Soft Clustering

Till Helge Helwig

Eberhard-Karls-University Tübingen

Lecture "Data Mining in Bioinformatics"

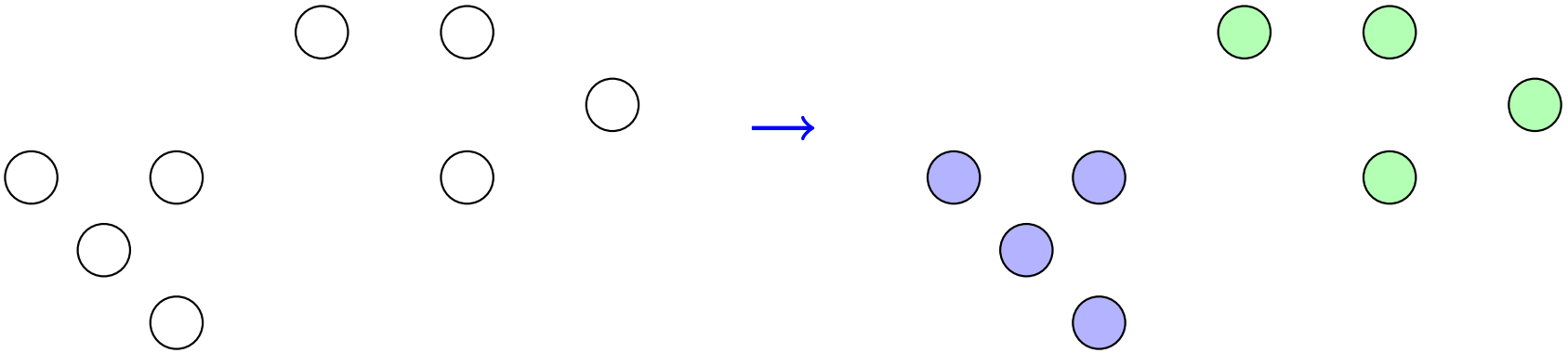
March 12th, 2010

Overview

Overview

Clustering

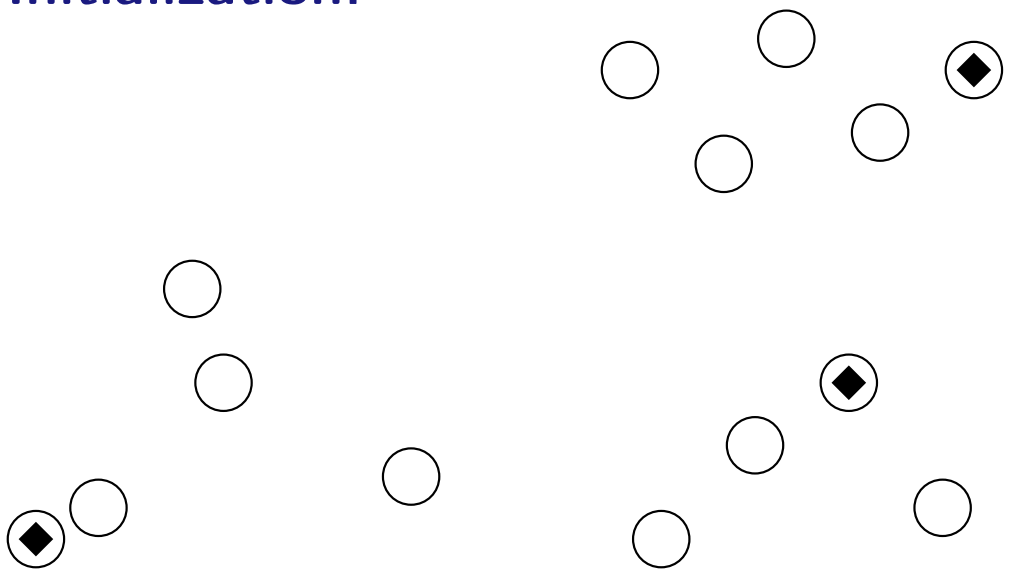
- Discovery of **classes** in a set of objects
- **Unsupervised** learning



