



Big Data Analytics

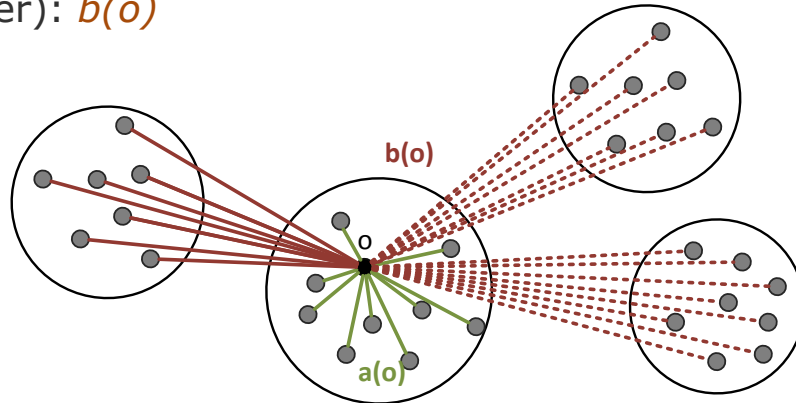
– Week 2: Clustering Evaluation & Expectation Maximization –

Knowledge Discovery and Data Mining
Prof. Dr. Emmanuel Müller

The Silhouette coefficient (1)

Basic idea:

- How good is the clustering = how appropriate is the mapping of objects to clusters
- Elements in cluster should be “similar” to their representative
→ measure the average distance of objects to their representative: $a(o)$
- Elements in different clusters should be “dissimilar”
→ measure the average distance of objects to alternative clusters
(i.e. second closest cluster): $b(o)$



The Silhouette coefficient (2)

$a(o)$: average distance between object o and the objects in its cluster A

$$a(o) = \frac{1}{|C(o)|} \sum_{p \in C(o)} \text{dist}(o, p)$$

$b(o)$: average distance between object o and the objects in its "second closest" cluster B

$$b(o) = \min_{c_i \neq C(o)} \left(\frac{1}{|C_i|} \sum_{p \in C_i} \text{dist}(o, p) \right)$$

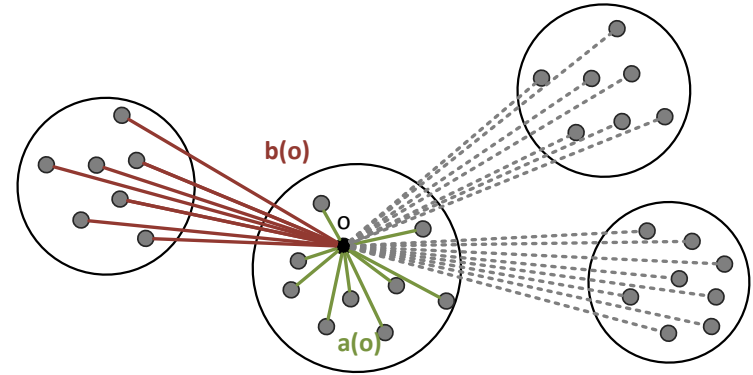
The silhouette of o is then defined as

$$s(o) = \begin{cases} 0 & \text{if } a(o) = 0, \text{ i.e. } |C_i| = 1 \\ \frac{b(o) - a(o)}{\max\{a(o), b(o)\}} & \text{else} \end{cases}$$

The silhouette of a clustering $\mathcal{C} = (C_1, \dots, C_k)$ is defined as:

$$\text{silh}(\mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum_{C_i \in \mathcal{C}} \frac{1}{|C_i|} \sum_{o \in C_i} s(o)$$

The values of the silhouette coefficient range from -1 to $+1$



The Silhouette coefficient (3)

„Reading“ the silhouette coefficient
how good is the assignment of o to its cluster

- $s(o) = -1$: bad, on average closer to members of B
- $s(o) = 0$: in-between A and B
- $s(o) = 1$: good assignment of o to its cluster A

Silhouette Coefficient s_C of a clustering: average silhouette of all objects

- $0.7 < s_C \leq 1.0$ strong structure
- $0.5 < s_C \leq 0.7$ medium structure
- $0.25 < s_C \leq 0.5$ weak structure
- $s_C \leq 0.25$ no structure

Expectation Maximization (EM)

