Statistical Databases and Registers with some datamining

a course in
Survey Methodology and Official Statistics
Pages in the book: 501-528

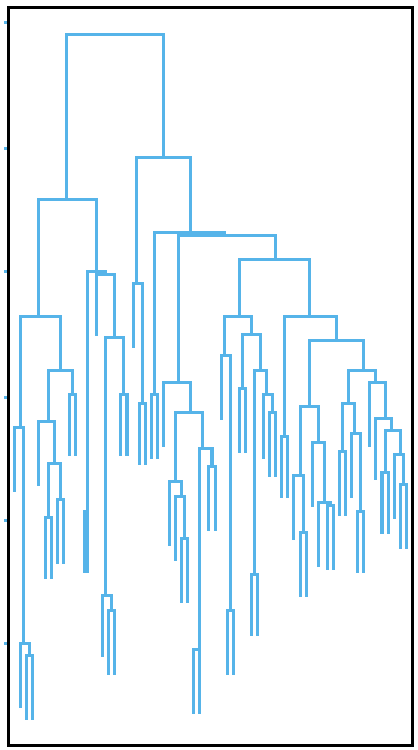
Department of Statistics
Stockholm University

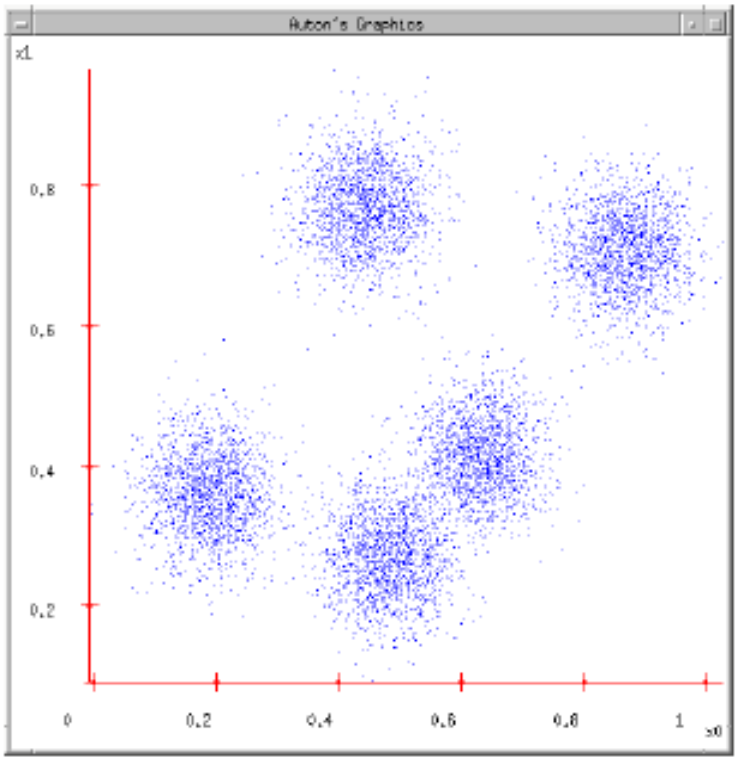
December 2011

Cluster analysis

Goals for cluster analysis are to

- arrange into a natural hierarchy
- group with respect to similarities





When is cluster analysis useful

- You need to identify customers with similar patterns of past purchases so that you can tailor your marketing strategies.
- You've been assigned to group television shows into homogeneous categories based on viewer characteristics. This can be used for market segmentation.
- You can use it in biology, to derive plant and animal taxonomies, to categorize genes with similar functionality, and to gain insight into structures inherent in populations.
- You're trying to examine patients with a diagnosis of depression to determine if distinct subgroups can be identified, based on a symptom checklist.

Tools for cluster analysis

Definition (Data matrix)

Set

$$x_{ij} = \text{measurement object } i \text{ variable } j$$

where $i = 1, 2, \ldots, n, j = 1, 2, \ldots, p$. We then have the $n \times p$ data matrix

$$\mathbb{X} = \begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix}$$

Think of j as product j and i as customer i. The x_{ij} may be amount sold or value of amount sold.

