Outline

- Hierarchical clustering
- Density-based clustering
- Hierarchical Density-based Clustering
- Model-based Clustering (EM-Clustering)

Model-based clustering

- Assumption: data have been generated by a statistical process
- Goal: find the statistical model that best fits the data
  - Statistical model = distribution (e.g., Gaussian) and its parameters
- Mixture models: model the data using a number of distributions
  - Each distribution corresponds to a cluster, the parameters of the distribution comprise the cluster description
- Procedure:
  - i) decide on the model
  - ii) find the parameters of the model from the data
- A particular kind of statistical model: Gaussian mixture models

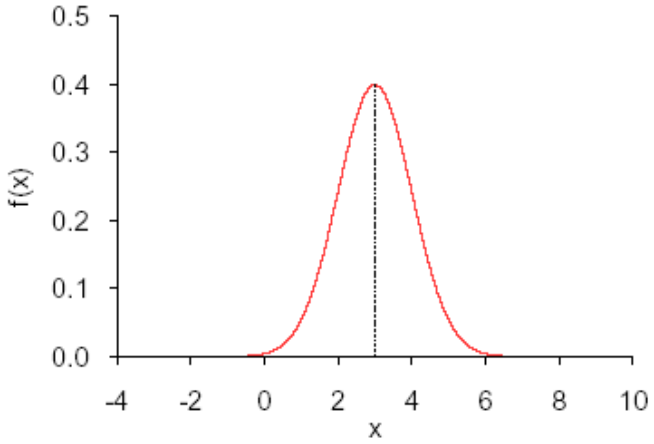
Gaussian Mixture Models

- Let X be a dataset of *N* d-dimensional objects
  - Objects are *d*-dimensional vectors x = (x_1, ..., x_d)
- Let's assume that objects are generated by a mixture of *k* (normal) distributions.
  - The dataset objects are independent and identically distributed samples from the mixture of the *k* distributions
- Each cluster refers to one multivariate Gaussian distribution
- Each cluster is represented by
  - Mean (centroid) μ_c
  - *d* x *d* covariance matrix Σ_c for the points in cluster *c*
- Probability density function of a Gaussian distribution

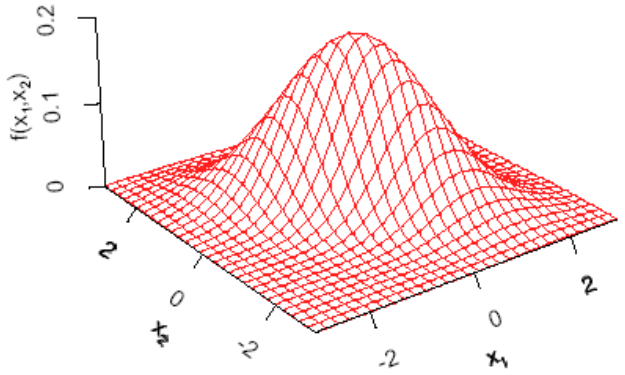
$$P(x \mid c) = \frac{1}{\sqrt{(2\pi)^d \mid \Sigma_c \mid}} e^{-\frac{1}{2}\cdot(x-\mu_c)^T \cdot \Sigma_c^{-1} \cdot (x-\mu_c)}$$

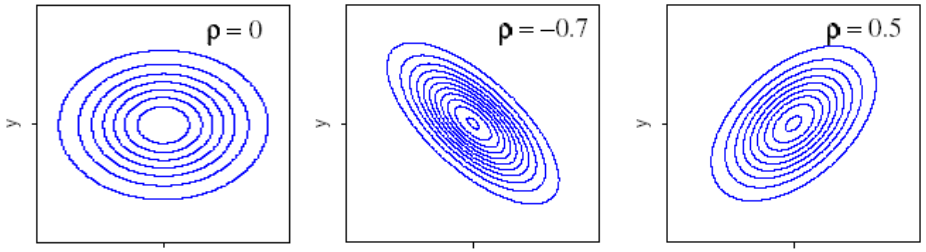
Multivariate normal distribution

Univariate normal distribution



Bivariate normal distribution





No covariance

Negative covariance

Positive covariance

Gaussian Mixture Models

- Probability of a cluster c

$$\mathrm{P}(c_l) = \frac{1}{N} \sum_{i=1}^{N} \mathrm{P}(c_l|x_i)$$

- Probability of observing an object x_i

$$\mathrm{P}(x_i) = \sum_{l=1}^{k} \mathrm{P}(c_l)\mathrm{P}(x_i|c_l)$$

  - where $P(x_i|c_l)$ is given by the probability density function of the Gaussian distribution