Additional Literature: C. M. Bishop „Pattern Recognition and Machine Learning“
Springer, 2009 (Chapter 9)

Statistical approach for finding
maximum likelihood estimates of parameters
in probabilistic models

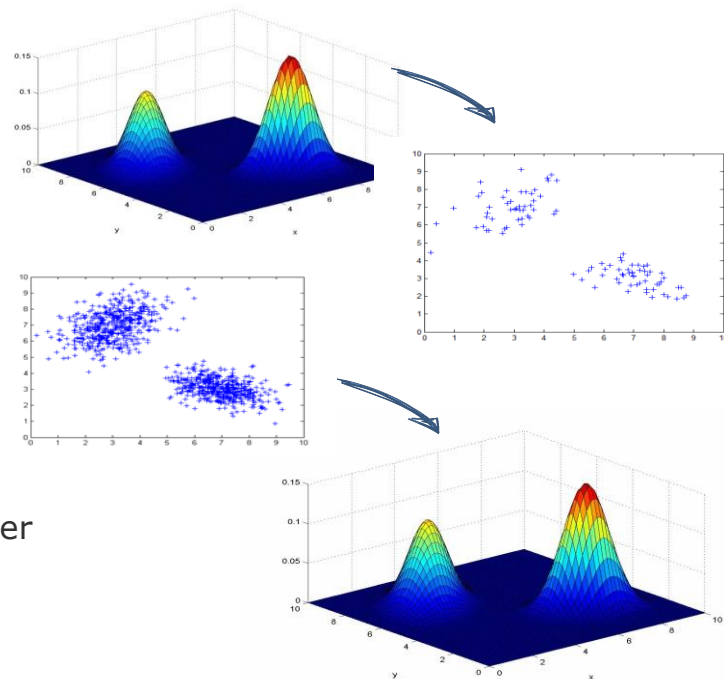
Here: using EM as clustering algorithm

Underlying assumption:

- Observations are drawn from one of several components of a mixture distribution.

Main idea:

- Define clusters as probability distributions
→ each object has a certain probability of belonging to each cluster
- Iteratively improve the parameters of each distribution
(e.g. center, “width” and “height” of a Gaussian distribution)
until some quality threshold is reached



Excursus: Gaussian Mixture Distributions

Note: EM is not restricted to Gaussian distributions, but they will serve as example in this course

Gaussian distribution:

- Univariate: single variable $x \in \mathbb{R}$:

$$p(x|\mu, \sigma^2) = \mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2\sigma^2} \cdot (x-\mu)^2}$$

mean $\in \mathbb{R}$

variance $\in \mathbb{R}$

- Multivariate: D -dimensional vector $\mathbf{x} \in \mathbb{R}^D$:

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \cdot e^{-\frac{1}{2} \cdot (\mathbf{x}-\boldsymbol{\mu})^T \cdot (\boldsymbol{\Sigma})^{-1} \cdot (\mathbf{x}-\boldsymbol{\mu})}$$

mean vector $\in \mathbb{R}^D$

covariance matrix $\in \mathbb{R}^{D \times D}$

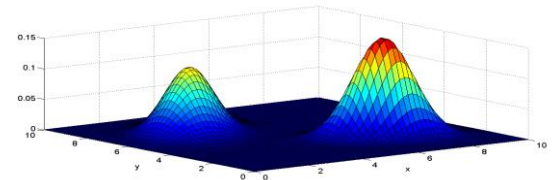
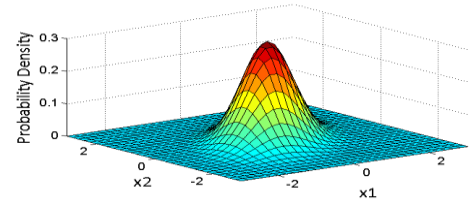
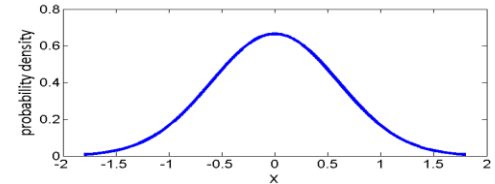
Gaussian mixture distribution with K components:

- D -dimensional vector $\mathbf{x} \in \mathbb{R}^D$:

$$p(\mathbf{x}) = \sum_{k=1}^K \pi_k \cdot \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$



mixing coefficients $\in \mathbb{R}$: $\sum_k \pi_k = 1$ and $0 \leq \pi_k \leq 1$