Definition (Dissimilarity matrix)

Set

$$\mathbb{D} = \begin{bmatrix} 0 & \cdots & \cdots & \cdots & \cdots \\ d(2,1) & 0 & \cdots & \cdots & \cdots \\ d(3,1) & d(3,2) & 0 & \cdots & \cdots \\ \cdots & \cdots & \cdots & \ddots & \cdots \\ d(n,1) & d(n,2) & \cdots & d(n,n-1) & 0 \end{bmatrix}$$

where $d(i,j)$ is the measured **difference** or **dissimilarity** between object i and j, $i,j = 1,2,\ldots,n$.

Here $d(i,j)$ is a distance function ie how far $\mathbf{x}_i$ is from $\mathbf{x}_j$

1. $d(i,j) \geq 0$
2. $d(i,i) = 0$
3. $\mathrm{d}(i,j) = d(j,i)$
4. $d(i,j) \leq d(i,k) + d(k,j)$

K-means

The algorithm k-means is a natural and easy method to find clusters, when you have the number of clusters.

It is based on Euclidean distance and the following observation

$$
\begin{aligned}
T &= \sum_{j=1}^{p}\sum_{i=1}^{n}\sum_{i'=1}^{n}\left(x_{ij}-x_{i'j}\right)^2 = \sum_{j=1}^{p}\sum_{i=1}^{n}\sum_{i'=1}^{n}\left(x_{ij}-\bar{x}_j+\bar{x}_j-x_{i'j}\right)^2 \\
&= \sum_{j}\sum_{i}\sum_{i'}\left(\left(x_{ij}-\bar{x}_j\right)^2+2\left(x_{ij}-\bar{x}_j\right)\left(x_{i'j}-\bar{x}_j\right)+\left(x_{i'j}-\bar{x}_j\right)^2\right) \\
&= \sum_{j}\left(\sum_{i}\sum_{i'}\left(x_{ij}-\bar{x}_j\right)^2+\sum_{i}\sum_{i'}\left(x_{i'j}-\bar{x}_j\right)^2\right)
\end{aligned}
$$

K-means (forts)

From this follows that

$$
\begin{aligned}
T &= \sum_{j=1}^{p} \left(n \underbrace{\sum_{i} (x_{ij} - \bar{x}_j)^2}_{\text{within cluster}} + n \sum_{i'} \left(x_{i'j} - \bar{x}_j \right)^2 \right) \\
&= 2n^2 \times \text{within cluster variance}
\end{aligned}
$$

The algorithm k-means minimizes the value T by finding its minimum under the constraint p clusters. Usually it finds a local minimum, which may or may not be the global minimum.

K-means (forts)