K-Means Clustering in a Nutshell

- Build k **clusters**
- Minimize **intra-cluster variance**

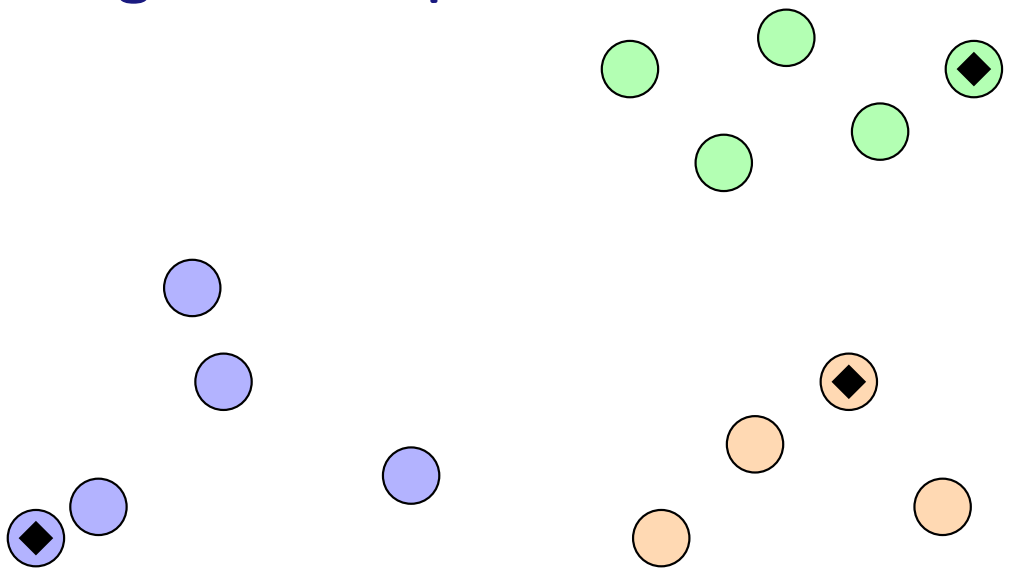
1. **Initialization:**



Pick k **random points** as means

K-Means Clustering in a Nutshell (2)

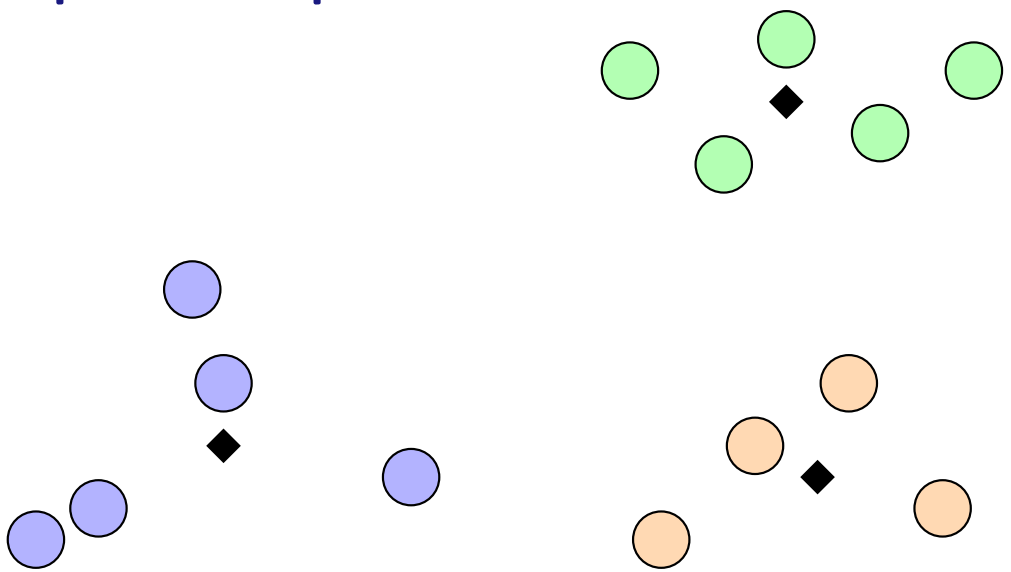
2. **Assignment step:**



Assign each point to the **nearest** mean

K-Means Clustering in a Nutshell (3)

