Machine Learning

Clustering

Fabio Vandin

December 15^{th}, 2023

Unsupervised Learning

In unsupervised learning, the training dataset is $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m)$
$\Rightarrow$ no target values!

We are interested in finding some interesting *structure* in the data, or, equivalently, to organize it in some meaningful way.

We are going to see the most common unsupervised learning approaches: *clustering*

We are going to focus on the most commonly used techniques:

- k-means
- linkage-based clustering,

There are also other general techniques: dimensionality reduction, association analysis,...

Clustering

Informal definition: the task of identifying meaningful groups among data points.

Definition

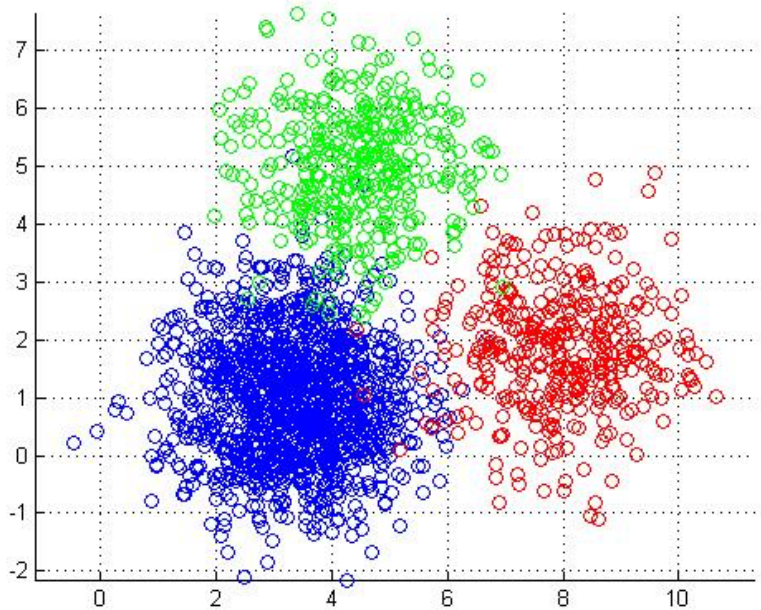
Clustering is the task of grouping a set of objects such that similar objects end up in the same group and dissimilar objects are separated into different groups.

Clustering

Informal definition: the task of identifying meaningful groups among data points.

Definition

Clustering is the task of grouping a set of objects such that similar objects end up in the same group and dissimilar objects are separated into different groups.



Example

- Data: features (e.g. product bought, demographic info, etc.) for a large number of customers
- Goal: **customers segmentation** = identify subgroups of homogeneous customers
- useful for: advertizing, product development, ...

Example (2)

Data:

- rows = genes ($\approx 20 \times 10^3$)
- columns = samples, cancer patients ($\approx 10^3 - 10^4$)
- values = expression of a gene in a patient ($\in \mathbb{R}$)

Goal: find similar cancer samples

- cluster colunms (samples) to find similar subgroups of patients (e.g., *disease subtypes*)

Goal: find genes with similar gene expression profiles

- cluster rows (genes) to deduce function of unknown genes from experimentally known genes with similar profiles

Other Applications

- **Information Retrieval**: clustering is used to *find* topics/categories of documents that are not explicitly given
- **Image Processing**: used for several tasks/applications, including: identification of different types of tissues in PET scans; identification of areas of similar land use in satellite pictures;...
- **Analysis of Social Networks**: detection of communities
- ...