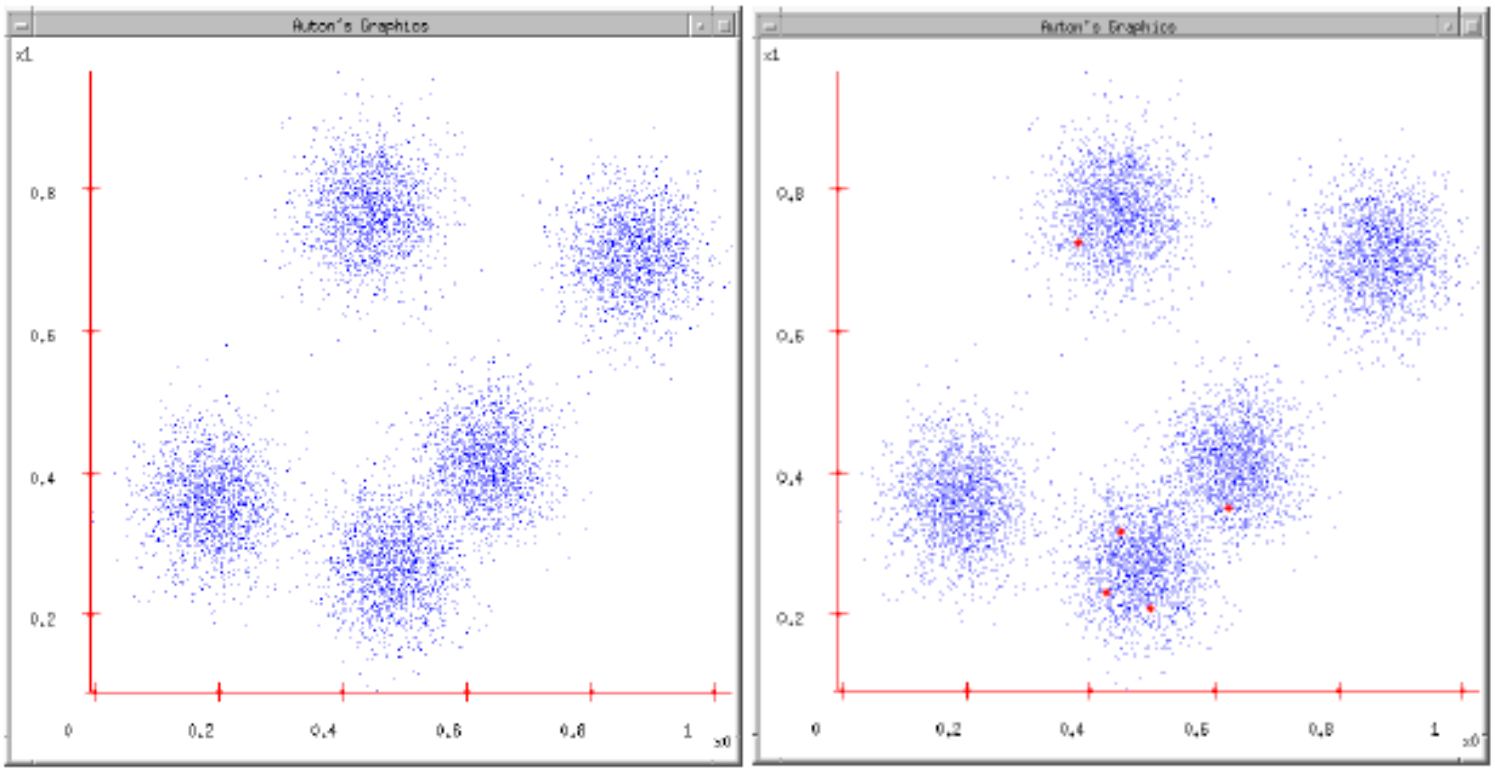
To find a minimum value for T the method k-means uses the following algorithm:

ALGORITHM

1. Choose points $m_1, \ldots, m_p$ and call them centroids
2. Partition the population into p subsets so that each point x_{ij} is attached to the m_k which is closest, Euclidean distance.
3. For subset k, $k = 1, 2, \ldots, p$, compute the center of gravity and set m_k equal to this center.
4. If at least one m_k changes then go to 2, otherwise stop.