3. **Update step:**



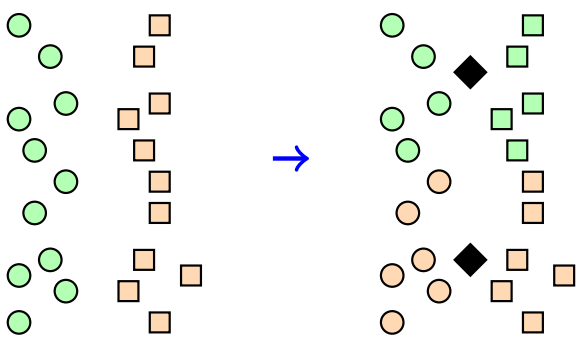
Recalculate means from corresponding points

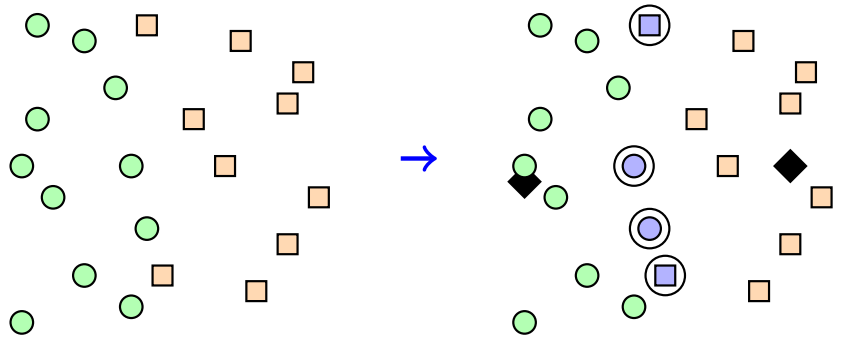
4. Go to step **2.** if at least for one point the cluster was changed

Overview

Difficult Cases

- Sometimes classes **can not** be defined by the least distance to a central point
- **Examples:**





- **Elongated clusters** often are not detected correctly
- **Points on the border** between two means should not be assigned to one cluster

Taking Care of Uncertainty

- Points can belong to **more than one** cluster
- Clustering should reflect the **degree of association** between points and means
- Replace hard (absolute) decisions in algorithms with **soft (relative) ones**
- The final clustering allows **interpretation** of uncertain points

Soft K-Means Clustering

- New representation of associations as **responsibility matrix**:

	m_1	$\dots$	m_k
p_1	$r_1^{(1)}$	$\dots$	$r_k^{(1)}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$
p_n	$r_1^{(n)}$	$\dots$	$r_k^{(n)}$

- $r_c^{(i)}$ describes the **responsibility** of cluster c for point i:

$$r_c^{(p)} = \frac{\exp\left(-\beta d\left(m_c, p_i\right)\right)}{\sum_{k'} \exp\left(-\beta d\left(m_{k'}, p_i\right)\right)}$$

$\Rightarrow$ **New parameter** β, which describes the "**stiffness**" of the clustering

Soft K-Means Clustering (2)

- **New update step:**
  Refine all clusters c via:

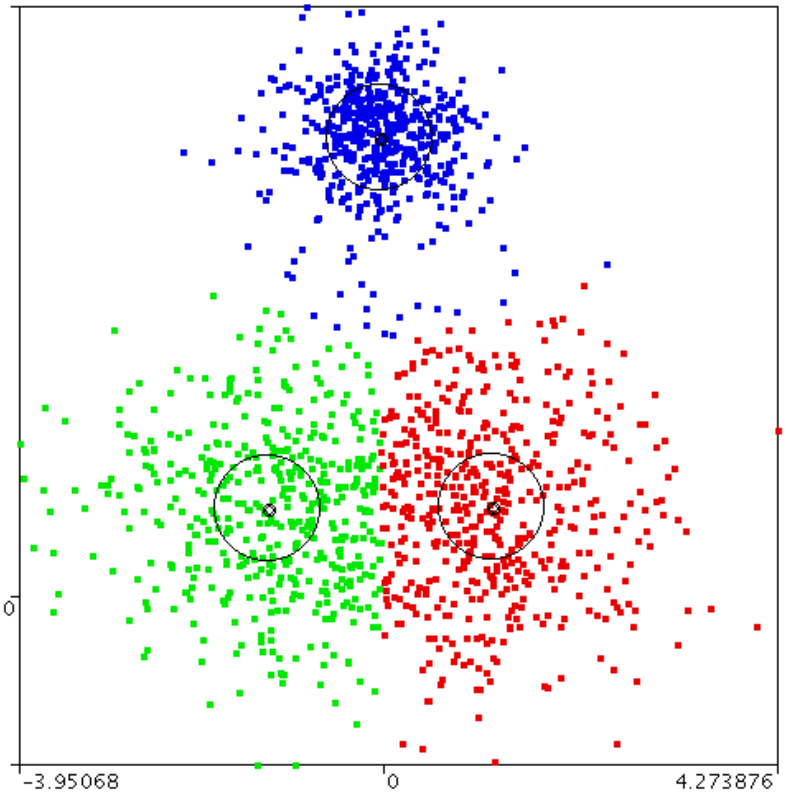
$$m_c = \frac{\sum_{i=1}^{n} r_c^{(i)} p_i}{\sum_{i=1}^{n} r_c^{(i)}} = \frac{\text{Weighted points sum}}{\text{Total responsibility}}$$

