Clustering

Unsupervised Machine Learning by tdi@ek.dk

Agenda

- Clustering vs Classification
- Clustering Algorithms
- Clustering Validation Measures

Intended Learning Outcomes

- To be able to explain the difference between supervised and unsupervised learning
- To be able to compare and combine the advantages of classification and clustering
- To get familiar with methods for clustering and be able to use them
- To learn about image segmentation

Supervised vs Unsupervised

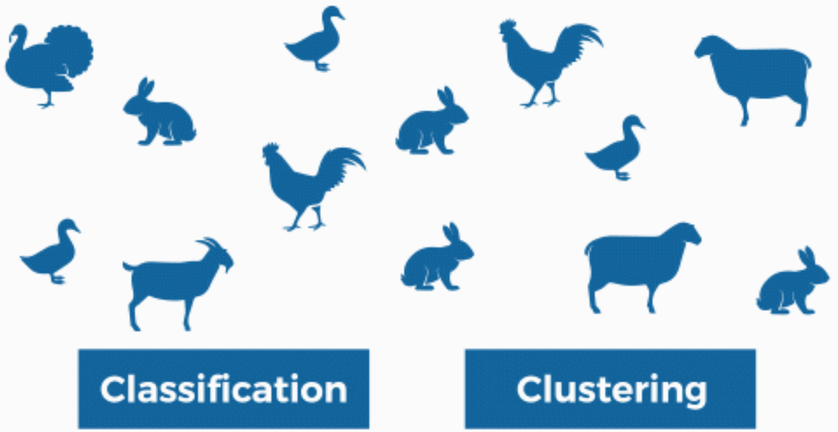
Clustering vs Classification

Classification

- association with a pre-defined classes

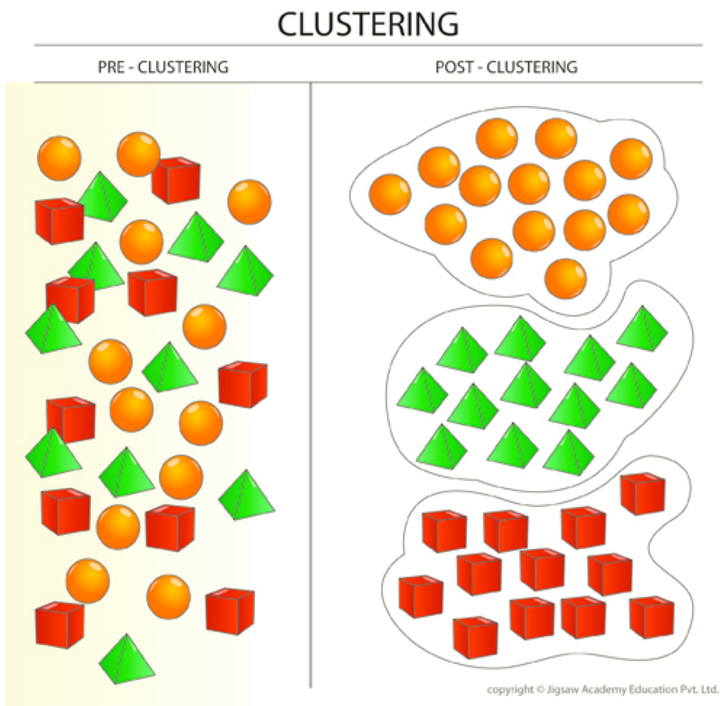
Clustering

- splitting into groups on similarities



https://blog.bismart.com/en/classification-vs.-clustering-a-practical-explanation

Clustering



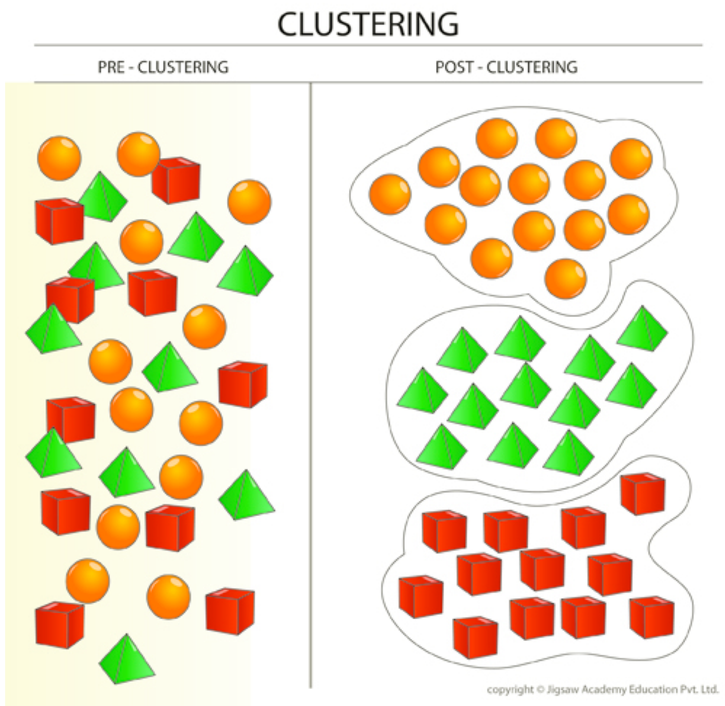
Separating objects into groups regarding their similarity and dissimilarity

- Input: objects
- Output: clusters

Objectives

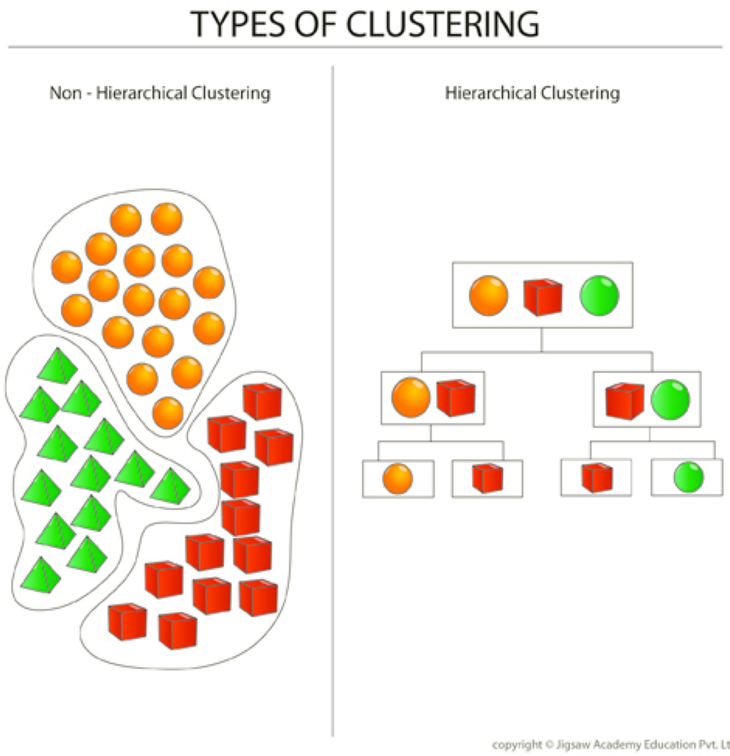
- creating partitions or clusters and
- then classifying objects into these clusters

Criteria for Success



1. All objects are associated to a cluster
2. The objects inside a cluster are much close to each other
3. The different clusters are much distant from each other

What for?



- For understanding the objects and the groups
- As utility for labelling clusters and then using them in supervised learning

- Applications
  - Market research
  - Pattern recognition
  - Information retrieval, taxonomies
  - Categorization, segmentation, aggregation of data
  - Image processing
  - Gaming