Data generation

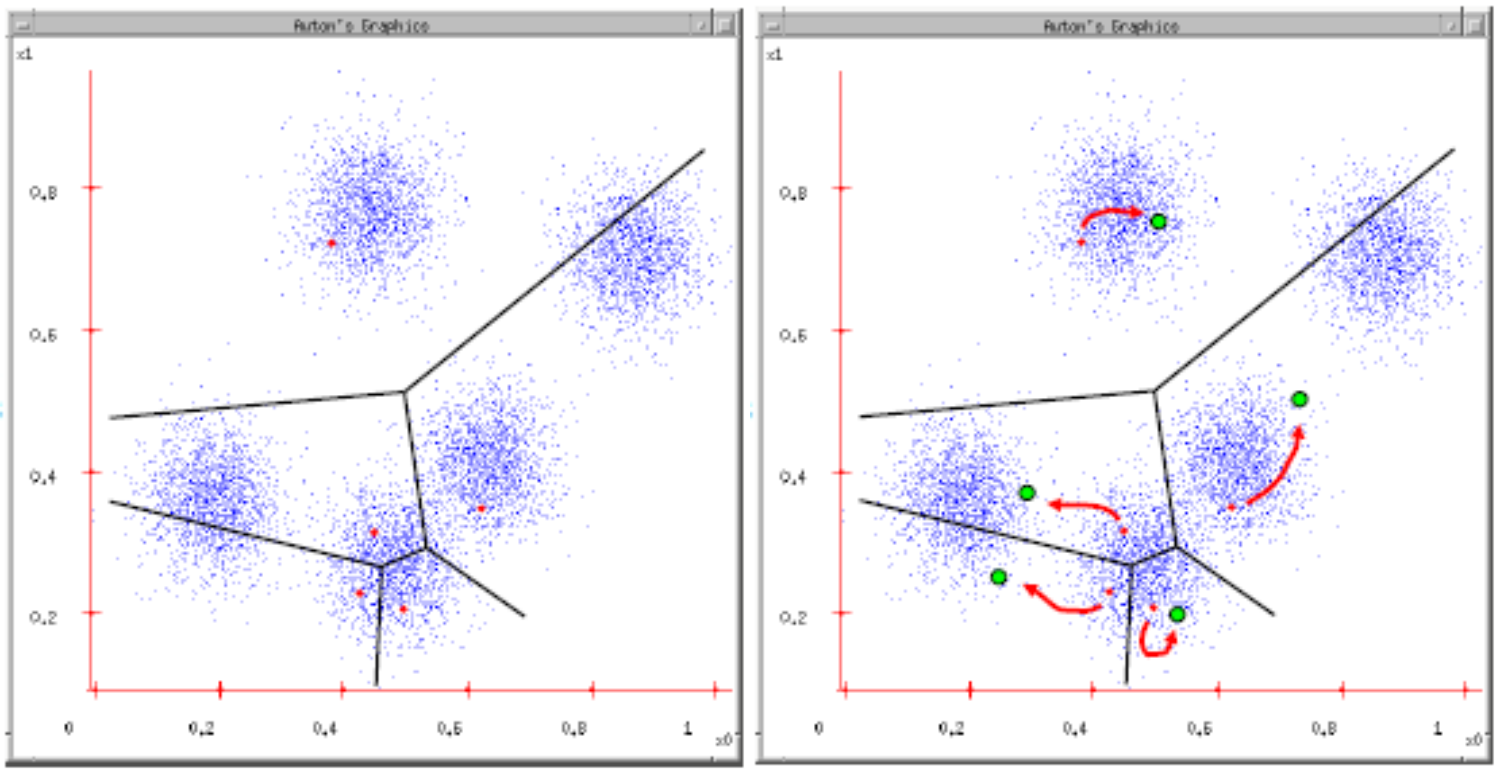
We generate a dataset from 5 different bi-normal distributions with 8000 points in it.



We will run 5-means on it and hence generate 5 starting points (the red ones).

Example: K-means (cont)

