methods scissors) starting from the set of all units. What you must be able to see in a dendrogram:

1. how to identify the subdivision into K groups;
2. which units belong to a group;
3. indications on the selection of an adequate number of groups to well represent the (possible) grouping structure;

Dendrogram





Aggregtive methods

The general operating principle is as follows:

- It is initialized considering $K = n$ distinct groups each consisting of one element starts from a matrix of distances/dissimilarities between the K groups
- aggregate the two groups that are less distant between all the $K(K-1) = 2$ possible pairs and the new distance matrix relating to $(K-1)$ is recalculated groups thus determined
- If $K = 1$ we stop otherwise we decrease the index K by 1 and go back to step 1.

- The strting matrix of distances necessary for the elaborations or is provided directly as input or it is calculated as a function of raw data.
- The calculation of the new matrix of distances / dissimilarity to when the number of groups and diminished by one, it presupposes that the notion of distance between sets of units (between two groups). For example the distance between group I and the group J we will denote it $d(I; J)$.
- When the two groups are made up of a single unit, the definition is natural. In the non-trivial case of groups consisting of several units, they can be used different definitions. Among these we point out the most frequently used:

$$d_{SL} = \min_{i \in I, j \in J} d(i, j)$$

$$d_{CL} = \max_{i \in I, j \in J} d(i, j)$$

$$d_{AL} = \text{mean } d(i, j)$$

Note that the operations in question can be calculated starting from the last matrix of distances obtained. If group I is obtained by aggregation of groups G_1 and G_2 from the last distance matrix obtained, I can identify the quantities a $d(G1; J)$ and $d(G2; J)$ from which I deduce:

- Single-Linkage or Neirest neighbor:

$$d_{SL} = \min_{i=1,2} d(G_i, J)$$

- Complete-Linkage:

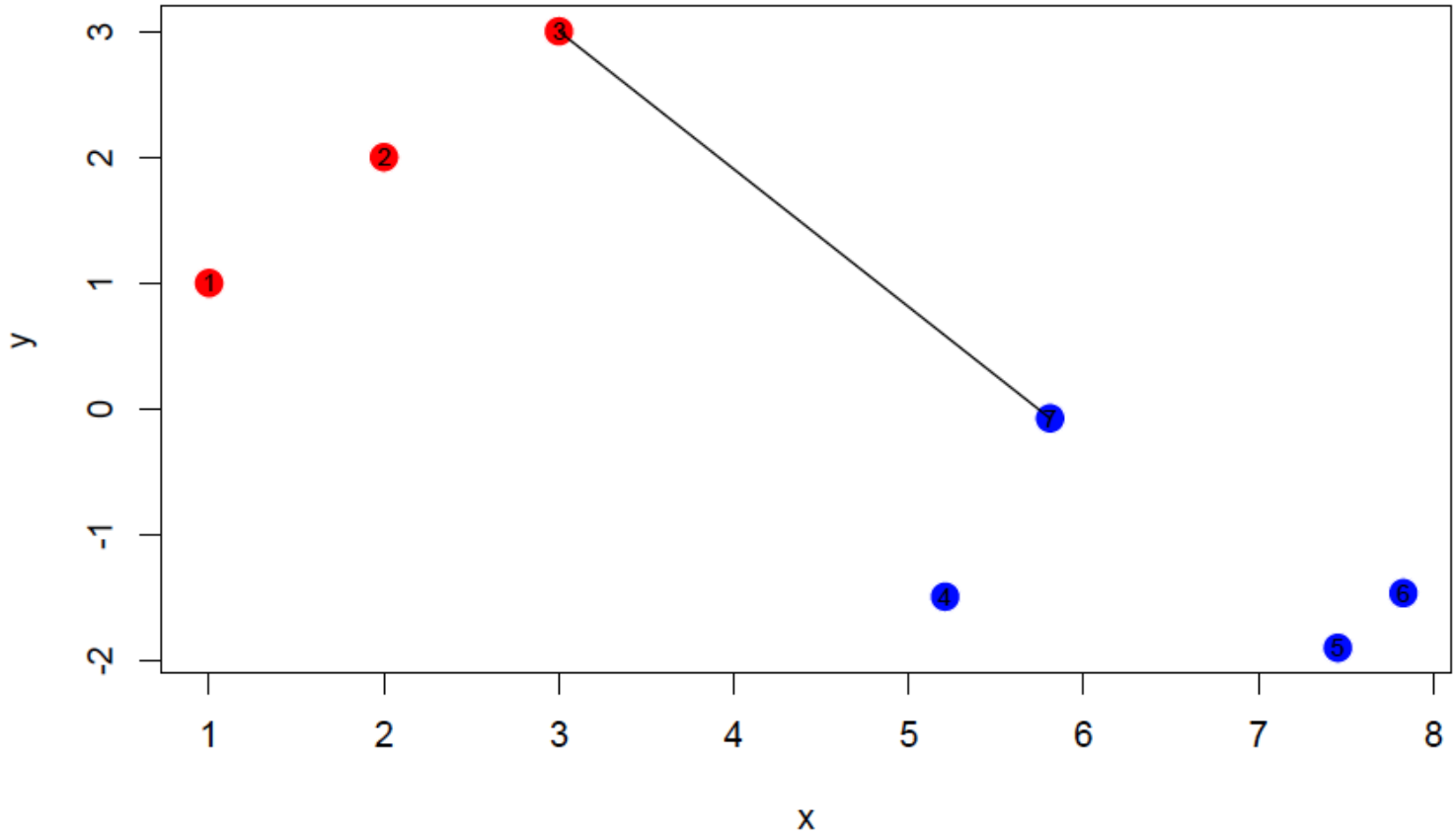
$$d_{CL} = \max_{i=1,2} d(G_i, J)$$

- Average-Linkage or Mean distance:

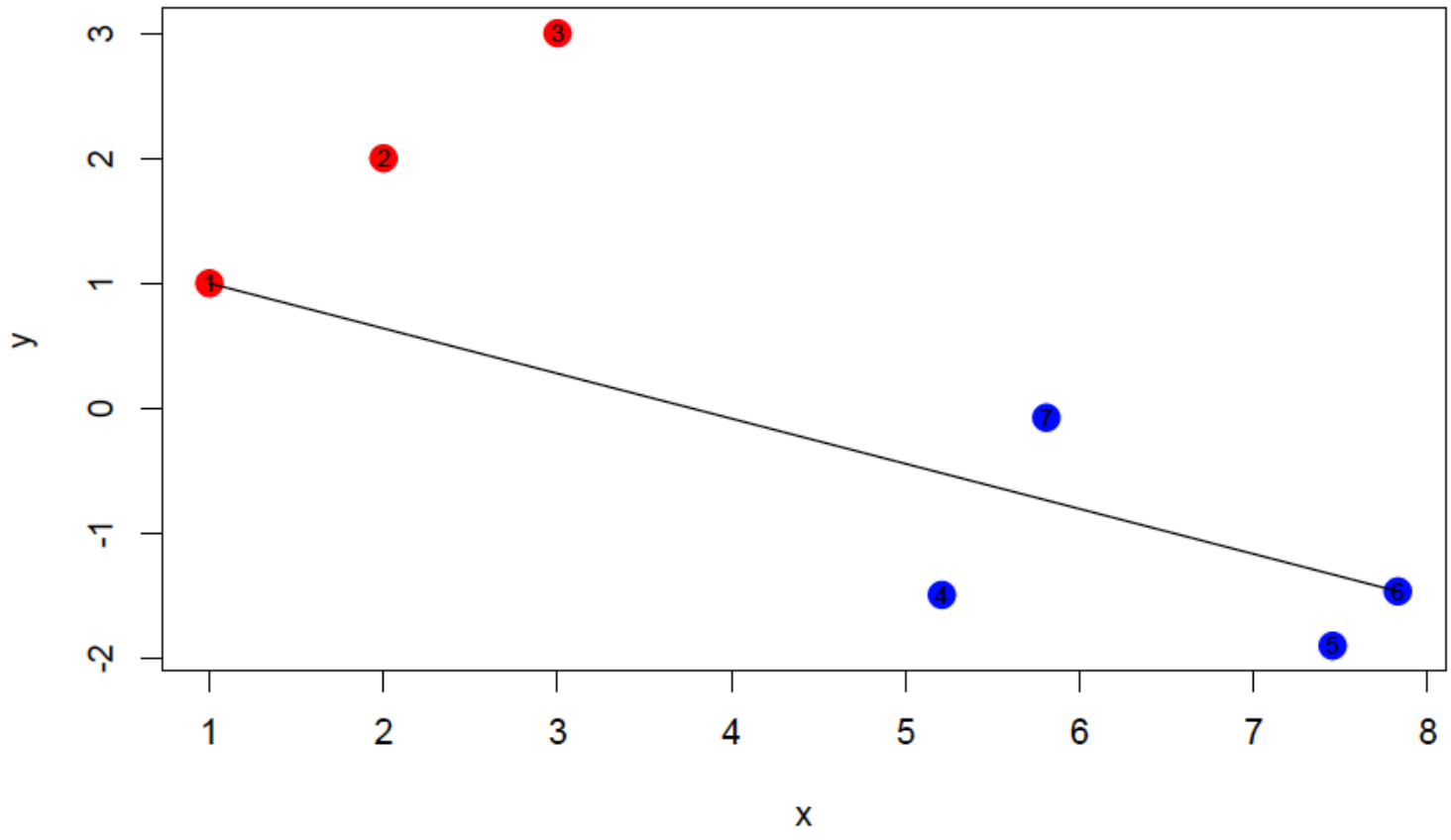
$$d_{AL} = \frac{n_{G_1}}{n_{G_1} + n_{G_2}} d(G_1, J) + \frac{n_{G_2}}{n_{G_1} + n_{G_2}} d(G_1, J)$$