K-Means

Machine learning method for unsupervised learning

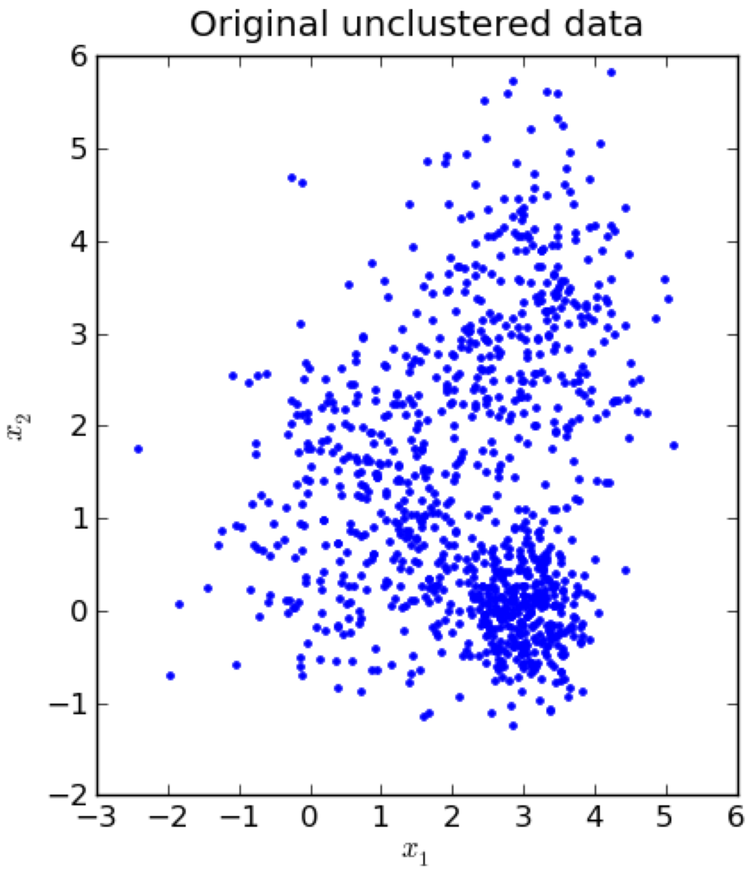
Input

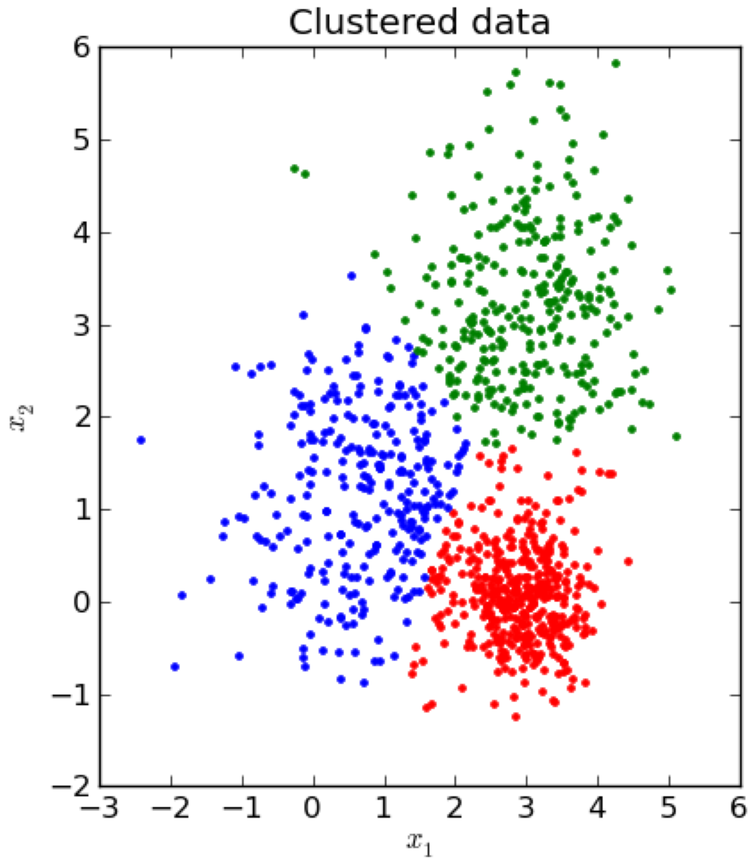
set of feature vectors

$X = (x_1, x_2, \ldots x_n)$

Output

number of clusters K

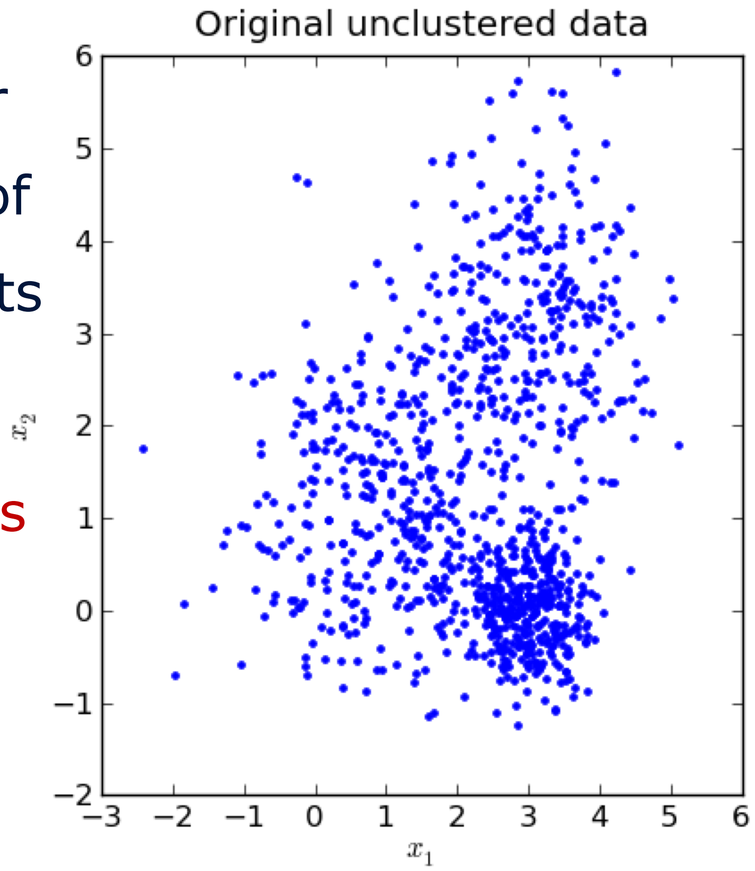


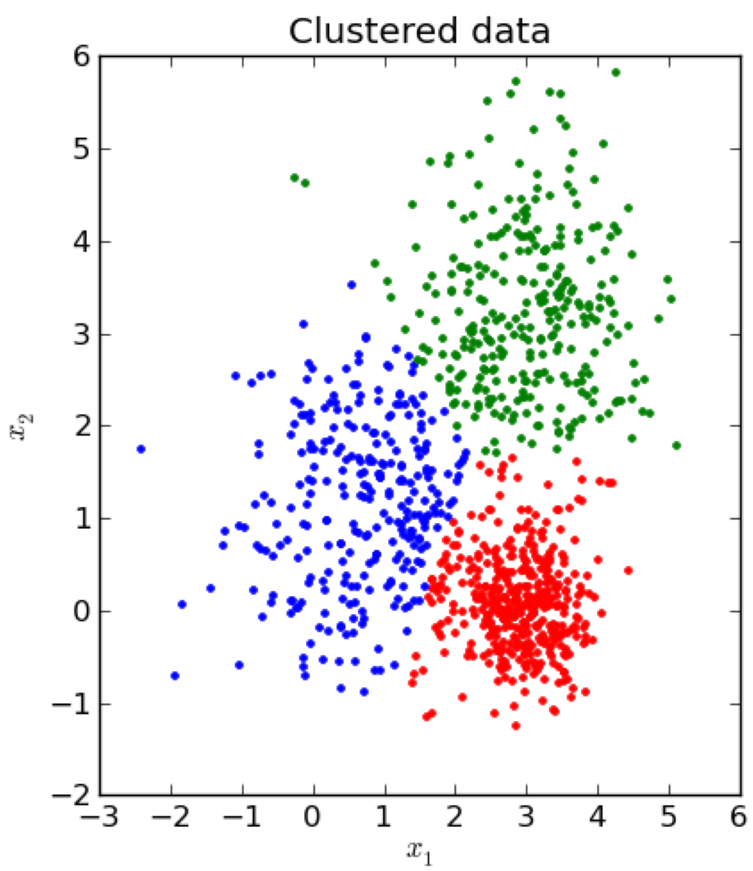


Clusters

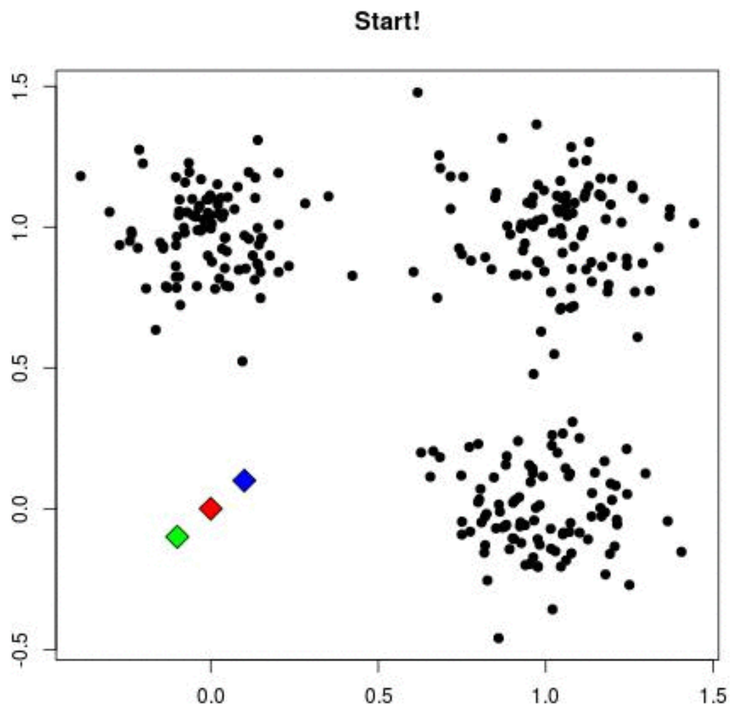
Assumptions

- The centroid of a cluster is the arithmetic mean of distances to all the points belonging to the cluster
- Each point is closer to its own cluster center than to other cluster centers





How Does K-Means Work?