Gaussian Mixture Models

- If the objects are generated in an independent manner, the probability of the whole set of objects X, |X|=N, is just the product of the probabilities of each x_i in X:

$$\mathcal{L} = \prod_{i=1}^{N} \mathrm{P}(x_i)$$

$$= \prod_{i=1}^{N} \sum_{l=1}^{k} \mathrm{P}(c_l)\mathrm{P}(x_i|c_l)$$

- We want to find the distribution parameters that maximize the likelihood that the observed objects follow the model.
- Using statistical methods, we can estimate the parameters of these distributions from the data, and thus describe the clusters.

Gaussian Mixture Models clustering

- How can we partition the data?
  - Choose the most likely cluster assignment for each object

- How to estimate the statistics of each cluster efficiently?
  - Use Expectation – Maximization (EM) algorithm
  - Original algorithm by [Dempster, Laird and Rubin, 1977]
  - A general method for method for finding the maximum-likelihood estimate of a data distribution, when the data is partially missing or hidden.
    - In our case, data *X* are fully observed
    - The cluster assignments of an object x_i though can be seen as hidden variables

EM algorithm

- Initialize cluster assignments
- Two alternating steps:
  - E-step:

    re-estimate the expected-values of the hidden data (cluster assignments) under the current estimate of the model

  - M-step

    re-estimate the model parameters such that the likelihood according to the current estimate of the complete data is maximized
- Until convergence