Expectation Maximization (EM)

A clustering $\mathcal{M} = \{C_1, \dots, C_K\}$ is represented by a mixture distribution with parameters $\Theta =$

$\{\pi_1, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \dots, \pi_K, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K\} :$

$$p(\mathbf{x}|\Theta) = \sum_{k=1}^K \pi_k \cdot \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Each cluster is represented by one component of the mixture distribution:

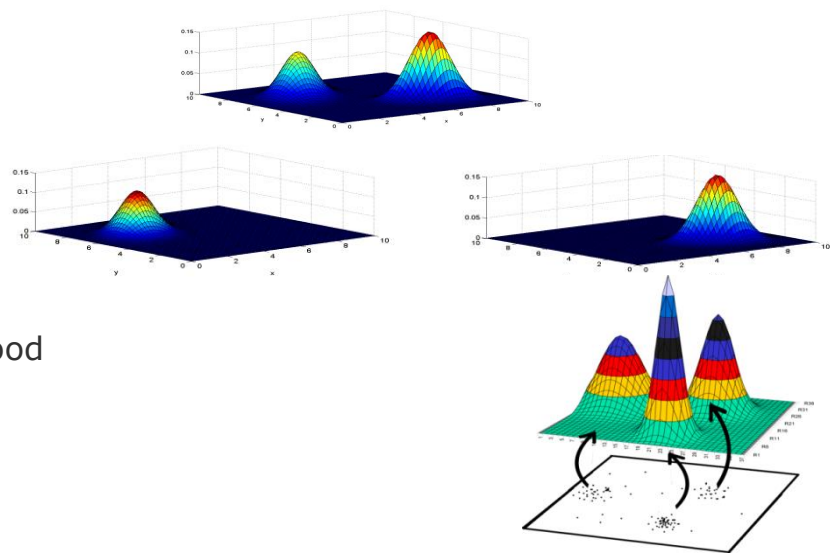
$$p(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Given a dataset $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\} \subseteq \mathbb{R}^D$, we can write the likelihood that all data points $\mathbf{x}_n \in \mathbf{X}$ are generated (independently) by the mixture model with parameters Θ as:

$$p(\mathbf{X}|\Theta) = \prod_{n=1}^N \sum_{k=1}^K \pi_k \cdot p(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

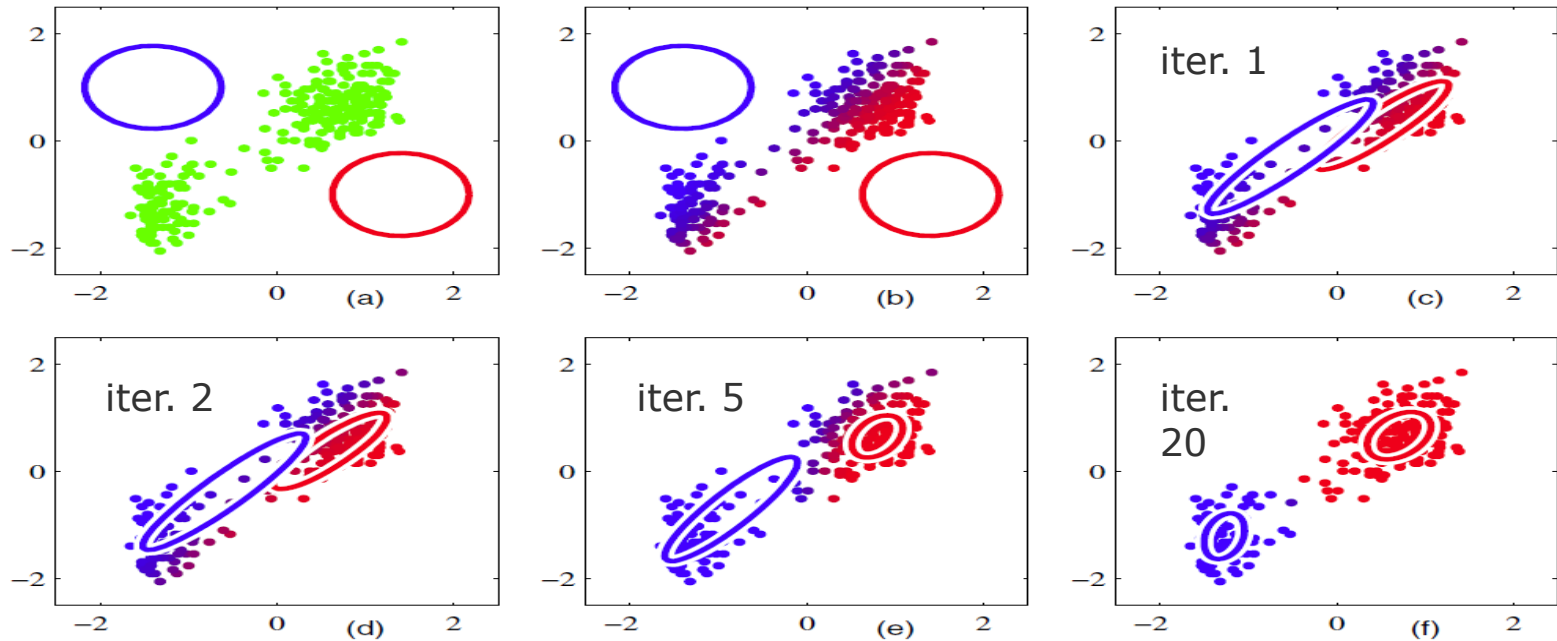
Goal: Find the parameters Θ_{ML} with maximal (log-)likelihood estimation (MLE)