- **New association step:**
  Update the responsibility matrix

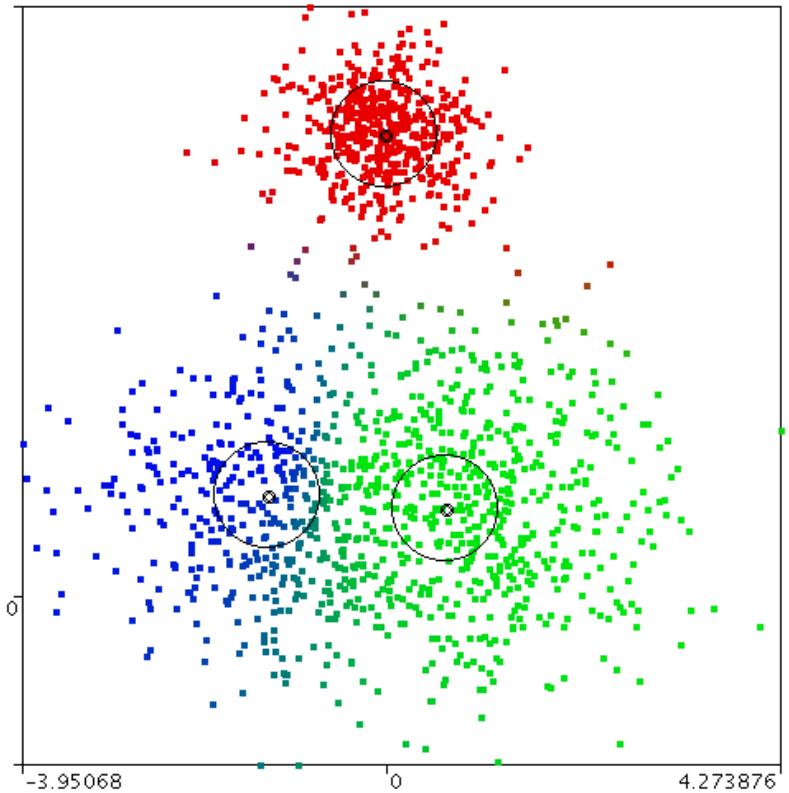
Stiffness

- The stiffness β influences the **difference** to the hard k-means clustering
- Soft k-means clustering with $\beta \to \infty$ would yield the **same result** as hard k-means clustering
- Figuring out the right value for β is **non-trivial** even with "try and error"

Example



K-means clustering



Soft k-means clustering ($\beta = 4$)

Applications

- In general soft clustering can reduce **information loss** due to discarding all clusters except one
- **Document clustering** e.g. for web search engines:
  - $\Rightarrow$ Soft clustering allows for documents to occur in **several topics**
- Analysis of **gene expression data** (microarray experiments):
  - $\Rightarrow$ Soft clustering decreases the **sensitivity towards noise**
- Prediction of molecule functions from **protein-protein-interaction networks**
  - $\Rightarrow$ Soft clustering can assign **several relevant functions** to a protein

Overview

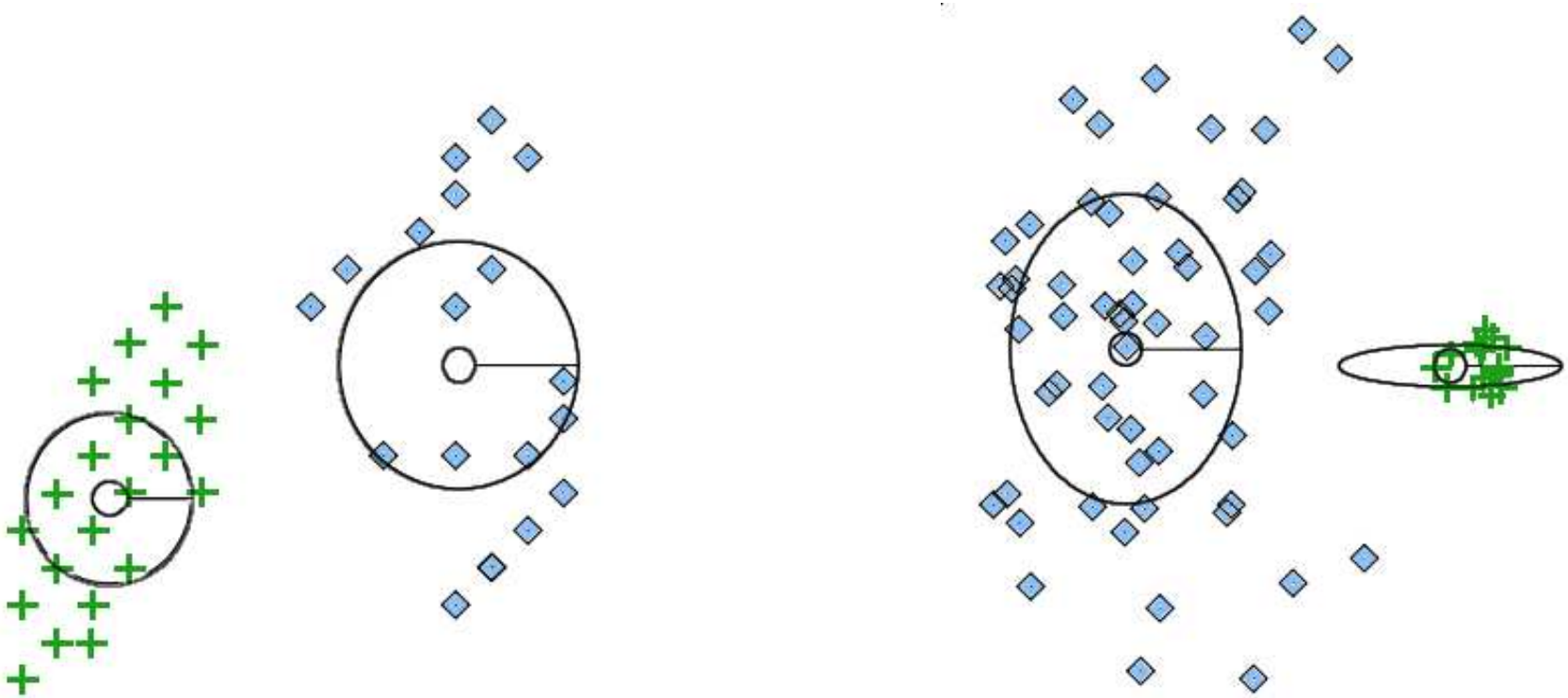
Further Improvements of Soft K-Means Clustering