$$\frac{\mathcal{L}_{new}}{\mathcal{L}_{old}} < 1 + \epsilon$$

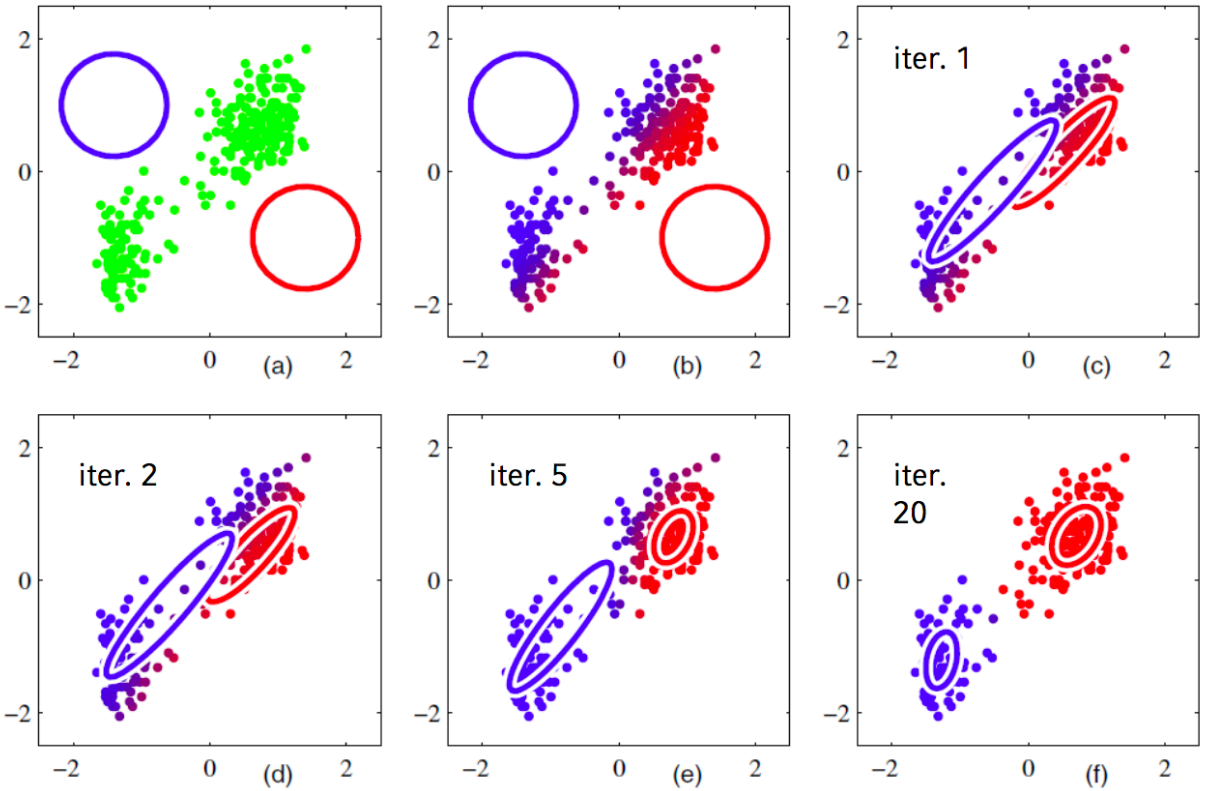
EM algorithm

- EM is similar to the k-Means algorithm
- k-Means for Euclidean data is a special case of EM for spherical Gaussian distributions with equal covariance matrices, but different means
- E-step (EM) → assign each object to a cluster step (k-Means)
  - In EM each object is assigned to a cluster with a probability
- M-step (EM) → compute cluster centroids step (k-Means)
  - In EM, the computation of the mean also considers the fact that each object belong to a distribution with a certain probability

EM algorithm

- Example of basic idea:

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



EM algorithm

- E-step: re-estimate the expected-values of the hidden data (cluster assignments) under the current estimate of the model

$$\mathrm{P}^{new}(c_l|x_i) = \mathrm{P}(c_l)\mathrm{P}(x_i|c_l)$$

EM algorithm

- M-step: re-estimate the model parameters such that the likelihood according to the current estimate of the complete data is maximized
  - Cluster densities:

$$\mathrm{P}^{new}(c_l) = \frac{1}{N}\sum_{i=1}^{N}\mathrm{P}^{new}(c_l|x_i)$$

  - Cluster means:

$$\mu_l^{new} = \frac{\sum_{i=1}^{N} x_i \mathrm{P}^{new}(c_l|x_i)}{\sum_{i=1}^{N}\mathrm{P}^{new}(c_l|x_i)}$$

  - Cluster covariances:

$$\sum_l^{new} = \frac{\sum_{i=1}^{N}(x_i-\mu_l^{new})(x_i-\mu_l^{new})'\mathrm{P}^{new}(c_l|x_i)}{\sum_{i=1}^{N}\mathrm{P}^{new}(c_l|x_i)}$$

EM (Gaussian Mixture Models) overview

- EM can be slow
- Not practical for models with a large number of components
- Problematic when clusters contain only a few points or if the points are nearly co-linear
- The choice of the exact model to use
- Difficulties with noise and outliers

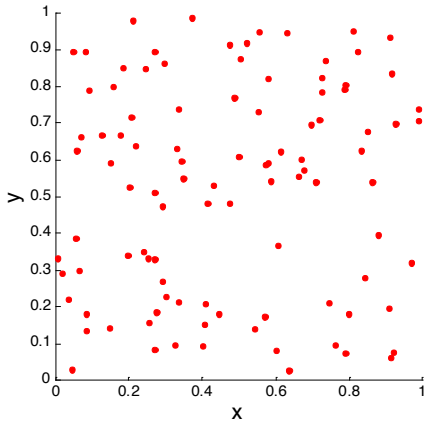
- More general than k-Means and fuzzy c-Means because they can use distributions of various types
- Thus, it can find clusters of different sizes and elliptical shapes
- It is easy to characterize the produced clusters