1. Plot the data
2. Decide on number of clusters
3. Select random centroids
4. Calculate the distance of each point to each centroid

K is selected experimentally

1. Assign each point to the closest centroid
2. Recalculate the centroids

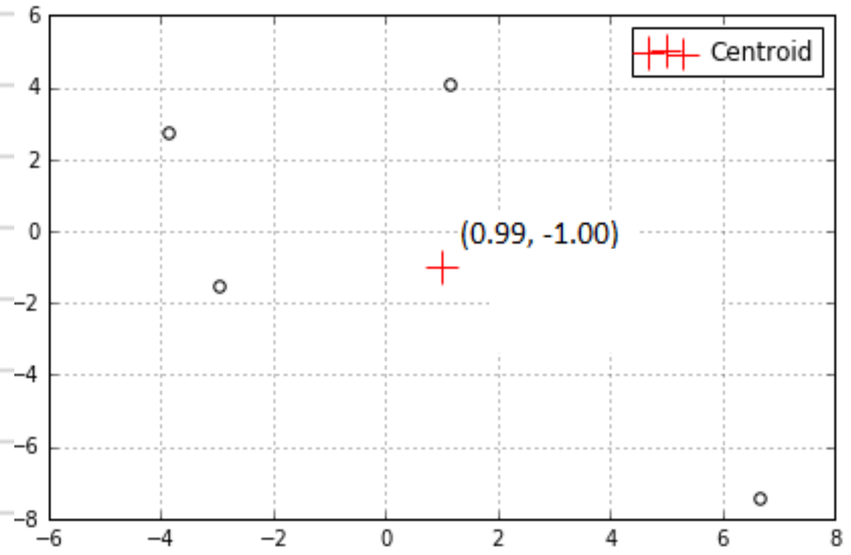
$$c_i = \frac{1}{|S_i|} \sum_{x_i \in S_i} x_i$$

3. Repeat the last two steps
4. Stop when there are no significant changes

Centroid

Centroid – a point, average representation of the other points

	DATA	
	X	Y
	4.03	-2.95
	-3.88	2.73
	1.15	4.10
	-2.98	-1.49
	6.65	-7.40
Means ->	0.99	-1.00



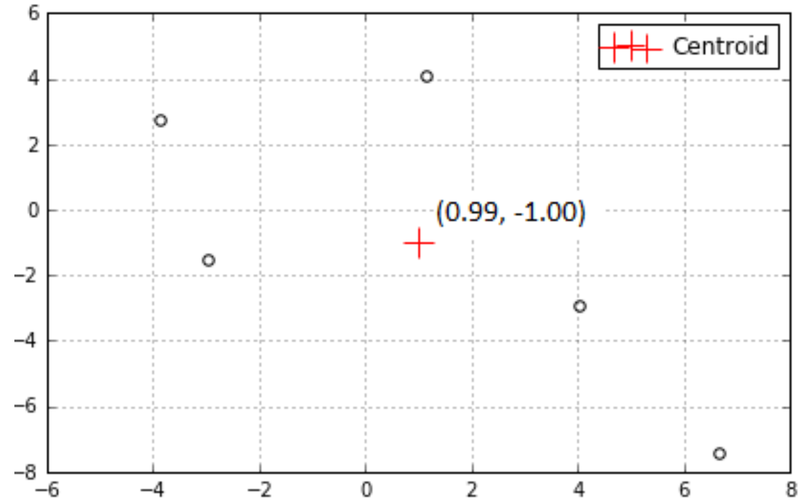
https://avidml.wordpress.com/2016/10/29/easily-understand-k-means-clustering/

Distortion

Distortion - estimation of the error of the centroid

- sum of square errors (SSE)

$$\sum \left(Data\ X_i - Centroid\ X\right)^2 + \left(Data\ Y_i - Centroid\ Y\right)^2 \ldots$$

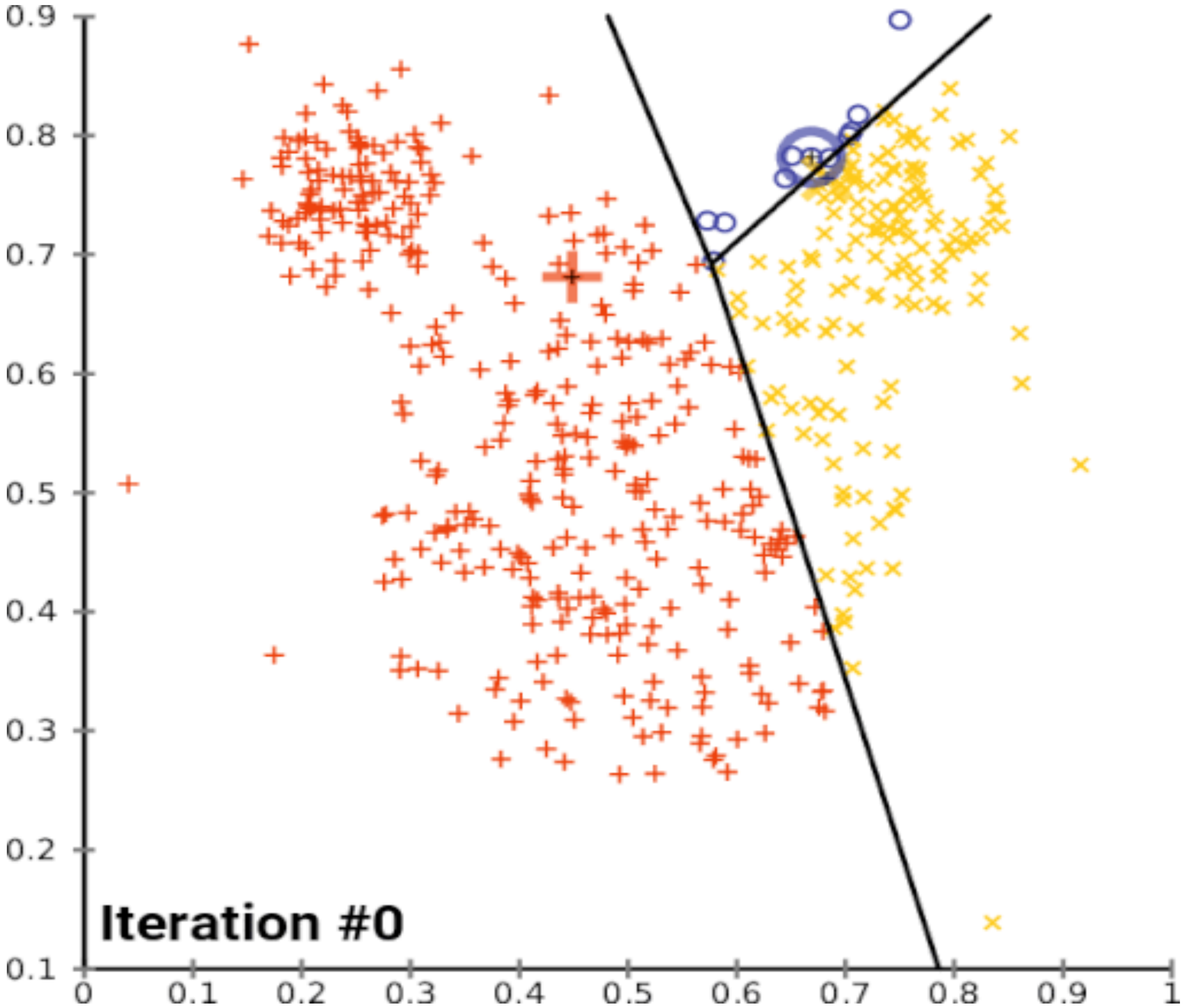



```
(4.03 - 0.99)^2 + (-2.95 - -1.00)^2 + (-3.88 - 0.99)^2 + (2.73 - -1.00)^2…
```

