Machine Learning

Clustering

Fabio Vandin

December 16^{th}, 2022

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

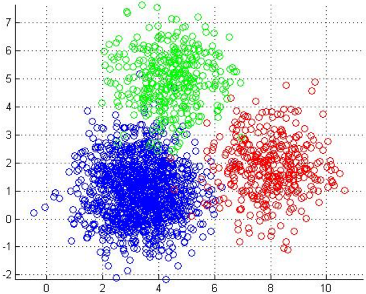
Clustering is the task of grouping a set of objects such that similar objects end up in the same group and dissimilar objects are separated into different groups.

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 columns (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
- ...

Clustering Definition

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.

Note: the definition above is not rigorous and may be ambigouos

$\Rightarrow$ different definitions have been proposed that may lead to different types of clustering. We will see only few of them.

Note: there are some difficulties that are somehow inherent in clustering...

Clustering: Difficulties

transitive property:

$a = b$

and

$b = c$

$\Rightarrow a = c$

Similarity is *not transitive*

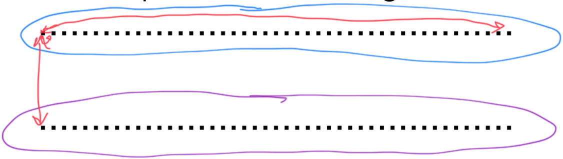
Clustering: Difficulties

Similarity is *not transitive*

$\Rightarrow$ "similar objects in same group" and "dissimilar objects into different groups" may contradict each other...

Example

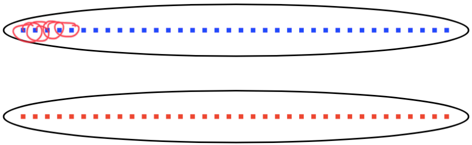
Assume we have data points in $\mathbb{R}^2$ as in figure



Assume we want to cluster the data into $k = 2$ clusters. How should we cluster the data?

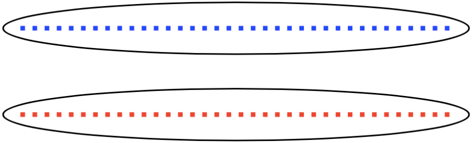
Clustering: Difficulties (continue)

If we focus on "similar objects in same group":

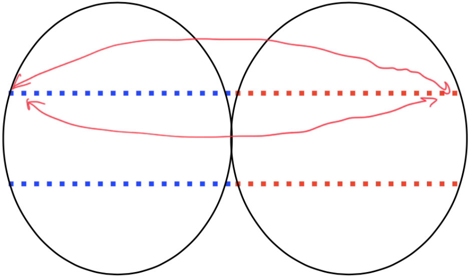


Clustering: Difficulties (continue)

If we focus on "similar objects in same group":



If we focus on "dissimilar objects into different groups":

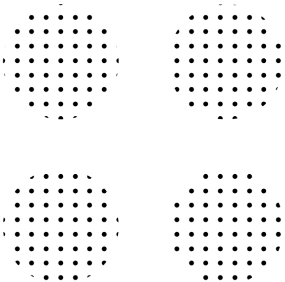


Clustering: Difficulties (continue)

In general we do not have a *ground truth* to evaluate our clustering (*unsupervised learning*)

Example

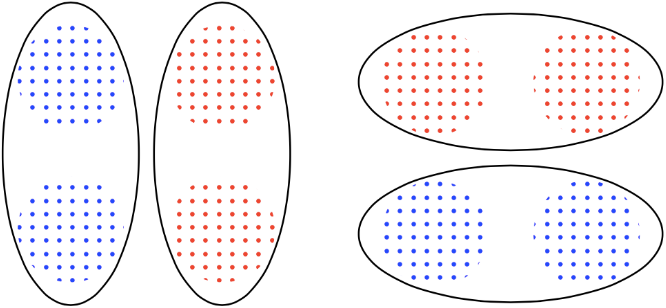
Assume we have data points in $\mathbb{R}^2$ as in figure



Assume we want to cluster the data into $k = 2$ clusters. What is a correct clustering?

Clustering: Difficulties (continue)

The following clusterings are different but both justifiable



In practice: a given set of objects can be clustered in various different *meaningful* ways

A Model for Clustering

Let's formulate the clustering problem more formally:

- **Input**: set of elements $\mathcal{X}$ and *distance* function $d : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_+$, that is a function that
  - is symmetric: $d(\mathbf{x}, \mathbf{x}') = d(\mathbf{x}', \mathbf{x})$ for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$
  - $d(\mathbf{x}, \mathbf{x}) = 0$ for all $\mathbf{x} \in \mathcal{X}$
  - d satisfies the triangle inequality: $d(\mathbf{x}, \mathbf{x}') \leq d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{x}')$

- **Output**: a partition of $\mathcal{X}$ into *clusters*, that is $C = (C_1, C_2, \ldots, C_k)$ with
  - $\cup_{i=1}^{k} C_i = \mathcal{X}$
  - for all $i \neq j$: $C_i \cap C_j = \emptyset$

- **Notes:**
  - sometimes the input also includes the number k of clusters to produce in output
  - sometimes, the output is a **dendrogram** (from Greek *dendron* = tree, *gramma* = drawing), a tree diagram showing the arrangement of the clusters

A Model for Clustering (continue)

Sometimes instead of a distance function we have a similarity function $s : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_+$, that is a function that:

- is symmetric: $s(\mathbf{x}, \mathbf{x}') = s(\mathbf{x}', \mathbf{x})$ for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$
- $s(\mathbf{x}, \mathbf{x}) = 1$ for all $\mathbf{x} \in \mathcal{X}$

Choice of distances/similarity:

- depends on the type of data
- different distances may be used for the same dataset
  $\Rightarrow$ choice of distances may have an impact on the results

Classes of Algorithms for Clustering

1. Cost minimization algorithms
2. Linkage-based algorithms