Unsupervised Learning
Clustering

Elena Bellodi
elena.bellodi@unife.it

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Regione Emilia-Romagna

FERRARIAE UNIVERSITAS · 13 91 · EX LABORE FRUCTUS

Università degli Studi di Ferrara

Outline

Introduction to Clustering

- Exclusive Clustering
- Overlapping Clustering
- Hierarchical Clustering
- Probabilistic Clustering

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Recap: supervised/unsupervised Learning

- Machine Learning Goals:
  - **Prediction** - what can we predict about this phenomenon?
  - **Description** - how can we describe/understand this phenomenon in a new way?

	Predictive model	*Descriptive model*
Supervised learning	classification, regression	subgroup discovery
Unsupervised learning	predictive clustering	descriptive clustering, association rule discovery

Table 1.1. An overview of different machine learning settings. The rows refer to whether the training data is labelled with a target variable, while the columns indicate whether the models learned are used to predict a target variable or rather describe the given data.

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Unsupervised Learning

1. Clustering learns a new labelling function g from unlabelled data which divides data into **clusters** (subgroups)
   - the function is a *predictive model* that can be applied to new data
   - the function is a *descriptive model* that only describes the unlabelled data

- Different kinds of clustering algorithms exist

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Unsupervised Learning

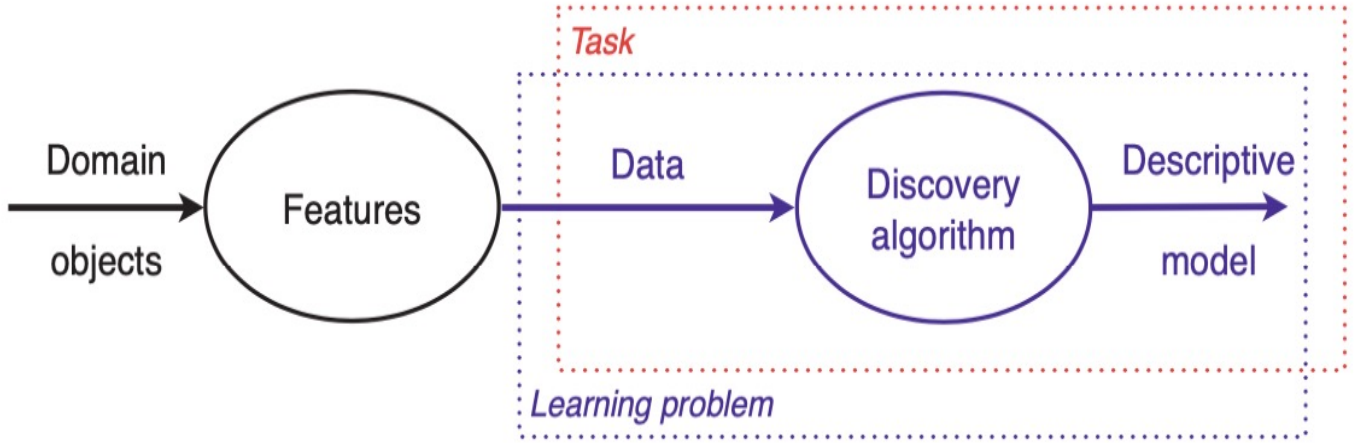
2. Association rules are patterns (*descriptive models*) that can be learned from unlabelled data
   - manifestation of underlying structure in the data
   - Data Mining

Association Rules	Support (%)	Confidence (%)	Lift
Soap → Perfume	30	75	1.87
Perfume → Soap	30	75	1.87
Bread → Jam	40	80	1.6
Jam → Bread	40	80	1.6
Chocolate → Bread	30	75	1.5
Beer → Snacks	30	100	2

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Unsupervised Learning: *descriptive* models

- *Descriptive learning* in general leads to the *discovery* of genuinely new knowledge



In descriptive learning the task and learning problem coincide: we do not have a separate training set, and the task is to produce a descriptive model of the data.