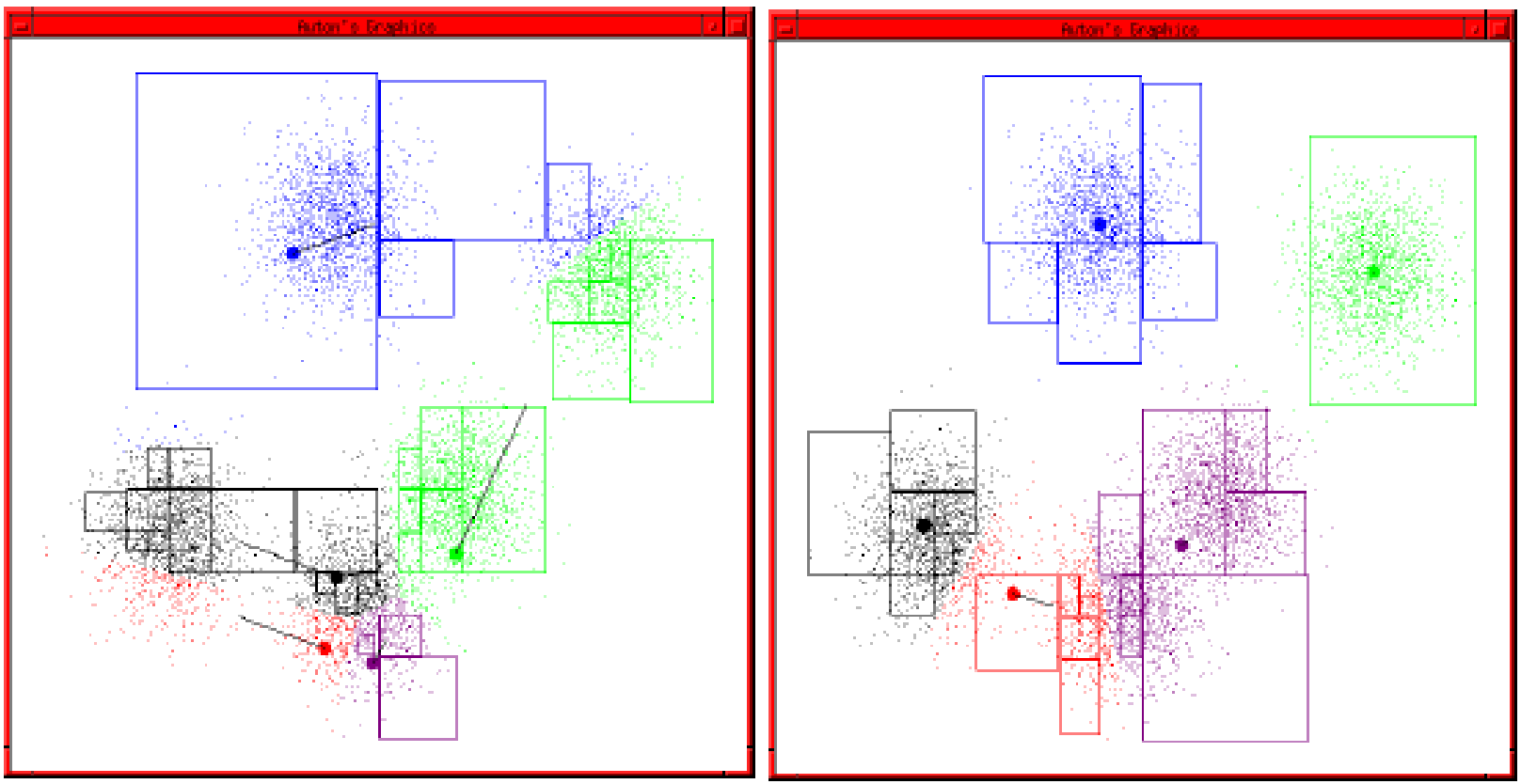
We associate each datapoint with the class center closest to it.



The first picture shows the induced partition, by the centroids, of subsets. New centroids in green.

Example: K-means (cont)

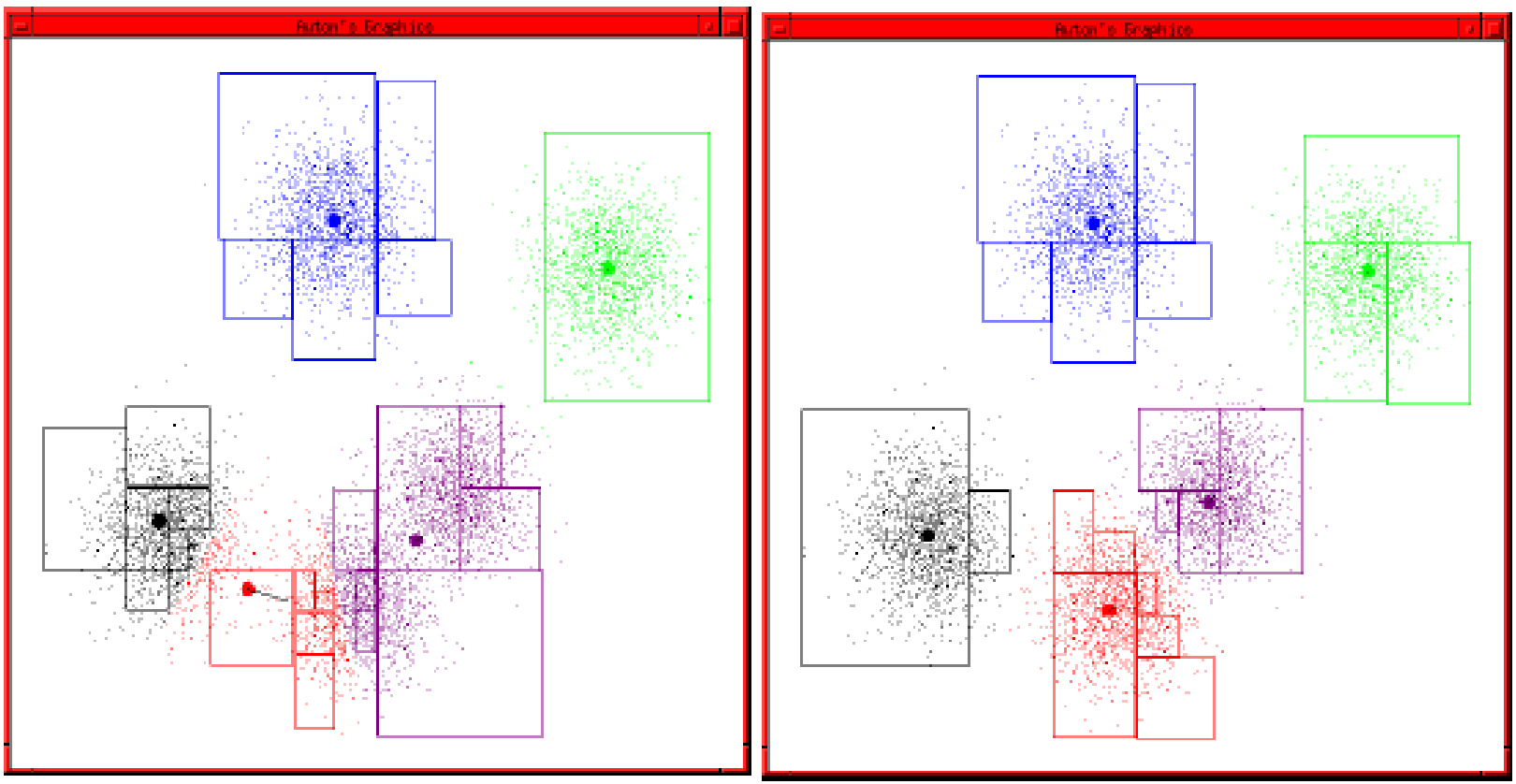
We indicate the direction that the points move by an arrow.