where n_{G_i} is the number of units composing the group G_i.

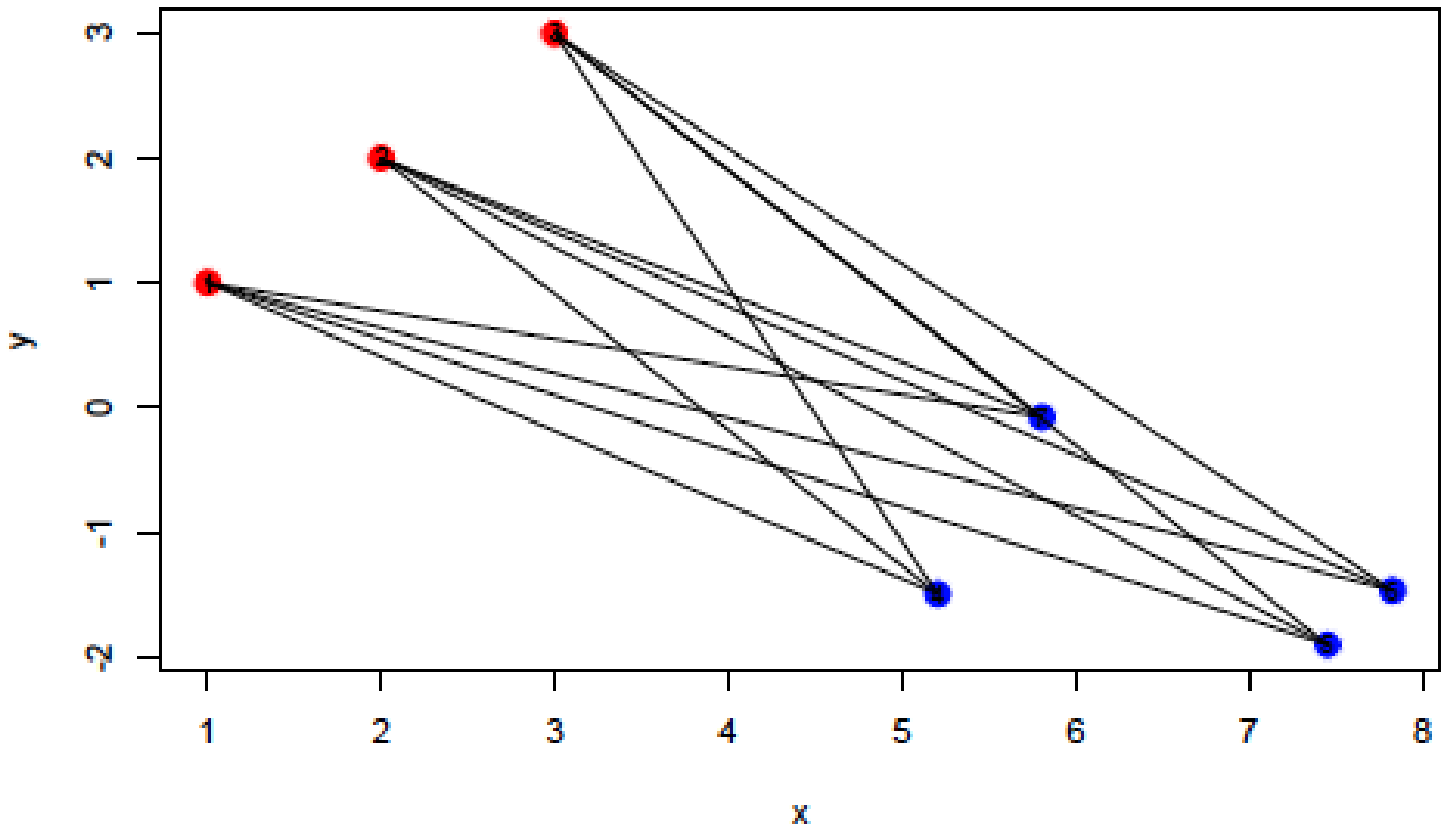
Single-Linkage or Neirest neighbor



Complete-Linkage



Average-Linkage or Mean distance



How to choose the number of groups

- Graphic inspection of the dendrogram
- Some tests based on formal criteria: Calinski and Harabasz
- Silhouette
- Gap statistics

Silhouette

The silhouette is an index which takes values in the interval $[-1, 1]$.

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

where $a(i)$ is the average dissimilarity in the cluster considered, $b(i)$ is the average dissimilarity w.r.t. any other clusters. If:

- $s \approx 1 \rightarrow$ the unit i is assigned to the right cluster
- $s \approx 0 \rightarrow$ the unit i should not be assigned to either the compared clusters
- $s \approx -1 \rightarrow$ the unit i is assigned to the wrong cluster

Silhouette in the choice of the number of clusters

The value of $s(i)$ depends on the partition π and so on the chosen number of clusters. We should adopt a notation of the type $s_\pi(i)$. We use as criteria for the choice of the number of groups the one based on the mean of the siluatte, therefore the number K which maximizes

$$\bar{S}_K = \frac{1}{n} \sum_{i=1}^{n} s_K(i).$$

Diagnostics

- Internal validation: uses data only and is based on the quantitative measures of distance used and / or on the definition of reference structures for the presence / absence of groupings.
  - Average silhouette ($\bar{S}$)
    - $\bar{S} \in (0.7, 1.00]$ the partition obtained is extremely reliable
    - $\bar{S} \in (0.5, 0.7)$ the partition obtained is reliable
    - $\bar{S} \in (0.25, 0.5)$ the partition obtained is not very reliable
    - $\bar{S} \in [-1, 0.25)$ the partition obtained is not reliable
  - Between deviance ratio on total deviance
- External validation: uses a priori information or structures of existing classification.
  - Rand index
  - Rand index adjusted

Probabilistic non hierarchical methods: mixture models

When analyzing a data set we assume that each observation comes from one specific distribution.

$$Y_i \sim N(\mu, \sigma^2) \quad \text{for } i = 1, \ldots, n$$

Then we proceed to estimate parameters of this distribution using maximum likelihood estimation, i.e.:

$$\frac{\partial \mathcal{L}(\mu, \sigma^2; Y)}{\partial \mu \partial \sigma^2} = 0 \quad \text{and} \quad \left(\hat{\mu}_{ML}, \hat{\sigma}^2_{ML}\right)$$

The assumption that each observation comes from one specific distribution may often be inadequate.

Mixture models: when?

In many cases, assuming that each sample comes from the same unimodal distribution is too restrictive and may not make intuitive sense. Often the data we are trying to model are more complex:

- Single or groups of observed individuals may have unobserved effects which may affect the estimates.

- Multimodality: multiple regions with high probability mass.