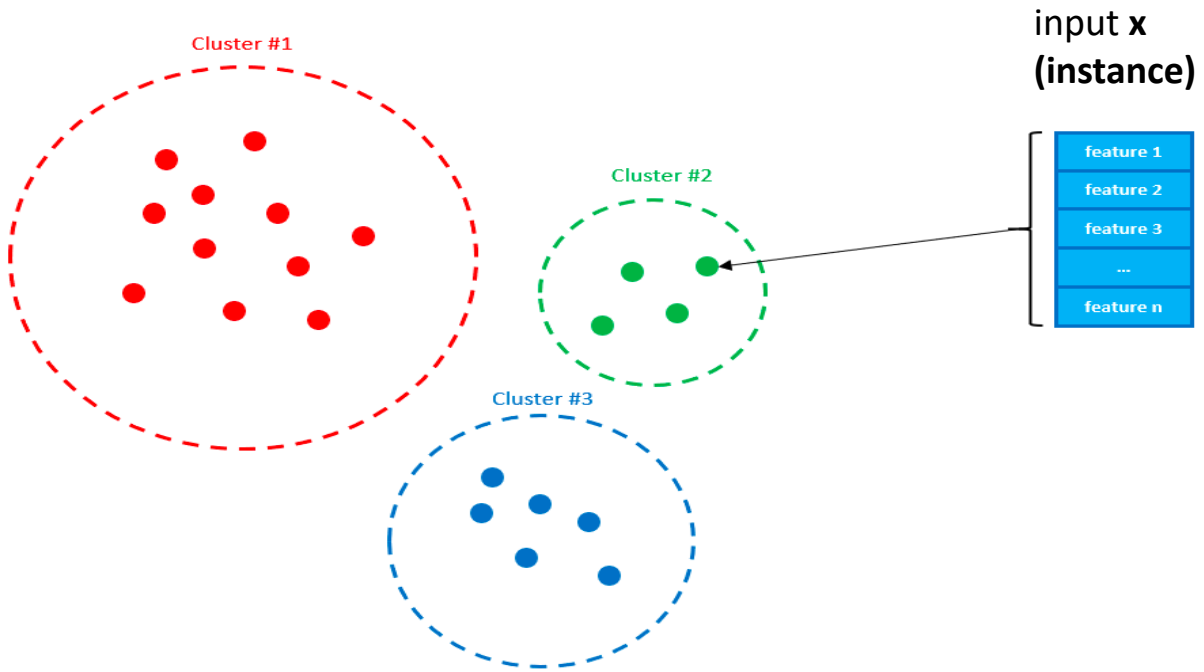
Unsupervised Learning: *performance*

- There is no test data:
  - In *clustering*, to evaluate a particular partition of data into clusters, one can calculate the average distance from the cluster centre;
  - In *association rules mining*, one can compute some metrics (support, confidence) to understand the popularity of an example in the data set and the likelihood of a rule.

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Introduction to Clustering

- A clustering algorithm works by assessing the similarity between instances (the things we're trying to cluster, e.g., e-mails) and putting similar instances in the same cluster and 'dissimilar' instances in different clusters



Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Introduction to Clustering

- Goals

1. Finding **homogeneous groups** in the data → see next

2. Finding unusual data objects (**outlier detection**). An object is an outlier if:
   - *Does the object belong to any cluster*? If not, then it is identified as an outlier.
   - *Is there a large distance between the object and the cluster to which it is closest*? If yes, it is an outlier.
   - *Is the object part of a small or sparse cluster*? If yes, then all the objects in that cluster are outliers.

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Distance Measures

- Range in 0 - +infinite
- Distance for data points in $\mathbb{R}^n$: *euclidean distance*

$$d_2(x,y) = \left(\sum_{k=1}^{n} (x_k - y_k)^2 \right)^{1/2} = \|x - y\|_2$$

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Similarity Measures

- Range in 0 - 1
- Cosine

$$s_{\cos}(x, y) = \frac{x^T y}{\|x\|\|y\|}$$

- Dice Coefficient

$$s_{Dice}(x, y) = \frac{2x^T y}{\left(\|x\|^2 + \|y\|^2\right)}$$

- Exponential Similarity

$$s_{\exp}(x, y) = \exp\left(-\|x - y\|^{\alpha}\right)$$

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Clustering Techniques

- **Exclusive Clustering**: each instance belongs to a unique cluster
  - ex: K-means
- **Overlapping Clustering**: based on fuzzy sets
  - each point may belong to two or more clusters with different degrees of membership
  - ex: Fuzzy C-means
- **Hierarchical Clustering**: builds a hierarchy of clusters
- **Probabilistic Clustering**
  - ex: Mixture of Gaussians

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Exclusive Clustering: K-Means

