School of Electronic Engineering and Computer Science
Queen Mary University of London

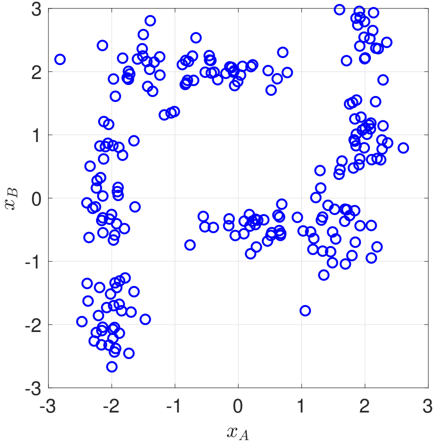
ECS7020P Principles of Machine Learning
Unsupervised learning: Structure analysis

Dr Jesús Requena Carrión

24 Nov 2022

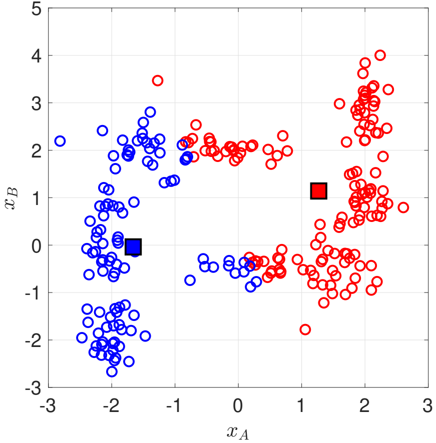
Agenda

Non-convex clusters



Non-convex clusters: K-means

K-means produces spherical clusters, as samples are arranged around a prototype.



Density-based clustering: DBSCAN

In a **non-convex** cluster, we can reach any sample by taking small jumps from sample to sample.

Non-convex scenarios suggest a different notion of cluster as group of samples that are **connected**, rather than simply close: *if I am similar to you, and you are similar to them, I am similar to them too*.

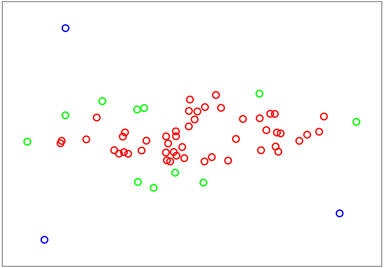
This notion of cluster as a group of connected samples is behind many clustering algorithms, such as DBSCAN (*density-based spatial clustering of applications with noise*).

DBSCAN belongs to the family of **density-based** algorithms, where an estimation of the density of samples around each sample is used to partition the dataset into clusters.

DBSCAN

DBSCAN defines two quantities, a **radius** r and a **threshold** t. A density is first calculated as the number of samples in a neighbourhood of radius r around each sample (excluding itself). Then, three types of samples are identified:

- **Core**: its density is equal or higher than the threshold t.
- **Border**: its density is lower than the threshold t, but contains a core sample within its neighbourhood.
- **Outlier**: Any other sample.



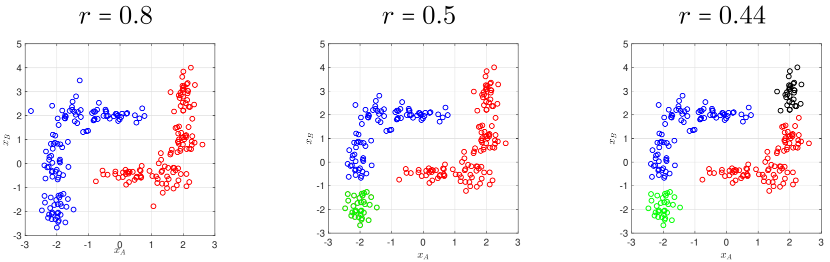
DBSCAN

The DBSCAN algorithm proceeds as follows:

- Identify core, border and outlier samples.
- Pair of core samples that are within each other's neighbourhood are connected. Connected core samples form the **backbone** of a cluster.
- Border samples are assigned to the cluster that has more core samples in the neighbourhood of the border sample.
- Outlier samples are not assigned to any cluster.

DBSCAN

Solutions for a threshold $t = 3$ and different radii.



Agenda

Hierarchical clustering

Given a dataset consisting of N samples, there exist two **trivial** clustering solutions: **one single cluster** that includes all the samples, and the solution where **each sample is a cluster** on its own.