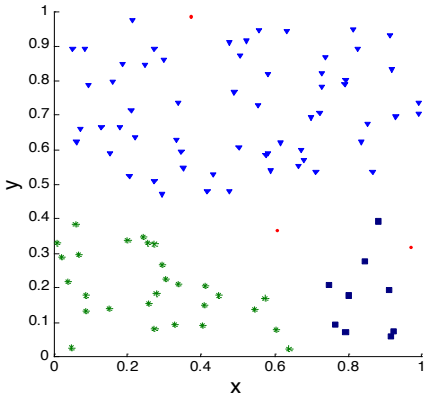
Cluster Validity

- In supervised learning, there is a variety of measures to evaluate how good a classifier is
  - accuracy, precision, recall, …
- For cluster analysis, the analogous question is how to evaluate the "goodness" of the resulting clusters?
- But "clusters are in the eye of the beholder"!
- Then why do we want to evaluate them?
  - To avoid finding patterns in noise
  - To compare clustering algorithms
  - To compare two sets of clusters
  - To compare two clusters

Clusters found in random data

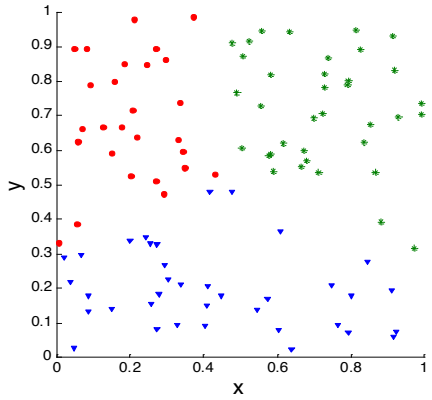
Random Points

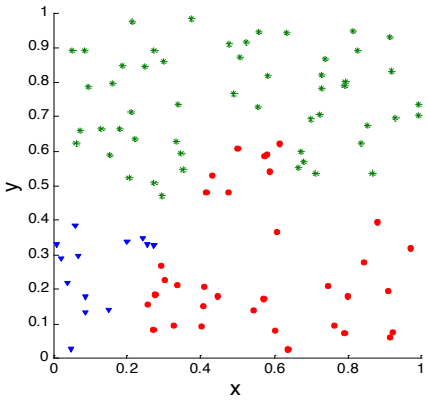




DBSCAN

K-means





Complete Link

Different Aspects of Cluster Validation

1. Determining the clustering tendency of a set of data, i.e., distinguishing whether non-random structure actually exists in the data.
2. Comparing the results of a cluster analysis to externally known results, e.g., to externally given class labels.
3. Evaluating how well the results of a cluster analysis fit the data *without* reference to external information.

    - Use only the data

4. Comparing the results of two different sets of cluster analyses to determine which is better.
5. Determining the 'correct' number of clusters.

For 2, 3, and 4, we can further distinguish whether we want to evaluate the entire clustering or just individual clusters.

Measures of Cluster Validity

- Numerical measures that are applied to judge various aspects of cluster validity, are classified into the following three types.
  - External Index: Used to measure the extent to which cluster labels match externally supplied class labels.
    - Entropy
  - Internal Index: Used to measure the goodness of a clustering structure *without* respect to external information.
    - Sum of Squared Error (SSE)
  - Relative Index: Used to compare two different clusterings or clusters.
    - Often an external or internal index is used for this function, e.g., SSE or entropy
- Sometimes these are referred to as criteria instead of indices
  - However, sometimes criterion is the general strategy and index is the numerical measure that implements the criterion.

Internal measures of cluster validity

- Idea: Check cluster characteristics, do not rely on external information
- Examples: cohesion and separation
- Cluster Cohesion: Measures how closely related are objects in a cluster
  - Example: SSE
- Cluster separation: Measures how distinct or well-separated a cluster is from other clusters
  - Cohesion is measured by the within cluster sum of squares (SSE)

$$WSS = \sum_i \sum_{x \in C_i} (x - m_i)^2$$

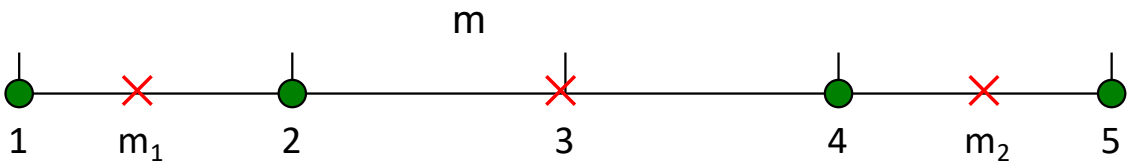
  - Separation is measured by the between clusters sum of squares

$$BSS = \sum_i |C_i| (m - m_i)^2$$

    - Where $|C_i|$ is the size of cluster i and m is the overall mean of all data points