$$\Theta_{ML} = \arg \max_{\Theta} \{\log p(\mathbf{X}|\Theta)\}$$



Expectation Maximization (EM): Exemplary Application

Example taken from: C. M. Bishop „Pattern Recognition and Machine Learning“, 2009



Discussion

Superior to k-Means for clusters of varying size or clusters having differing variances

→ more accurate data representation

Convergence to (possibly local) maximum

Computational effort for N objects, K derived clusters, and t iterations (t is quite high in many cases):

$$O(t \cdot N \cdot K \cdot D^2)$$

Both - result and runtime - strongly depend on

- the initial assignment

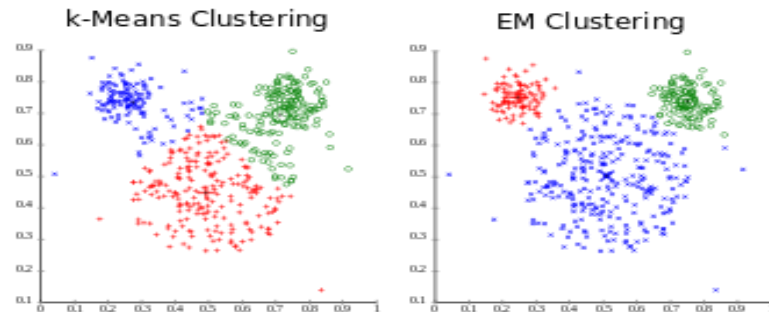
- do multiple random starts and choose the final estimate with highest likelihood

- Initialize with clustering algorithms (e.g., K-Means usually converges much faster)

- Local maxima and initialization issues have been addressed in various extensions of EM

- a proper choice of parameter K (= desired number of clusters)

- Apply principals of model selection (see next slide)



Model Selection for Determining Parameter K

[MP'00] G. McLachlan and D. Peel. *Finite Mixture Models*. Wiley, New York, 2000.

Classical trade-off problem for selecting the proper number of components K

- If K is too high, the mixture may overfit the data
- If K is too low, the mixture may not be flexible enough to approximate the data

Idea: determine candidate models θ_K for a range of values of K (from K_{min} to K_{max}) and select the model

$$\theta_{K^*} = \max\{\text{qual}(\theta_K) | K \in \{K_{min}, \dots, K_{max}\}\}$$

- Silhouette Coefficient (as for k -Means) only works for partitioning approaches.
- The MLE (Maximum Likelihood Estimation) criterion is non-decreasing in K

Solution: deterministic or stochastic model selection methods[MP'00]
which try to balance the goodness of fit with simplicity.

- Deterministic: $\text{qual}(\theta_K) = \log p(\mathbf{X} | \theta_K) + \mathcal{P}(K)$
where $\mathcal{P}(K)$ is an increasing function penalizing higher values of K
- Stochastic: based on Markov Chain Monte Carlo (MCMC)