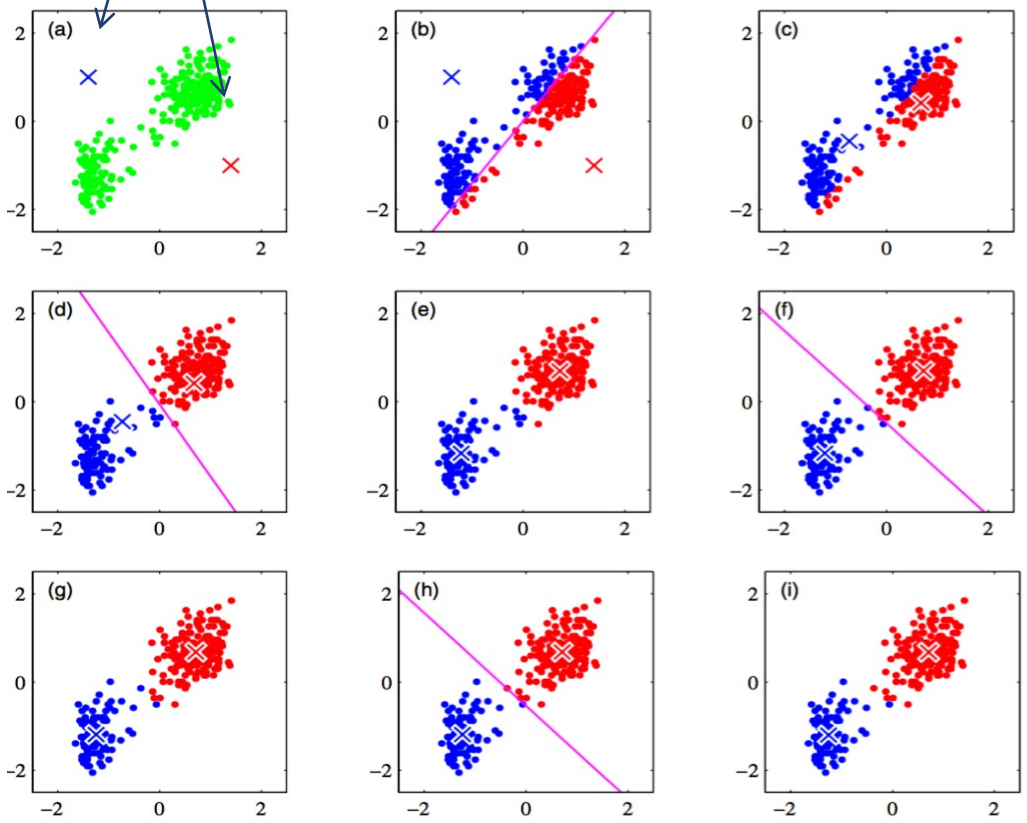
- It works in numeric domains
- It finds:
  - K unique clusters
  - an assignment of data points to clusters
  - **the center of each cluster μ_k is the mean of the values in that cluster** such that the sum of the squares of the distances of each data point to its closest μ_k is a minimum
    - K-means minimizes the objective function:

$$e^2 = \sum_{j=1}^{k} \sum_{i \in cluster(j)} d^2(x_i, c_j)$$

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Exclusive Clustering: K-Means

(a) initial choices for centres



(b) each data point is assigned either to the red cluster or to the blue cluster, according to which cluster centre is nearer

(c) each cluster centre c_j is re-computed to be the mean of the points m_j assigned to the corresponding cluster

$$c_j = \frac{\sum_{i=1}^{m_j} x_i}{m_j}$$

(d) the algorithm converges

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Initial choices for centres

Various choices are possible:

- The first k instances in the dataset
- Label instances with numbers from 1 to m (number of instances) and choose those with numbers m / k, 2m / k, ..., (k-1) m / k and m
- Randomly choose k instances
- Generate k points by randomly choosing the values of each coordinate in the coordinate range
- Generates a partition of the dataset into k mutually exclusive subsets and considers the centroids of the subsets

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Learning Stages for Clustering

General approach to k-means clustering

1. Collect: Any method.
2. Prepare: Numeric values are needed for a distance calculation, and nominal values can be mapped into binary values for distance calculations.
3. Analyze: Any method.
4. Train: Doesn't apply to unsupervised learning.
5. Test: Apply the clustering algorithm and inspect the results. Quantitative error measurements such as sum of squared error (introduced later) can be used.
6. Use: Anything you wish. Often, the clusters centers can be treated as representative data of the whole cluster to make decisions.

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Stage 5 for K-means: «Test»

How do you know that the clusters are good clusters?

- **SSE (Sum of Squared Error)** is the sum of the squared differences between each observation and its group's mean:

$$SSE = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- **A lower SSE means that points are closer** to their centroids → good clustering

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

K-Means: Pros and Cons

(+) Easy to implement

(-) The *K* -means problem is NP-complete, which means that there is no efficient solution to find the global minimum

- while *K* -means converges to a stationary point in finite time, **no guarantees can be given about whether the convergence point is in fact the global minimum, or if not, how far we are from it**
- an unfortunate initialisation of the centroids can lead to a sub-optimal solution. In practice it is advisable to run the algorithm a number of times and **select the solution with the smallest within-cluster scatter.**

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

K-Means: Pros and Cons

(-) Large number of dimensions and/or large datasets increase **time complexity**

(-) For methods distance-based, if an obvious distance measure doesn't exist we must "define" it, which is not always easy, especially in multi-dimensional spaces

(-) The performance on a data set will be determined by the distance measure chosen

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Applications

- **Marketing**: finding groups of customers with similar behavior given a large database of customer data containing their properties and past buying records;
- **Biology**: classification of plants and animals given their features;
- **Libraries**: book ordering;
- **City-planning**: identifying groups of houses according to their house type, value and geographical location;
- **Earthquake studies**: clustering observed earthquake epicenters to identify dangerous zones;
- **WWW**: document classification; etc...