	DATA		Centroid	
	X	Y	X	Y
	4.03	-2.95	0.99	-1.00
	-3.88	2.73		
	1.15	4.10		
	-2.98	-1.49		
	6.65	-7.40		
Means ->	0.99	-1.00		

https://avidml.wordpress.com/2016/10/29/easily-understand-k-means-clustering/

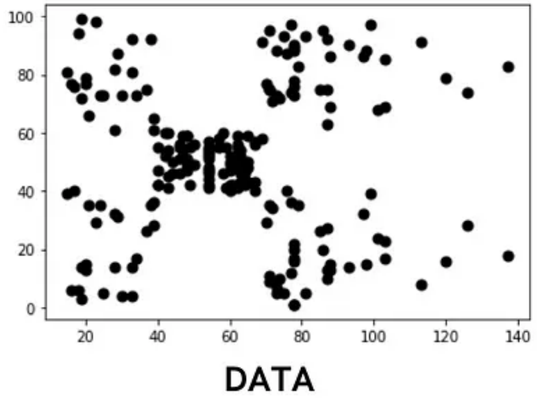
Demo

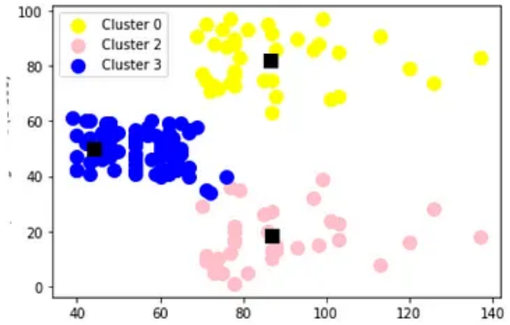




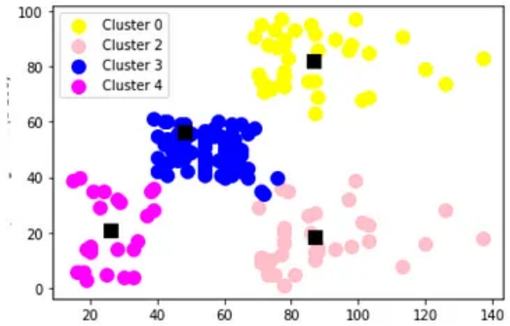



Best K?

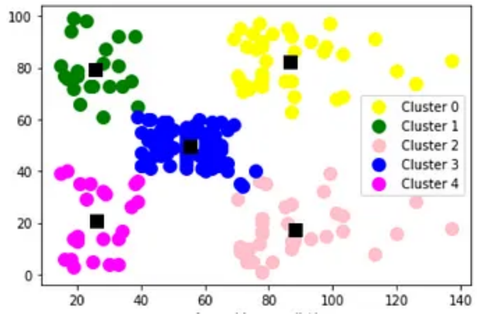




k – 3



k – 4

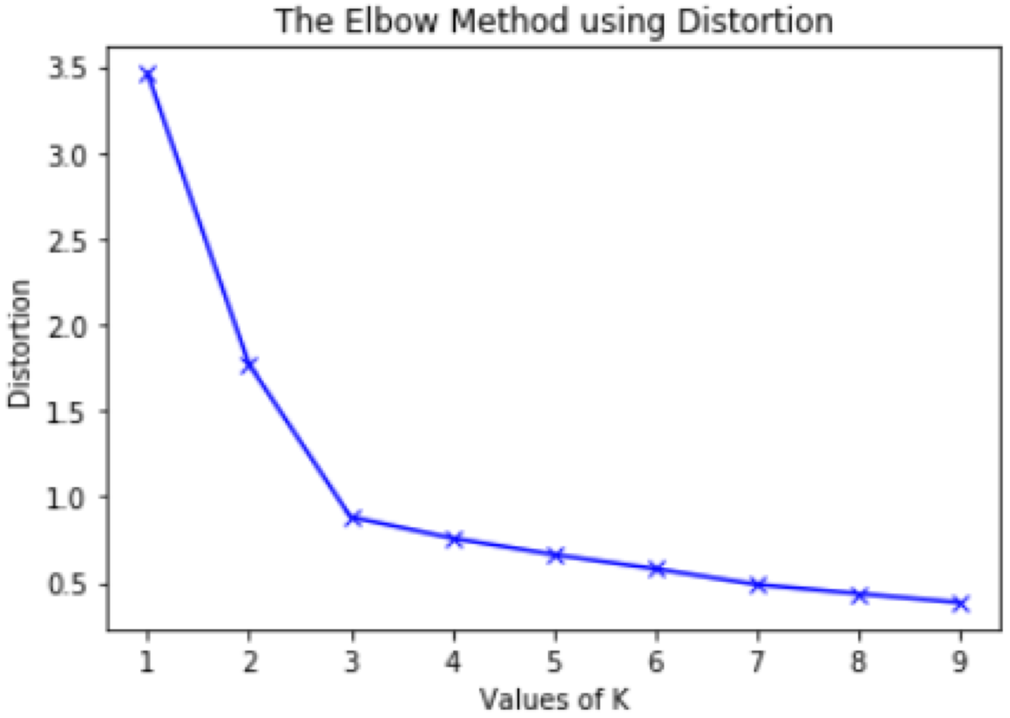


k – 5

Best K?

Elbow Method

- Measuring the distortion of various potential clusters
- The chosen number of clusters is the one, which produces minimal distortion



Best K?

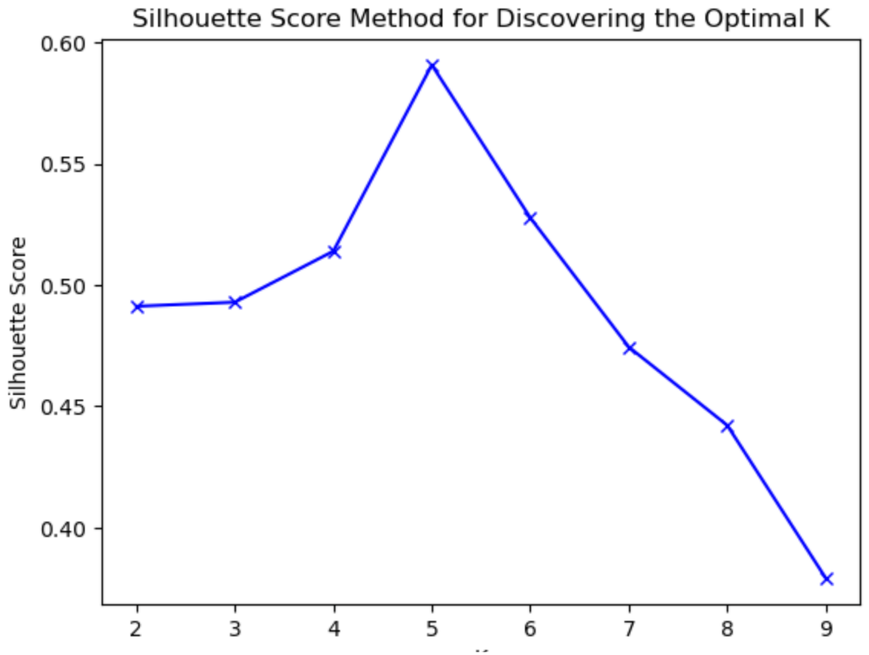
Silhouette Score

Estimates how close a point is to the other points in the cluster, and how far is it from another cluster

```
silhouette_score = (p - q) / max(p,q)
```

`p` – mean distance to the nearest cluster

`q` – mean distance to the points in the own cluster



Understanding the Silhouette Score

The silhouette score is a metric used to evaluate the quality of a clustering result, ranging from -1 to +1.

- Interpretation

  +1: data points are well-suited to their own cluster and are distinctly separated from other clusters