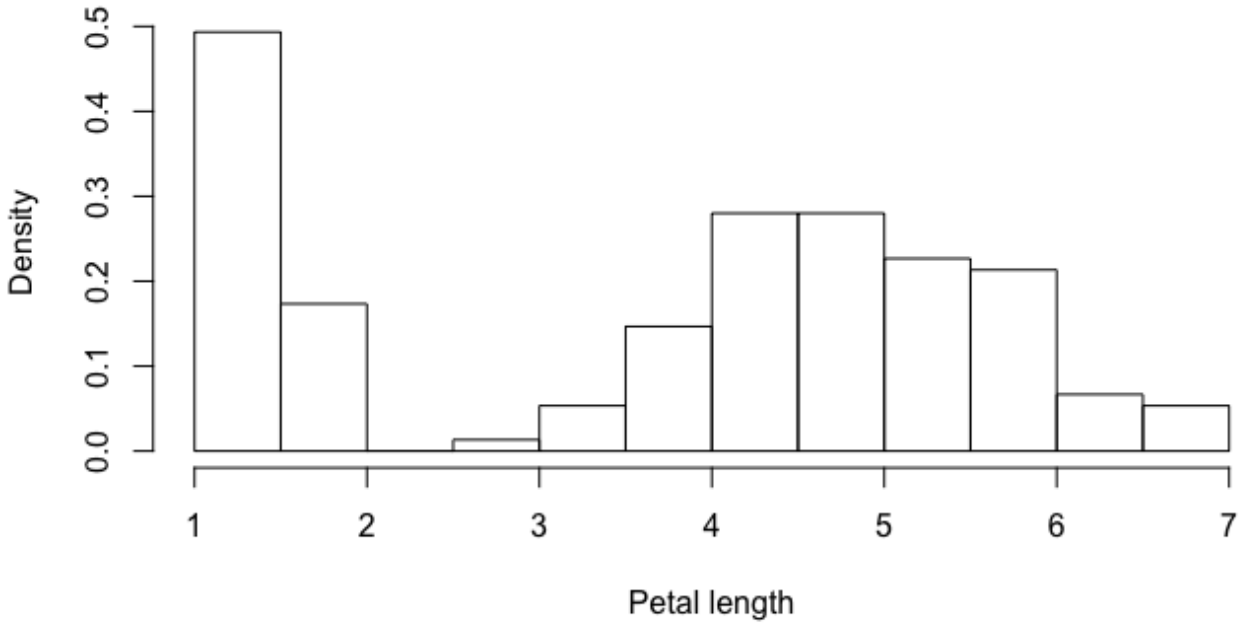
Mixture models: overview

A mixture model is a probabilistic model for representing the presence of subpopulations within an overall population.

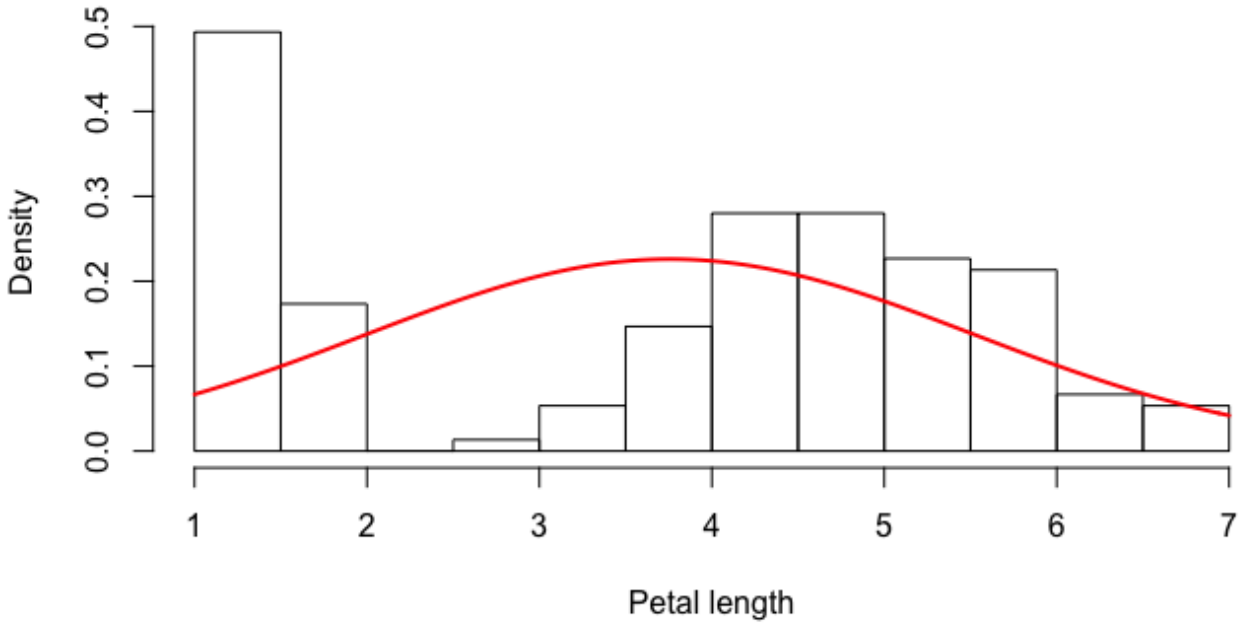
It does not require that an observed data set should identify the sub-population to which an individual observation belongs, i.e:

- we observe a sample of n individuals;
- we assume that in our sample there are K subpopulations;
- we do not specify the subpopulation to which each individual belongs.

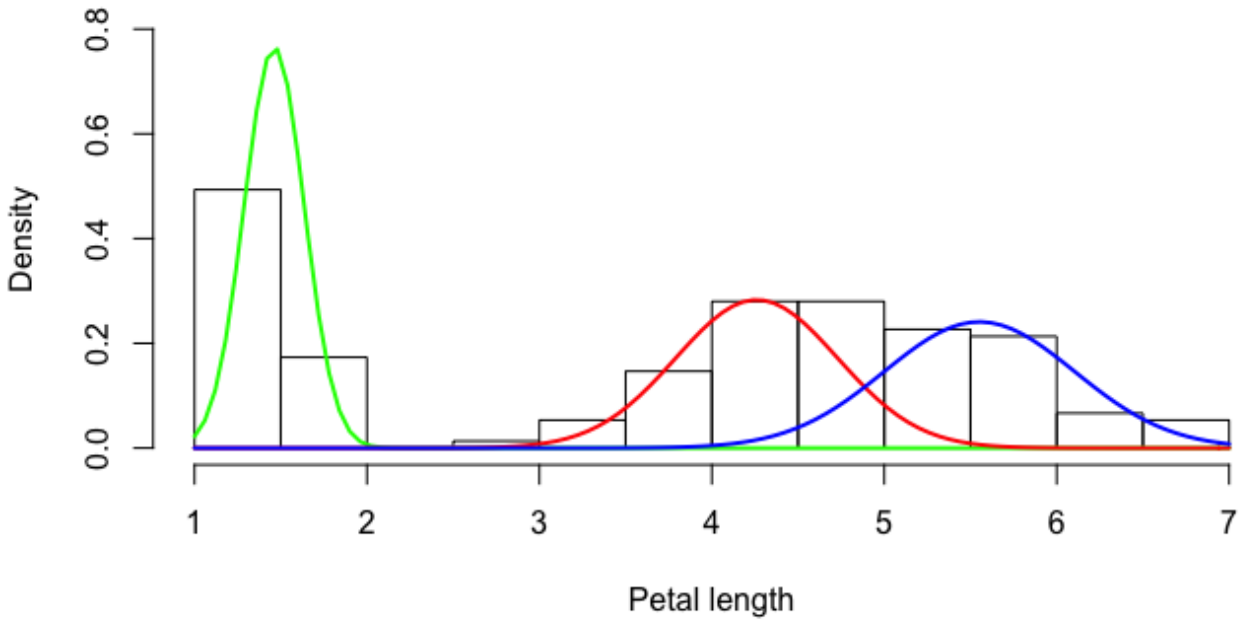
Mixture models: Iris data example



Mixture models: Iris data example



Mixture models: Iris data example



Finite mixture models

We define a mixture distribution $f_\theta \in \Omega$, with $\theta \in \Theta$:

$$f_\theta(y) = \sum_{k=1}^{K} w_k f_{\theta_k}(y)$$

where:

- $f_{\theta_k} \in \Omega \quad \forall k \in K$;
- $\sum_{k=1}^{K} w_k = 1$.

Note that in a sample of n observed individuals $K \leq n$.