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

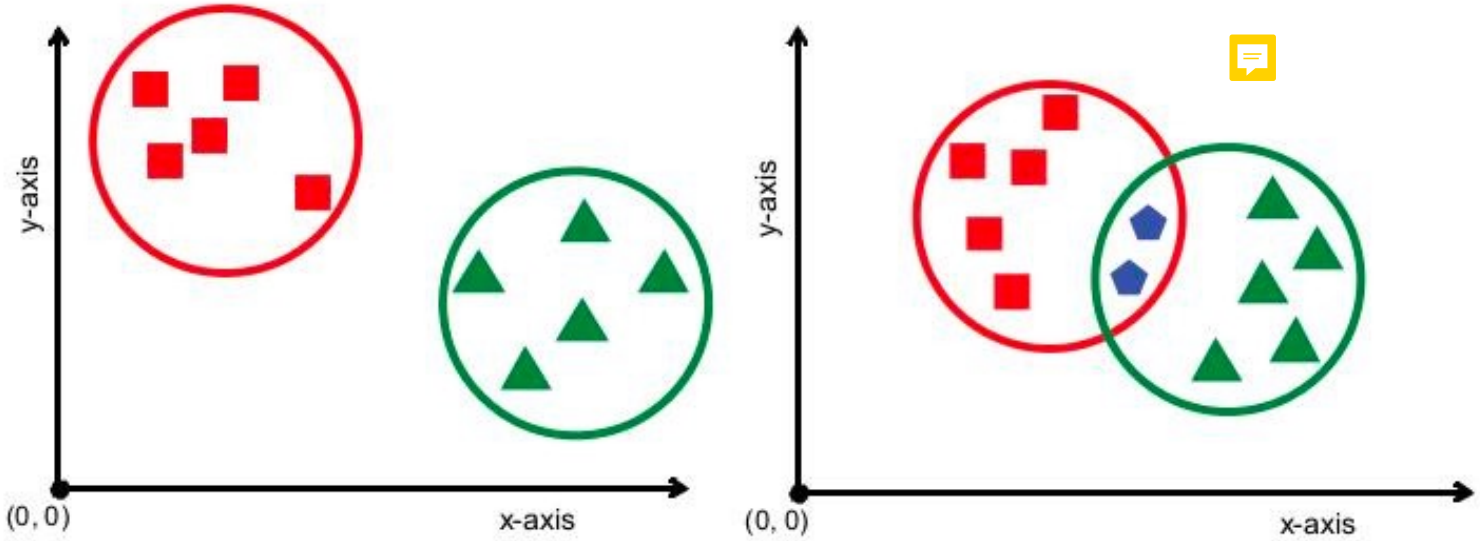
Applications

- **Image segmentation (**partition an image into regions each of which has a reasonably homogeneous visual appearance or which corresponds to objects or parts of objects**)**

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Overlapping Clustering

Exclusive Clustering *Overlapping Clustering*



Exclusive clustering versus overlapping clustering with two centers. In the former, squares and triangles each have their own cluster, and each belongs only to one cluster. In overlapping clustering, some shapes, like pentagons, can belong to both clusters with some probability, so they're part of both clusters instead of having a cluster of their own.

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Overlapping Clustering

- **Fuzzy c-means** (FCM, 1973) is a method of clustering which allows one data point to belong to two or more clusters
- Numerous applications of FCM in virtually every major application area of clustering
- Widely used in Pattern Recognition
  - pattern recognition analyzes the vast amount of available data and extracts useful knowledge from it in the form of patterns → clustering plays a key role in finding the structures in data
- **Different initializations cause different evolutions of the algorithm**. In fact it could converge to the same result but probably with a different number of iteration steps.

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Overlapping Clustering: algorithm

- The FCM algorithm attempts to partition a finite collection of *n* elements $X=\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$ into a collection of *c* fuzzy clusters with respect to some given criterion
- Given a finite set of data, the algorithm returns a list of *c* cluster centres $C=\{\mathbf{c}_1,\ldots,\mathbf{c}_c\}$ and a **partition matrix** $W = w_{i,j} \in [0,1]$, i = 1,...n, j= 1,....c, where each element $w_{i,j}$ **tells the degree** to which element $\boldsymbol{x}_i$ belongs to cluster $\mathbf{c}_j$
- The FCM aims to minimize an objective function

$$\arg\min_{C} \sum_{i=1}^{n} \sum_{j=1}^{c} w_{ij}^{m} \|\mathbf{x}_i - \mathbf{c}_j\|^2$$

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Overlapping Clustering: algorithm

- $1 \leqslant m < +inf$ is a parameter for determining the the degree of fuzziness. In the limit $m \longrightarrow 1$ $w_{i,j}$ converges to 0 or 1 and the FCM coincides with that of K-means
- larger m results in fuzzier clusters
- In the absence of experimentation or domain knowledge, m is commonly set to 2