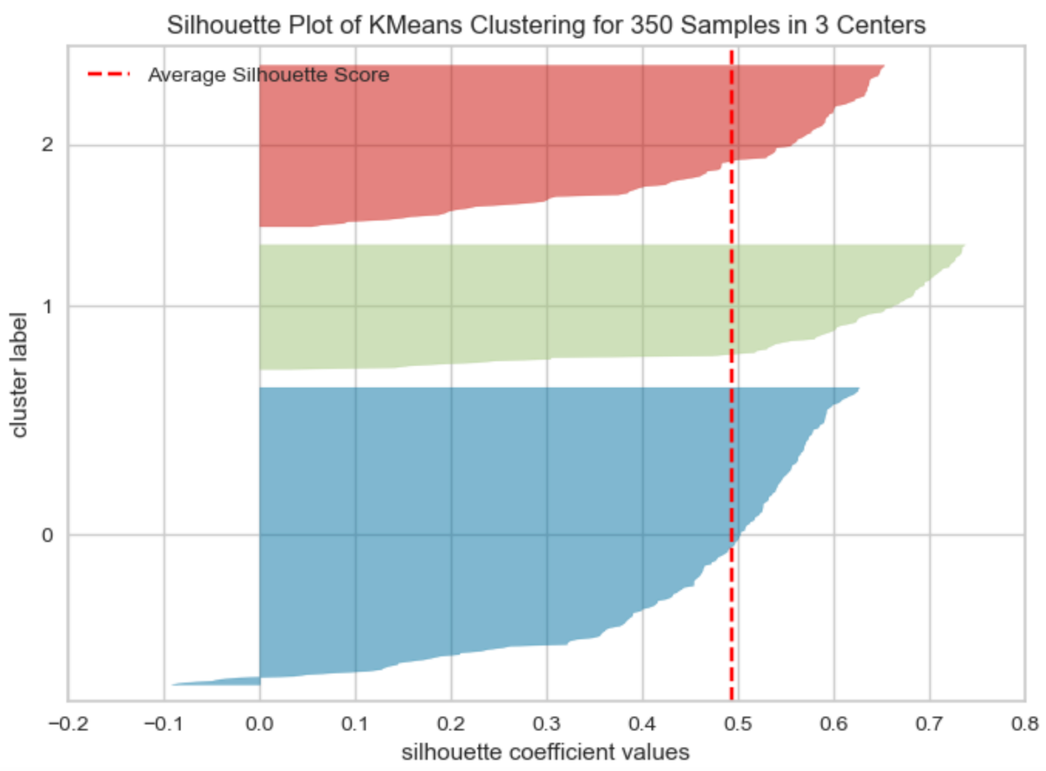
  0: the clusters are indifferent, overlapping, or the points are on the boundary between clusters.

  -1: a data point is likely assigned to the wrong cluster

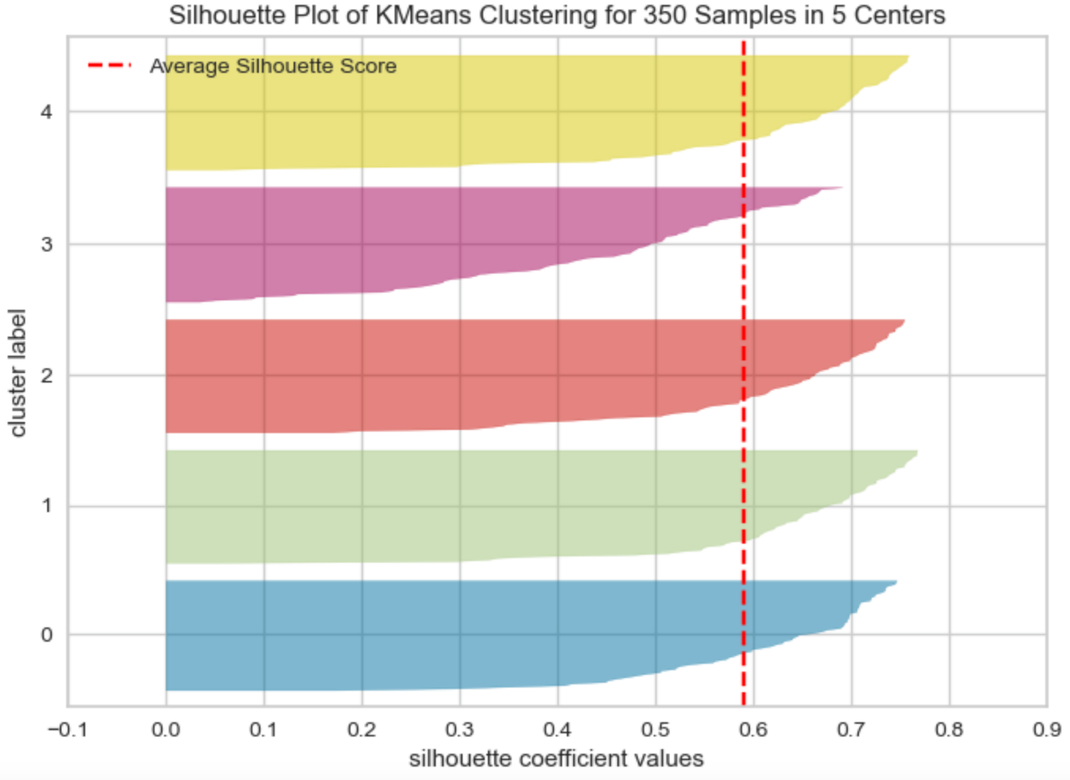
- An average silhouette score of over 0.7 is considered strong, while a score over 0.5 is reasonable, and over 0.25 is weak.

Silhouette Scores of All Clustered Data

Not Good Clustering

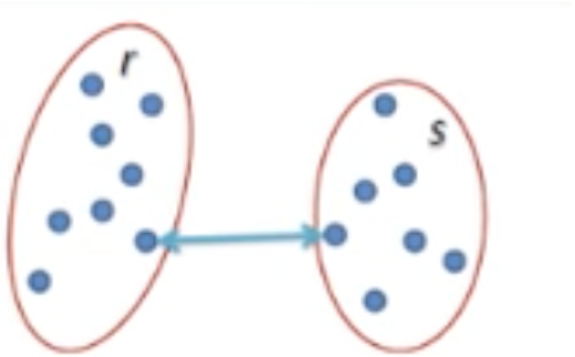


Good Clustering



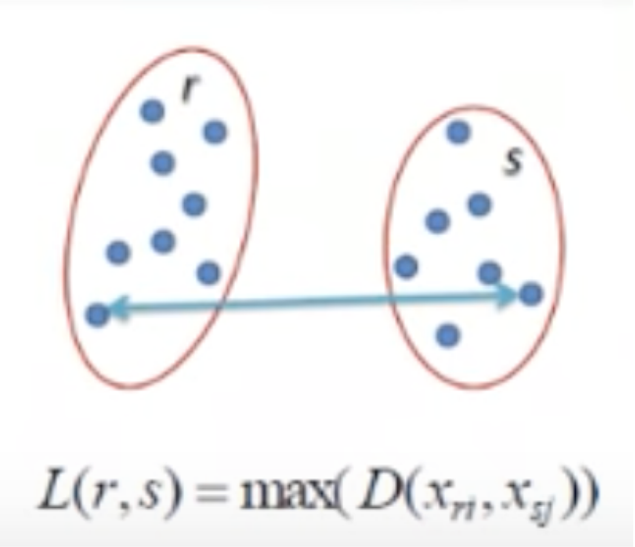
Measures of Dissimilarity Between Clusters