- Choice of β decides about **usefulness** of the results
  - Modification using **Gaussian maximum likelihood** function
  - Assumption: Each cluster is **Gaussian sphere** with its own width
  - During the update step the algorithm **recalculates** β itself

  $\Rightarrow$ Clusters with **different sizes** can be detected
- Similar enhancement using **axis-aligned Gaussians** is possible

  $\Rightarrow$ Clusters with elongated shapes can be detected

Examples



Source: MacKay, 2003

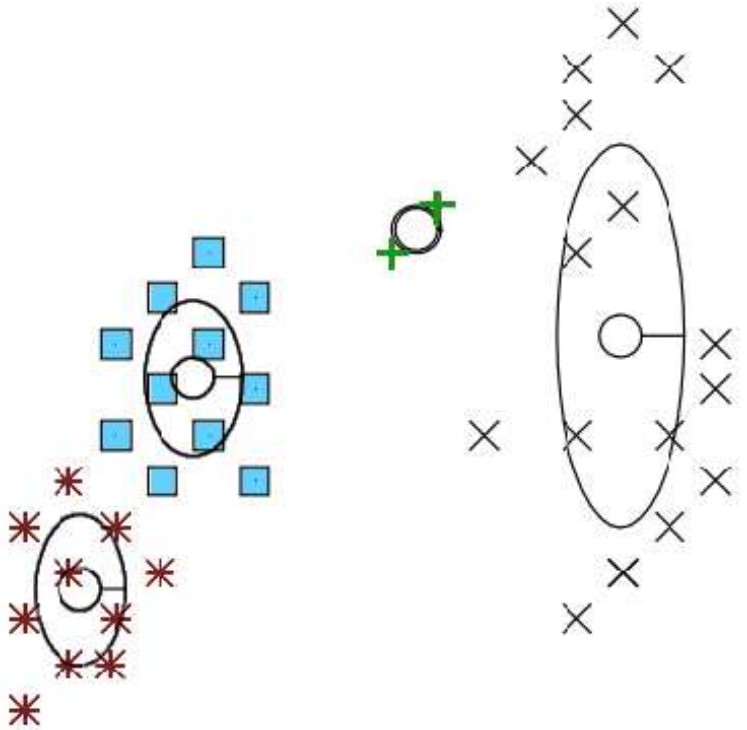
Overview

The Last Slide

- **Soft decisions** can improve the chances for finding **clusters with special shapes**
- This can go terribly wrong and make the **results worse**
- **Association** of points to clusters is the crucial step in k-means clustering

⇒ Many more **improvements** and **modifications** possible

- **Applications** for soft clustering are innumerable



Source: MacKay, 2003

Thank you for your attention.

Questions? Remarks?

References:

K. Borgwardt.
Data mining in bioinformatics.
Lecture, March 2010.

D.J.C. MacKay.
Information theory, inference, and learning algorithms.
Cambridge Univ Press, 2003.