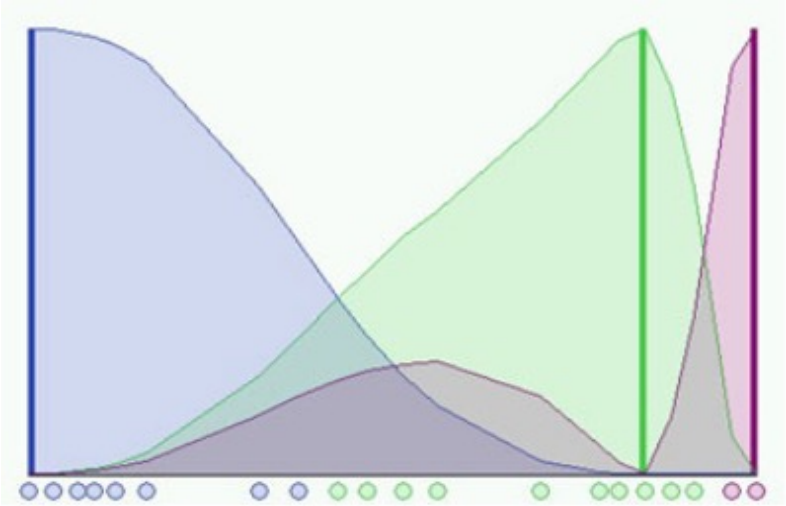
- Disadvantages: the **minimum** is a local minimum, and the results depend on the initial choice of weights

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

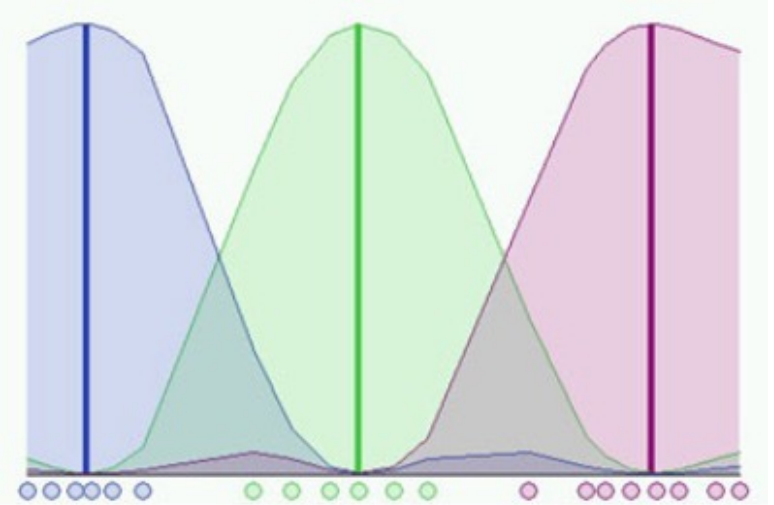
Overlapping Clustering

- 20 data and 3 clusters are used to initialize the algorithm.
- Figure shows the membership value for each datum and for each cluster. **The color of the data is that of the nearest cluster according to the membership function**.

clustering sub-ottimale



clustering ottimale



Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Hierarchical Clustering

Agglomerative bottom-up technique *(1967)*

Given a set of N items to be clustered, and an N*N distance (or similarity) matrix:

1. Start by assigning each item to a cluster, so that if **you have N items, you have N clusters**
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** 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
5. The set of clusters obtained along the way forms a hierarchical clustering

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Hierarchical Clustering

Agglomerative bottom-up technique *(1967)*

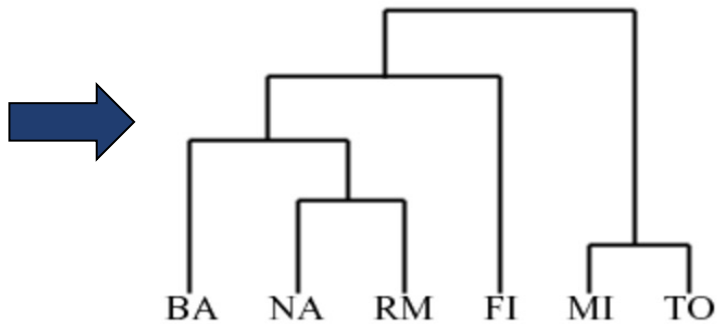
- Step 3 can be done in different ways
- The function used to determine the distance between two clusters, known as the *linkage function*, can be:
  - *single-linkage*
  - *complete-linkage*
  - *average-linkage*
- The formula used to calculate inter-cluster distances is different

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Hierarchical Clustering

- Example: a hierarchical clustering of distances in km between some Italian cities
- Result: a binary tree called **dendrogram**
- The height of each node in the dendrogram is proportional to the dissimilarity between its children

	BA	FI	MI	NA	RM	TO
BA	0	662	877	255	412	996
FI	662	0	295	468	268	400
MI	877	295	0	754	564	138
NA	255	468	754	0	219	869
RM	412	268	564	219	0	669
TO	996	400	138	869	669	0



Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

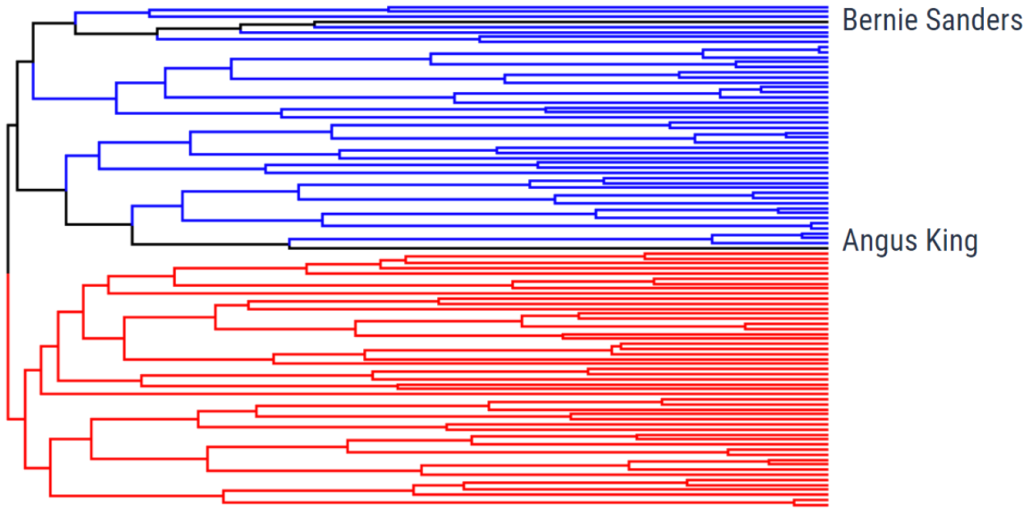
Disadvantages

- they do not scale well: time complexity of at least $O(n^2)$, where n is the number of total objects
- they can never undo what was done previously
- the addition of just a single new instance can radically alter the entire clustering structure

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Applications

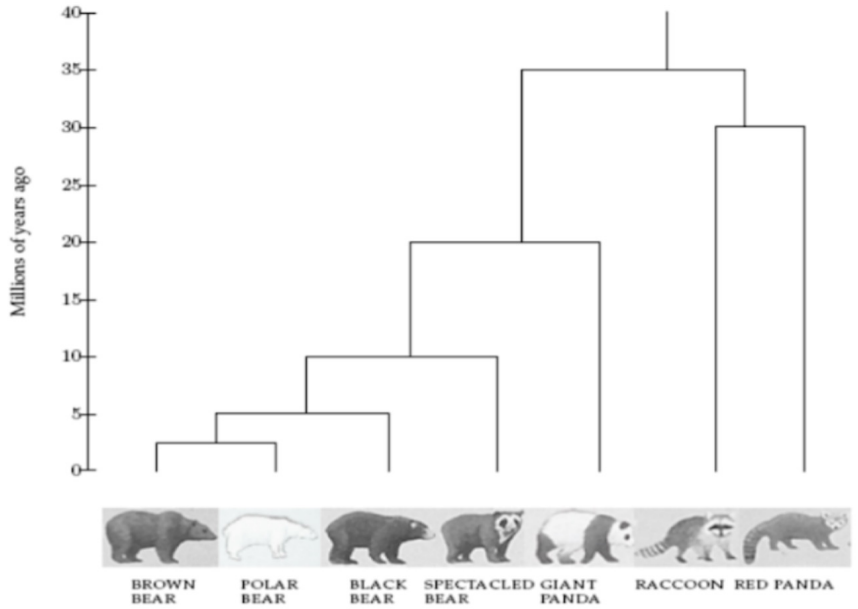
- **Clustering through Twitter**
  - data: which American senators follow which senators
  - the senator similarity is the number of times you end up at certain senator starting from a senator



Reds are Republicans, Blues are Democrats, Blacks are independent

Applications

- **Charting Evolution through Phylogenetic Trees**
  - Data: DNA sequences
  - Calculate the DNA similarities