Linkage Criteria

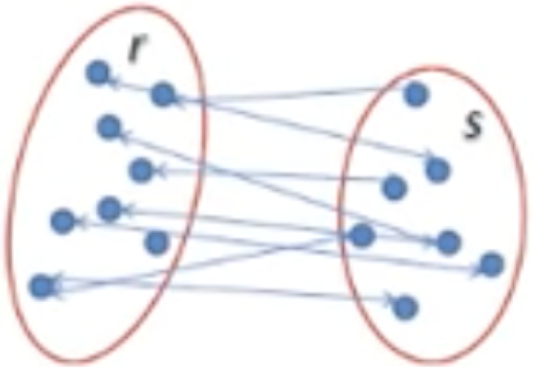


$$L(r,s) = \min(D(x_{ri}, x_{sj}))$$

Single linkage



$$L(r,s) = \max(D(x_{ri}, x_{sj}))$$

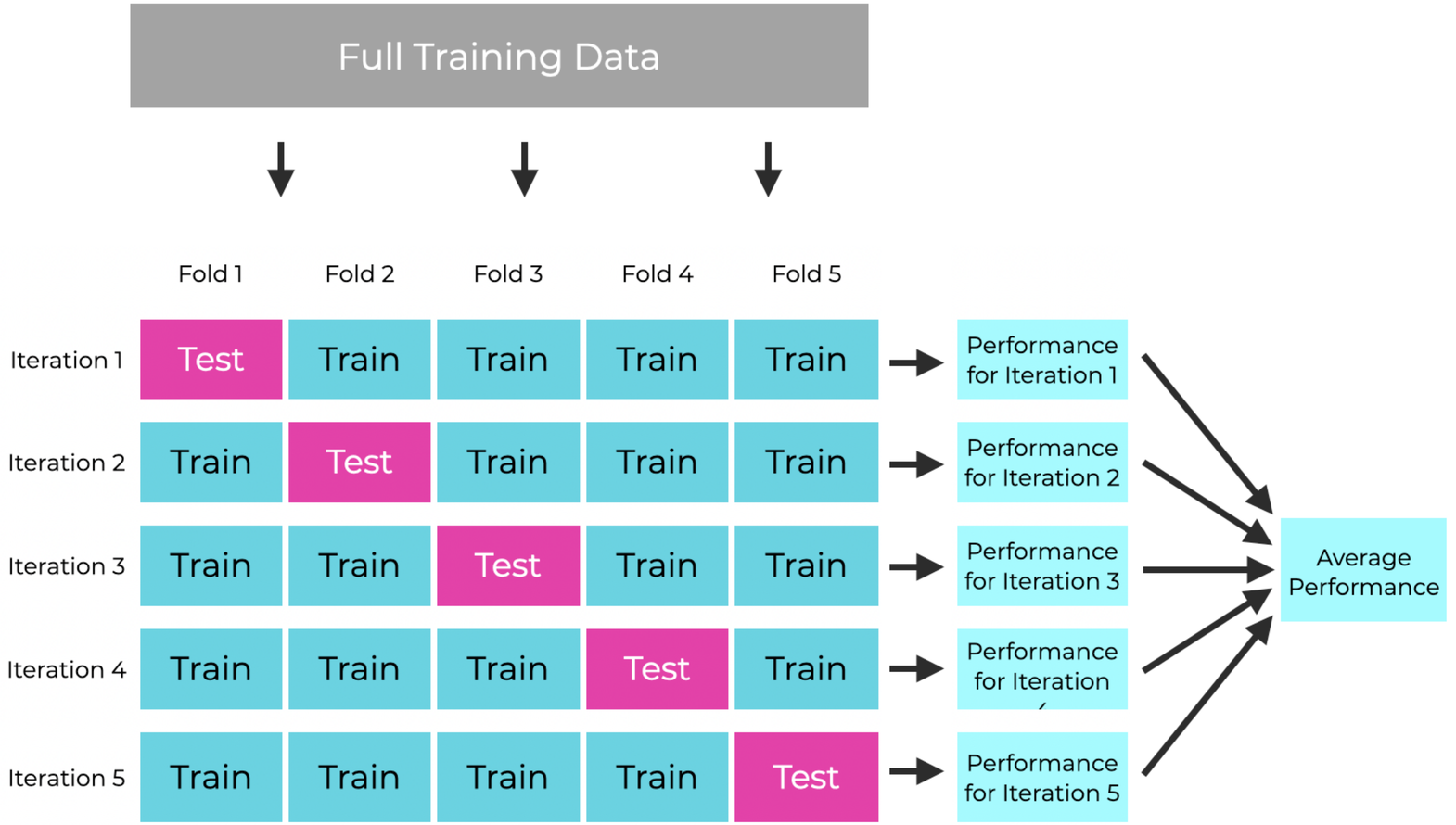


$$L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj})$$

Complete linkage

https://medium.com/100daysofmlcode/day-32-of-100daysofml-a36a120cbb1e

Improving Quality by Cross-Validation



See also:

https://www.youtube.com/watch?v=fSytzGwwBVw&t=25s

K-Means Features

Advantages

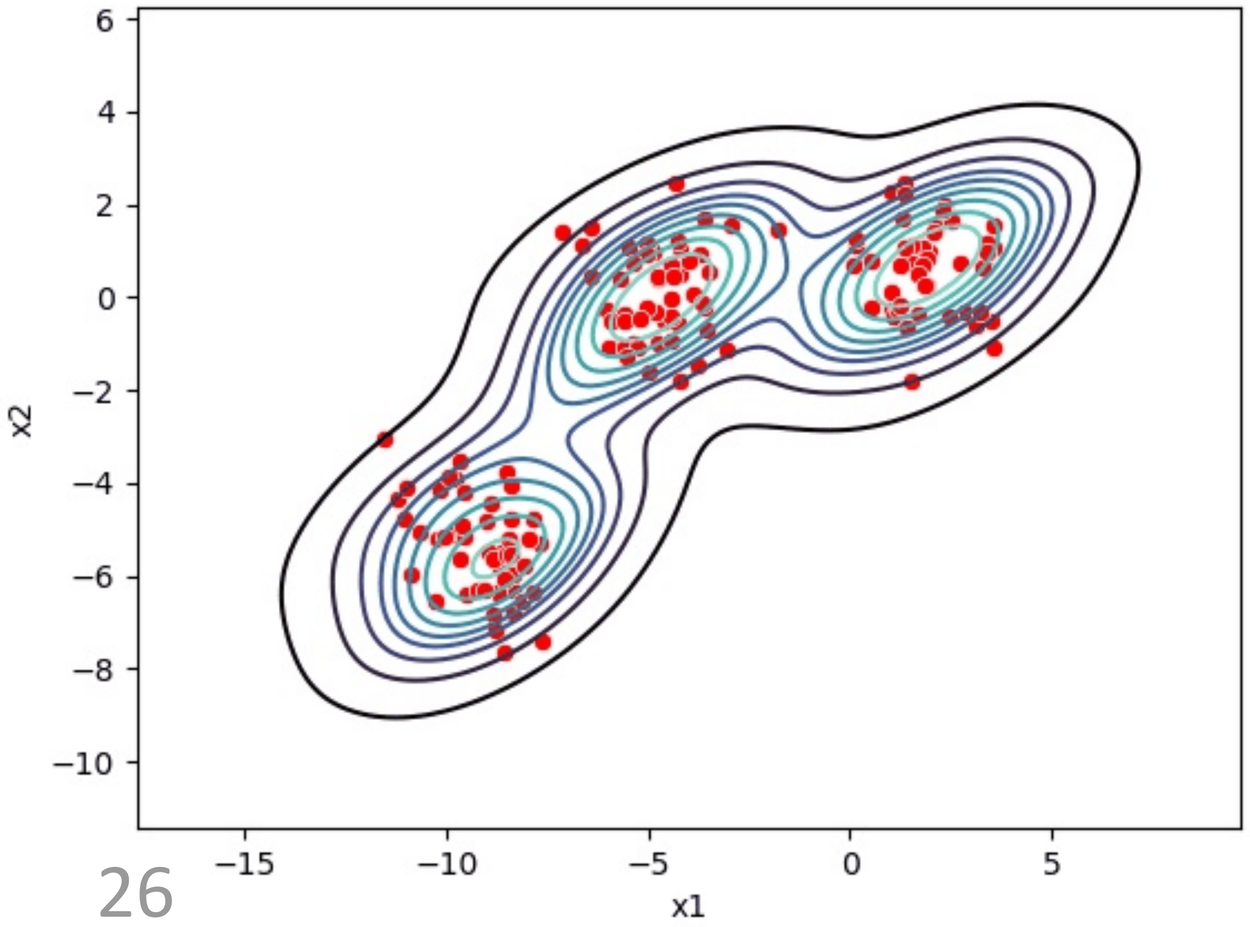
- Simple
- Effective – always gives solution (not always good enough ;)
- Efficient – quick

Disadvantages

- Good for compact clusters only, sensitive with outliers and noise
- Uses numeric data only
- Features should have similar measure scale
- Difficult to estimate the quality

Mean Shift

Machine learning method for unsupervised learning



Mean Shift

Random points come distributed in different density regions

Mean Shift algorithm is an unsupervised clustering algorithm that aims to discover blobs in a smooth density of samples.

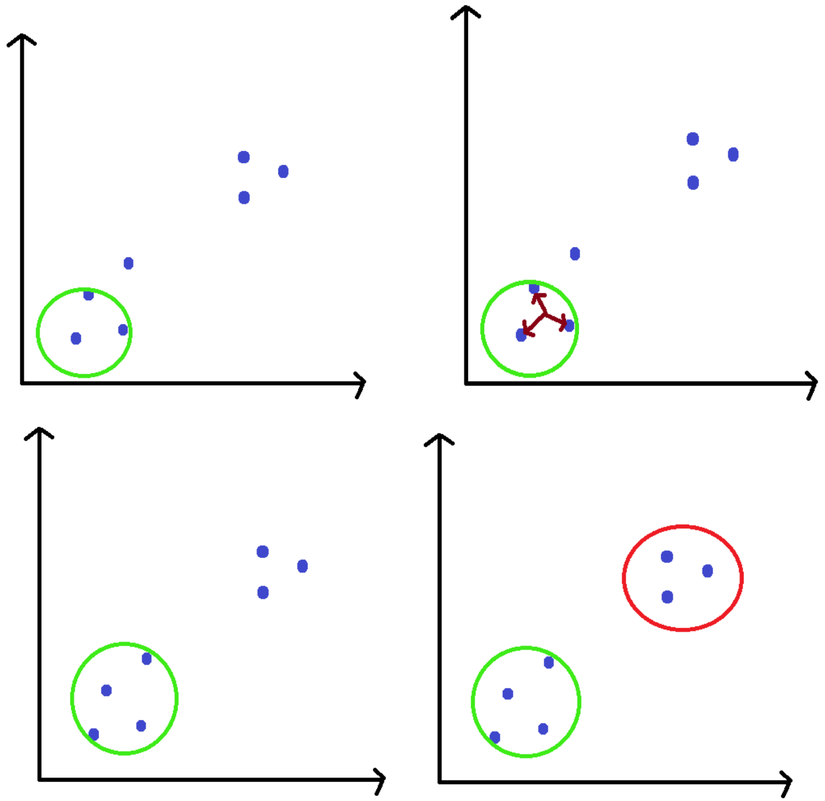
It is a centroid-based algorithm that works by updating candidates for centroids to be the mean of the points within a given region (also called bandwidth)

(https://ml-explained.com/blog/mean-shift-explained)

How Does It Work?