Finite mixtures in regression analysis

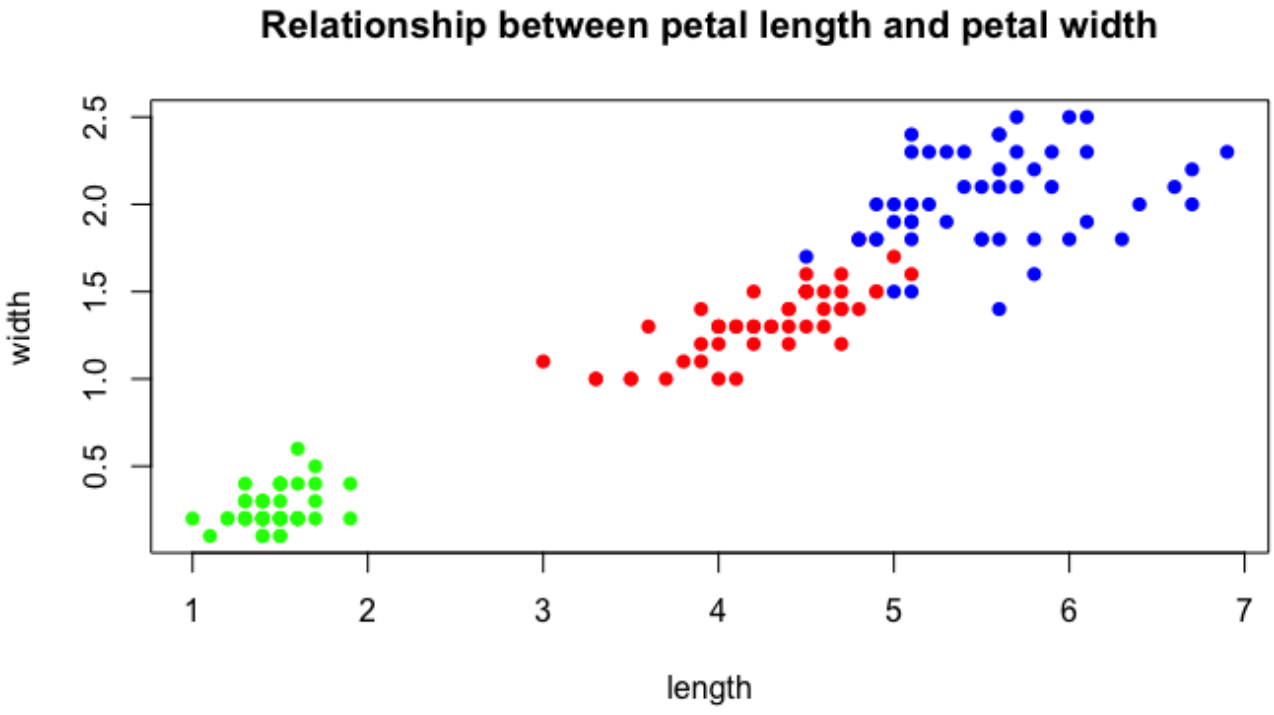
Let y be a dependent variable and let x represent a covariate. We define a mixture of regression models as a distribution $f_\theta \in \Omega$, with $\theta = (\alpha, \beta) \in \Theta$:

$$f_\theta(y|x) = \sum_{k=1}^{K} w_k f_{\theta_k}(y|x)$$

where:

- $f_{\theta_k} \in \Omega \quad \forall k \in K$;
- $\sum_{k=1}^{K} w_k = 1$;
- $g(E(y|x)) = \alpha_k + \beta_k x$.

Mixture in regression analysis: Iris data example



Finite mixture models

Let assume to have a sample of N individuals. We define a mixture model as a *hiearchical model* composed by:

- N random variables that are observed, each distributed according to a mixture of K components belonging to the same parametric family of distributions, but with different parameters;
- N random **latent** variables specifying the identity of the mixture component of each observation, each distributed according to a *K-dimensional categorical distribution*;
- A set of K mixture weights w, which are probabilities that sum to 1.
- A set of K parameters, each specifying the parameter of the corresponding mixture component.

Finite mixture models: inference

Most of the approaches for finite mixture estimation that have been proposed focus on maximum likelihood methods.

Two scenarios to consider:

- if K is assumed to be known: expectation maximization (EM) is the most popular technique used to determine the parameters of a mixture with an a priori given number of components;
- if K is assumed to be uknown: (in general) methods to determine the number and functional form of the mixture components are distinguished from methods to estimate the corresponding parameter values.