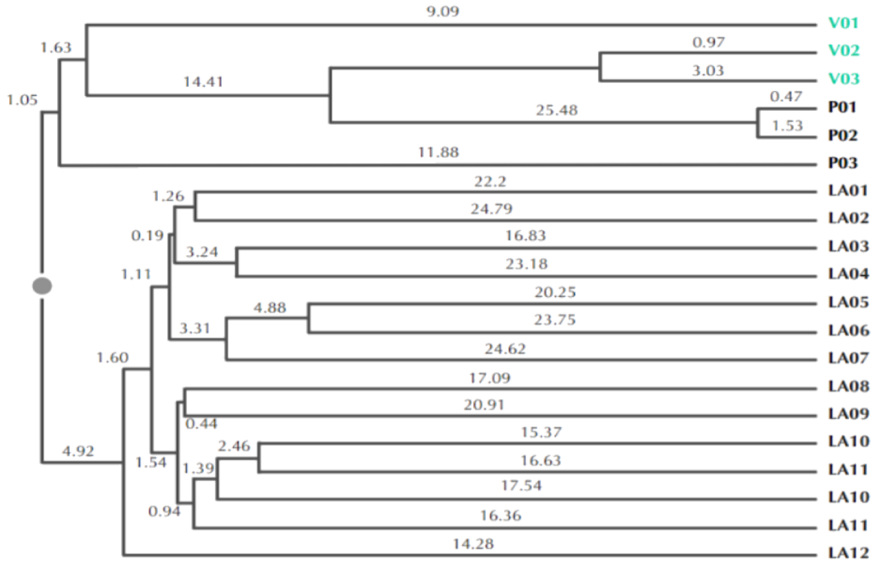


researchers were able to place the giant pandas closer to bears

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Applications

- **Tracking Viruses through Phylogenetic Trees**
  - Data of HIV's DNA sequence used in a court case (HIV of the victim and the accused patient)
  - HIV has high mutation rates → the similarity of the DNA sequence of the same virus depends on the time since it was transmitted



V1–3 are victim's strands, P1–3 are accused patient's, and LA1–12 are the control group

the victim's HIV-1 sequences was most closely related to the patient's HIV-1 sequences (https://www.pnas.org/content/99/22/14292)

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Probabilistic Clustering

Clustering as a Mixture of Gaussians

- each cluster can be mathematically represented by a parametric distribution, like a Gaussian (continuous) or a Poisson (discrete)
- The entire data set is therefore modelled by a *mixture* of these distributions
- The most widely used clustering method of this kind is the one based on learning a *mixture of Gaussians*
- A *Gaussian Mixture* is a function that is comprised of several Gaussians, each identified by $k \in \{1,\ldots, K\}$, where K is the number of clusters of our dataset

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

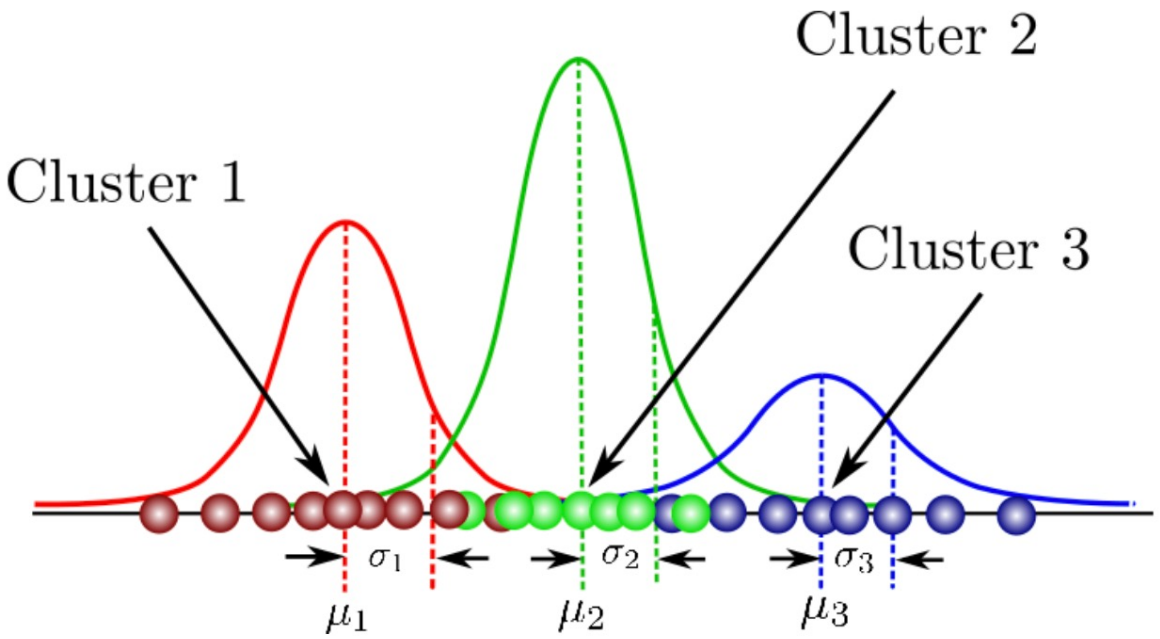
Probabilistic Clustering

Clustering as a Mixture of Gaussians

- A mean μ that defines its centre
- A covariance s that defines its width
- The probability distribution, each for one cluster, describes the distribution of values for members of that cluster
- For a given set of data points, our Mixture of Gaussians would identify the probability of each data point belonging to each of these distributions

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Clustering as a Mixture of Gaussians