- Based on a sliding window – kernel - a circle with a predefined radius - bandwidth
- Attempts to find the most dense areas of data points
  - tries to locate the center points of densed clusters
  - iteratively updating candidates for center point
  - center point is the mean of the points within the sliding window

How Does It Work?



- Starts with a window centered at randomly selected point X1 and small radius r
- The density within the window is proportional to the number of points inside
- Calculates the mean of all points inside the window as a new center
- Shifts the window by moving the original center to the calculated mean point
- Re-calculates the mean again and moves towards the new point, which would provide higher density
- Stops, when there are no more areas with higher density, than the current one - convergence

https://iq.opengenus.org/mean-shift-clustering-algorithm/

Advantages and Disadvantages

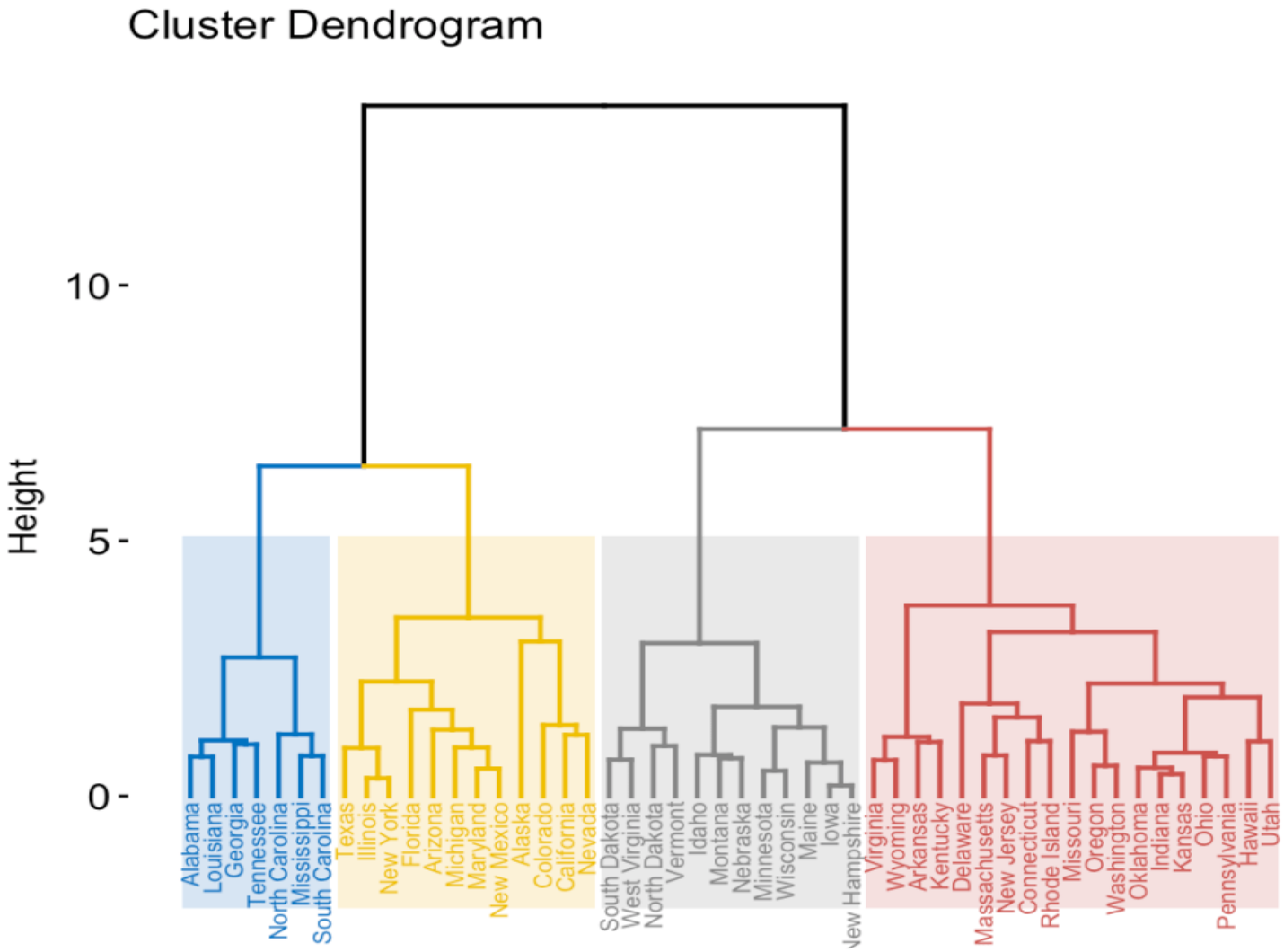
Advantages

- automatically discovers the number of clusters
- cluster centers automatically converge towards the points of maximum density

Disadvantage

- the radius has to be pre-defined
- slow

Hierarchical Clustering



Hierarchical Clustering

Two categories

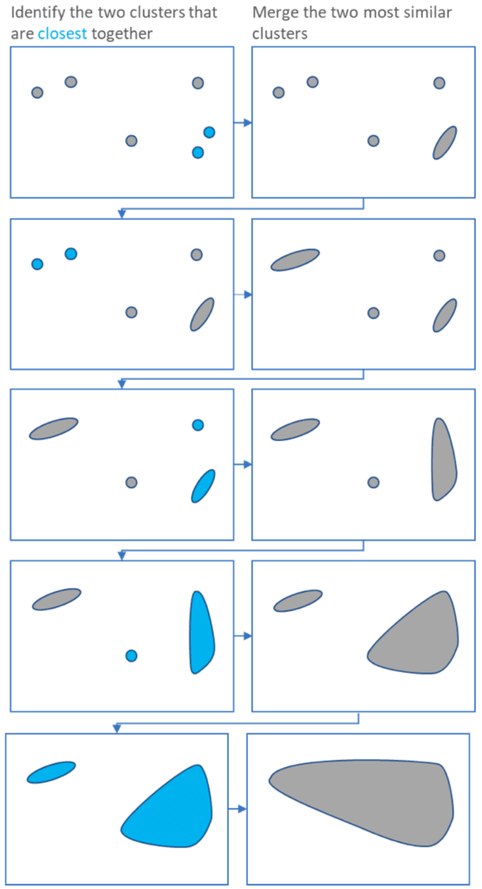
- bottom-up
  - AGglomerative NESting - AGNES
- top-down
  - DIvisive ANAlysis – DIANA

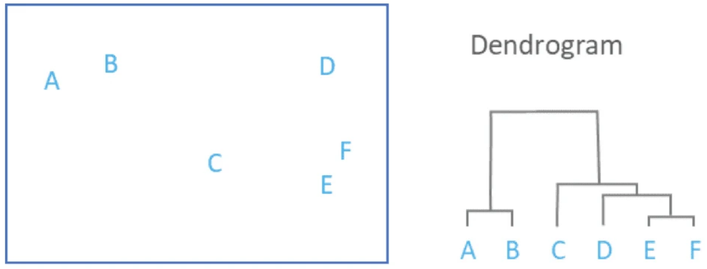
AGNES

1. Starts with the assumption that each data point as a single cluster – in total N clusters
2. Measures the distance between clusters and then merges (agglomerates) the closest pairs into one
3. Repeat step 2 until all points get into one cluster

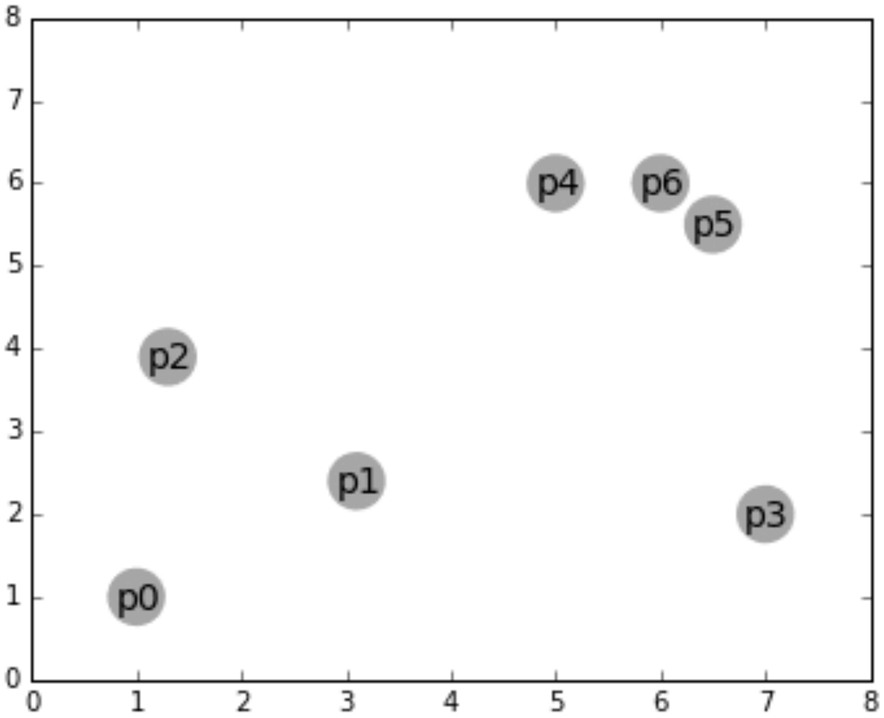
Dendrogram – the hierarchy of clusters, built in the process

