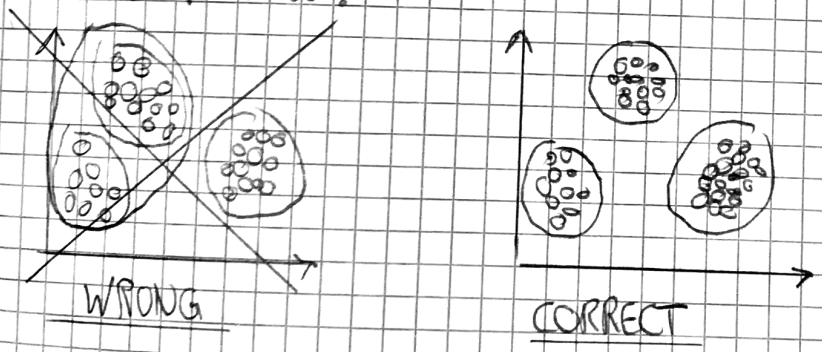
Question on clustering

Describe in your own words what three clustering approach we learned and use appropriate IT-Jargon to describe. Further point out from each algorithm the advantages and disadvantages.

1) We have learned K-Means clustering, Hierarchical clustering and DBSCAN,
all of them are unsupervised learner which is used when data are have no pre-defined labels.

2) Kmeans is a prototype based learner which means that each cluste is represented by a prototype with a:
- Centroid (average) → continous features
- Medoid (The most representative or most frequently occuring point.)
  ↳ In case of categorical features.

Further it is a partitioneal learner, which basically means a top-down approach, and divides data into non-overlapping subsets of cluser.

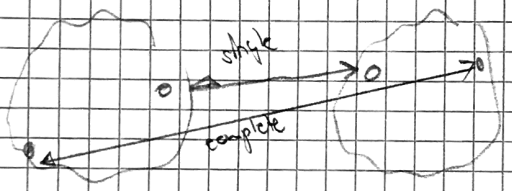


K-means is an optimisation problem as he tries to minimize the points.

3) ⊕ simple to understand
⊕ works well on small and large datasets
⊕ performant
⊖ Not suitable for special structures
⊖ sensitive to initiated seeds
⊖ sensitive to outliers

3) Hierarchical clustering generates sets of cluster Until at last all of the points are all in a single cluster or in other terms a set of nested clusters organized as a hierarchical tree.

- There are two main approaches to hierarchical clustering
  - Divisive (TOP Down)
  - Agglomerative (Bottom up Approach)
- To link the datapoints to a cluster there is linkage (Distance between the cluster) used also, there are two kinds of approaches
  - Single linkage (closest points)
  - Complete linkage (Furthest points)



an advantage is that the number of clusters are not fixed in advance, and the data can be visualized by a dendrogram, this is a Tree-like visual representation of the observations.

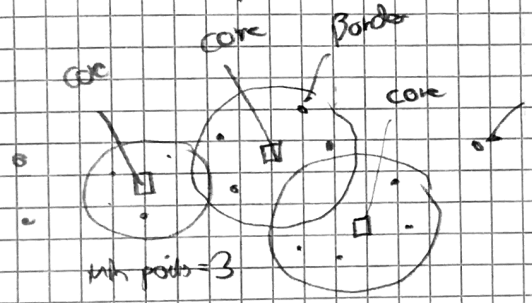
⊕ Dendrogram
⊕ No need to specify #cluster in advance
⊖ Not appropriate for large Datasets

4) DBSCAN

The Density-based Spatial Clustering of Application with noise is a much more advanced algorithm and is much more performant on data with outliers.
However many things need to be given at initial stage such as the specification of radius and the minimum of numbers of Points.

- The notion of density in DBSCAN is defined as the number of points within a specified radius E.

- There are three kinds of special labels



- Core points = where density is at max
- Border = has fewer points as neighbors as min points.
- Noise = All other points that are neither core nor border

⊕ good for outliers
⊕ arbitrary spherical shapes
⊕ different from K-means and hierarchical
⊕ Input parameter.