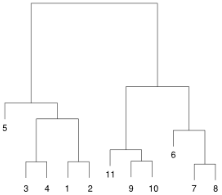
K-means produces K clusters, but we need to choose K within $1 \leq K \leq N$. In DBSCAN clusters are discovered automatically, but the final number of clusters depends on the values of the radius r and the threshold value t.

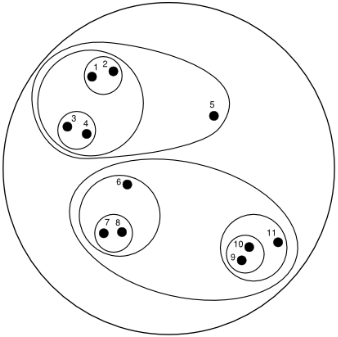
This ambiguity ultimately reveals that **the structure of a dataset can be explored** at different levels that expose different properties.

Hierarchical clustering

Hierarchical clustering is a family of clustering approaches that proceed by progressively building clustering arrangements at **different levels**.

The resulting collection of clustering arrangements is hierarchical in the sense that a cluster in one level contains all the samples from one or more clusters in the level below.





Hierarchical clustering

The representation of the relationship between clusters at different levels is called a **dendrogram**. At the bottom we find the arrangement where each sample is one cluster and at the top, the whole dataset.

There exist two basic strategies to build a **dendrogram**:

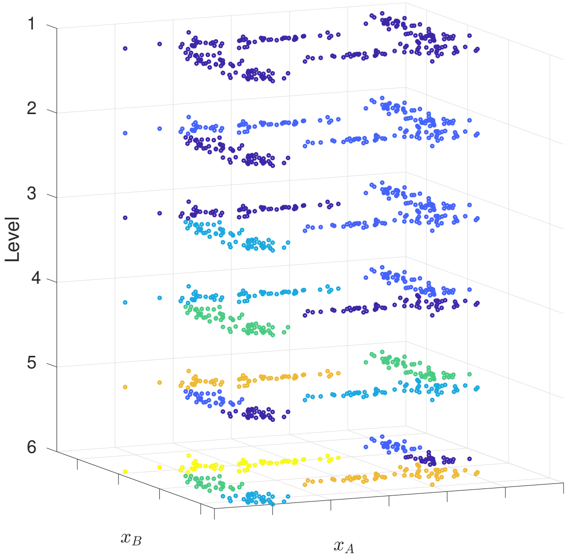
- The **divisive** or top-down approach splits clusters starting from the top of the dendrogram and stops at the bottom level.
- The **agglomerative** or bottom-up merges two clusters, starting from the bottom until we reach the top level.

Hierarchical clustering

There are different options to decide which clusters to merge or split at each level. Common strategies in agglomerative clustering include:

- **Single linkage**: uses the distance between the two closest samples from two clusters. This option results in clusters of arbitrary shapes.
- **Complete linkage**: uses the distance between the two further samples from each pair of clusters. This choice produces clusters that tend to have a spherical shape.
- **Group average**: uses the average distance between samples in two cluster and also produces spherical shapes, although they are more robust to outliers.

Hierarchical clustering



Agenda

Unsupervised learning

- Unsupervised learning provides with answers for the basic question *where is my data? (in the attribute space)*
- Our answer is a mathematical/computer model. This model can tell us where in the space we have samples (clustering) or the probability to find a sample in a region within the space (density estimation).
- Sometimes we say that our data is unlabelled. What we **really mean** is that we don't treat any attribute as a label that we want to predict. Datasets are neither labelled nor unlabelled.
- Lacking such a target as a label means that **our quality metric is not obvious**.

Clustering

- **K-means** is a prototype-based clustering that produces spherical clusters where K is a hyperparameter that has to be set.
- **DBSCAN** is a density-based option suitable for non-convex scenarios and does not require specifying the number of clusters. We need to set r and t and they determine the final number of clusters.
- **Hyerarchical clustering** allows to explore the structure of a dataset at multiple levels.

Comparing clustering solutions

- We could consider **comparing** the solutions from two different algorithms. However, if they use different definitions of clustering quality, this comparisons will make little sense.
- Clustering is ultimately implemented with an **application** in mind so we should create a final notion of clustering quality based on the specific goals of the application.

What about component analysis?

Component analysis allows us to identify the **directions in the space** that our data are aligned with. This can be useful to transform our dataset, clean it and reduce its dimensionality.