In the first picture black and red moves to the left, green and blue up.

Example: K-means (stop)

Left picture another iteration



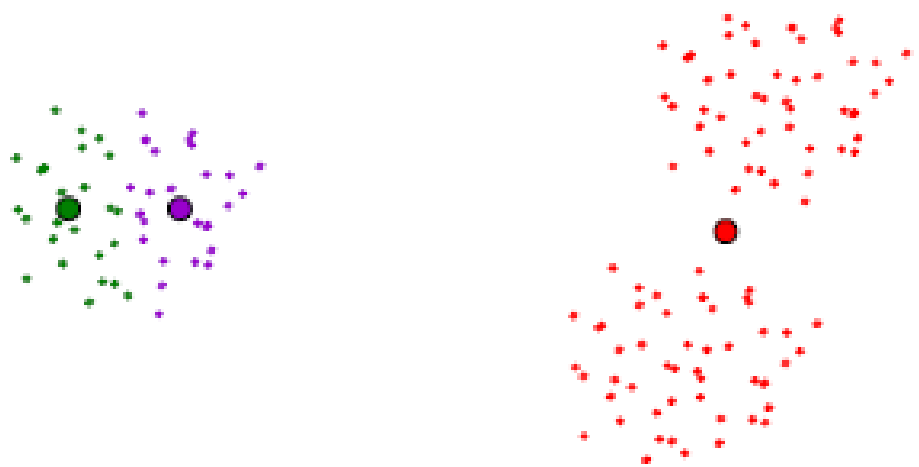
The right picture shows the final result.

Example: K-means (cont)

In our example we had 5 clusters and five centroids and we found the five best cluster.

It is not always true that we will find the best partition even though we have the proper amount of centroids.

Contemplate the following figure

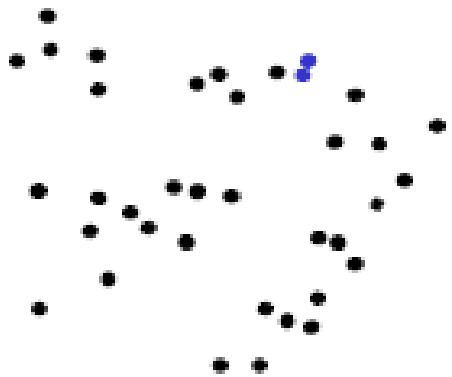


Will the given centroids ever move?