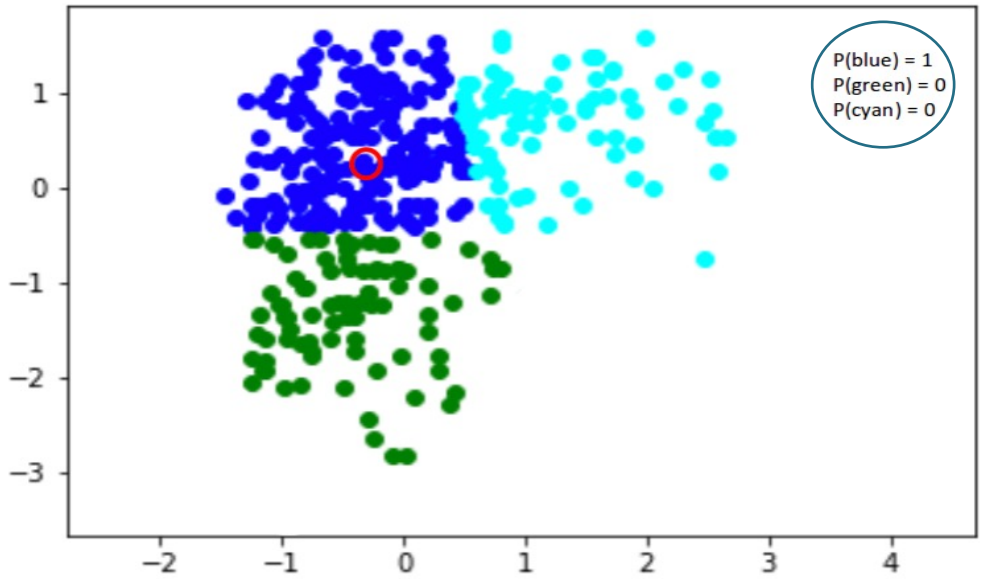
Here, we can see that there are three Gaussian functions, hence $K = 3$. Each Gaussian explains the data contained in each of the three clusters available.

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Clustering as a Mixture of Gaussians

- Example: three Gaussian distributions GD1, GD2, and GD3, with a certain mean (μ1, μ2, μ3) and variance (σ1, σ2, σ3)
- 3 clusters denoted by 3 colors – Blue, Green, and Cyan

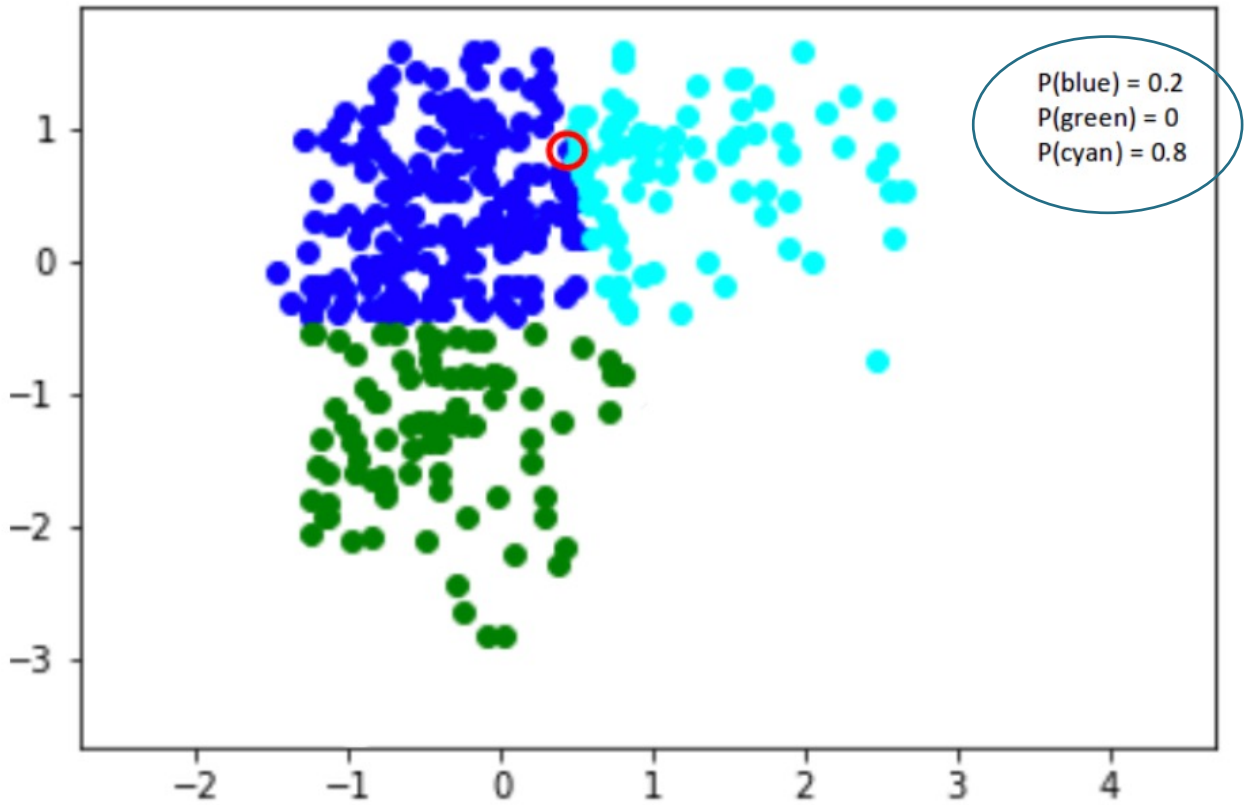
1)



Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Clustering as a Mixture of Gaussians

2)



Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021