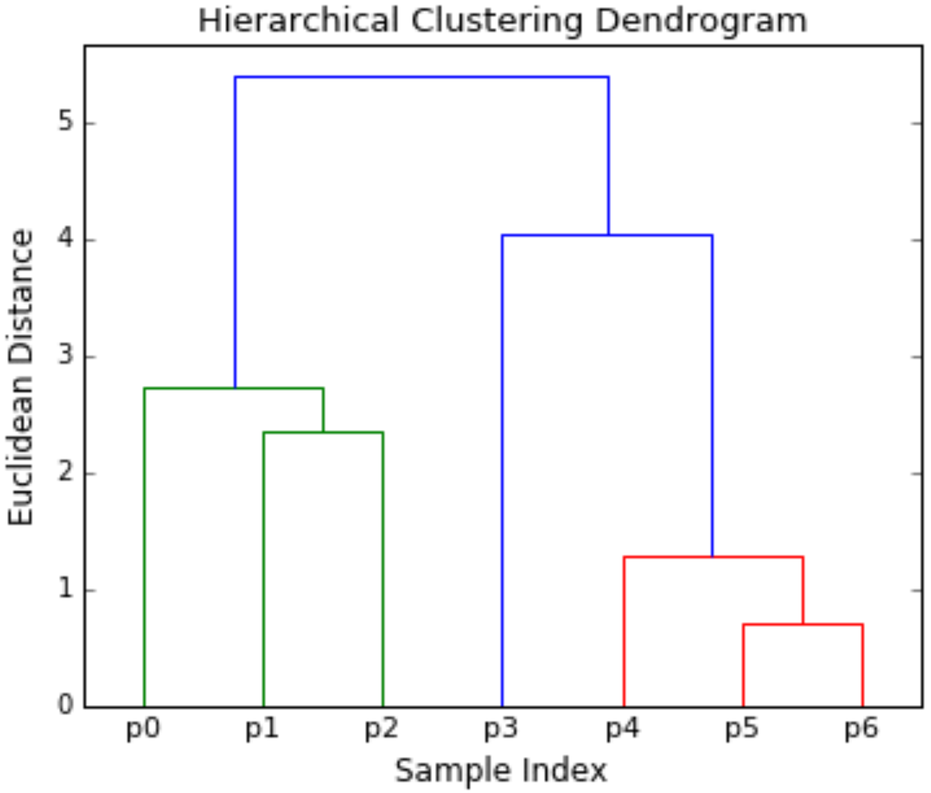
- root: a cluster with all points
- leaves: clusters with a single point each
- any horizontal line in the dendrogram shows a number of clusters and the value of clustering feature



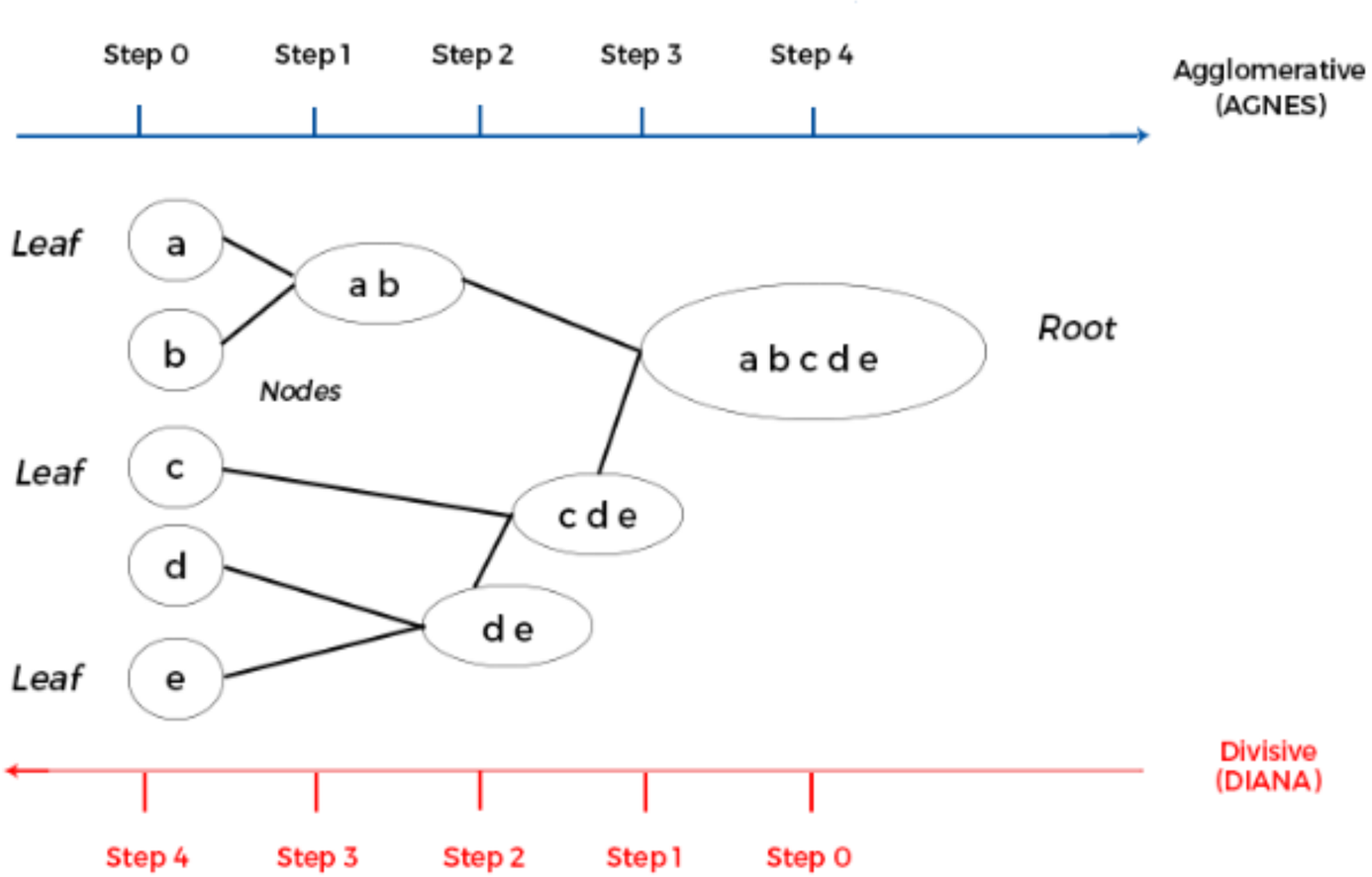


Cutting the Tree into Clusters





AGNES vs DIANA



Advantages and Disadvantages

Advantages

- hierarchical clustering does not require pre-defined number of clusters - we can select the optimal number of clusters afterwards, from the dendrogram
- the algorithm doesn't depend on distance metrics

Best for

- when the input data has a hierarchical structure, the algorithm recovers it

Disadvantage

- lower efficiency
- slow for big datasets

Reference

- Prateek Joshi, Artificial Intelligence with Python
- Russell & Norvig, AI: A Modern Approach, 3rd Ed.
- http://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
- https://www.educba.com/course/cluster-analysis-course/
- http://geek.sg/step-by-step-pokemon-clustering/
- http://www.onmyphd.com/?p=k-means.clustering
- http://www.learnbymarketing.com/methods/k-means-clustering/
- https://jakevdp.github.io/PythonDataScienceHandbook/05.11-k-means.html
- https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68
- https://dzone.com/articles/10-interesting-use-cases-for-the-k-means-algorithm
- https://nikkimarinsek.com/blog/7-ways-to-label-a-cluster-plot-python
- https://www.youtube.com/watch?v=QXOkPvFM6NU
- https://www.youtube.com/watch?v=9991JlKnFmk
- https://www.youtube.com/watch?v=4b5d3muPQmA