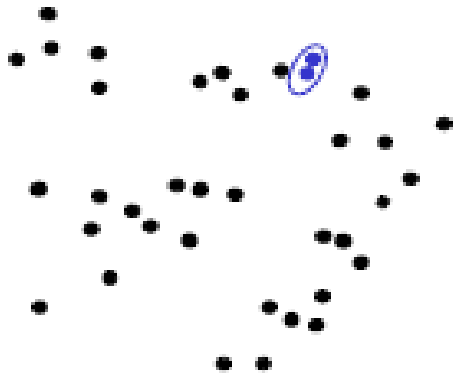
Refined k-means

There is a new K-means algorithm that do not demand initial centroids and still find the "best" clusters.
See autonlab.com and the animation **Kmeans**.

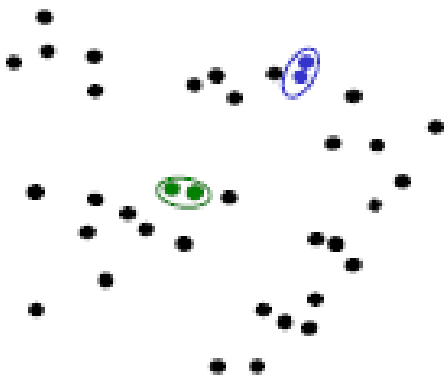
Example: Hierarchical clustering



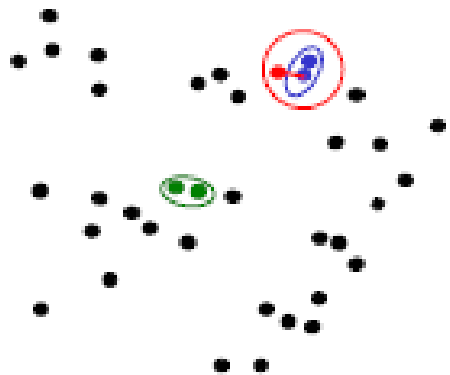
Step 1)



Step 2)



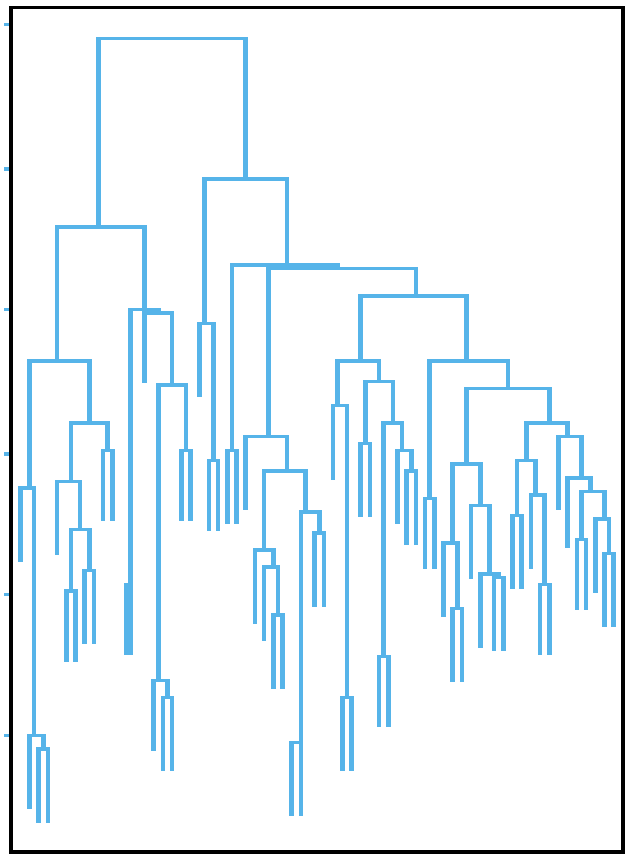
Step 3)



Step 4)

Example: Hierarchical clustering

Such clustering technique give rise to dendrograms



Useful e.g. when making a taxonomy of animals, plants and so on.

Interesting sites

Auton Lab
Neural Networks and Applications
Pattern recognition – Universität Heidelberg
The mathematical monk (observe the playlist at the bottom of the screen)