Example

- Example: SSE



K=1 cluster:

$$WSS = (1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2 = 10$$

$$BSS = 4 \times (3-3)^2 = 0$$

$$Total = 10 + 0 = 10$$

K=2 clusters:

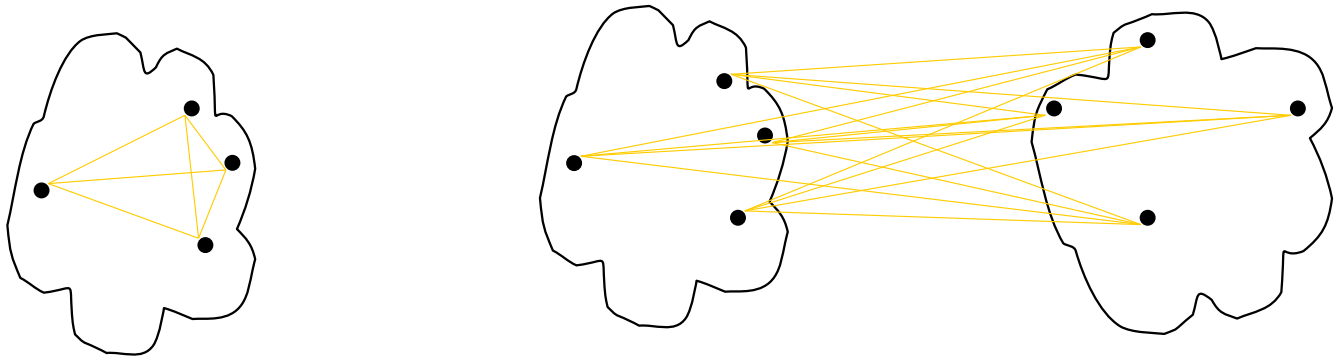
$$WSS = (1-1.5)^2 + (2-1.5)^2 + (4-4.5)^2 + (5-4.5)^2 = 1$$

$$BSS = 2 \times (3-1.5)^2 + 2 \times (4.5-3)^2 = 9$$

$$Total = 1 + 9 = 10$$

Internal measures of cluster validity

- A proximity graph based approach can also be used for cohesion and separation.
  - Cluster cohesion is the sum of the weight of all links within a cluster.
  - Cluster separation is the sum of the weights between nodes in the cluster and nodes outside the cluster.



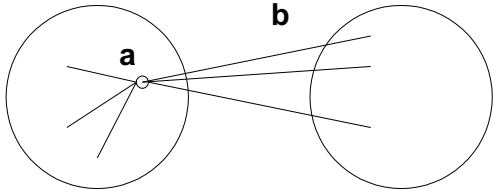
cohesion separation

Internal Measures: Silhouette Coefficient

- Silhouette Coefficient combine ideas of both cohesion and separation, but for individual points, as well as clusters and clusterings
- For an individual point, *i*
  - Calculate ***a*** = average distance of *i* to the points in its cluster
  - Calculate ***b*** = min (average distance of *i* to points in another cluster)
  - The silhouette coefficient for a point is then given by

    s = 1 – a/b if a < b, (or s = b/a - 1 if a ≥ b, not the usual case)

  - Typically between 0 and 1.
  - The closer to 1 the better.



- Can calculate the Average Silhouette width for a cluster or a clustering

External measures of cluster validity

- Idea: Measure the extent to which cluster labels match externally supplied class labels.
- Examples: entropy, purity

Table 5.9. K-means Clustering Results for LA Document Data Set

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Entropy	Purity
1	3	5	40	506	96	27	1.2270	0.7474
2	4	7	280	29	39	2	1.1472	0.7756
3	1	1	1	7	4	671	0.1813	0.9796
4	10	162	3	119	73	2	1.7487	0.4390
5	331	22	5	70	13	23	1.3976	0.7134
6	5	358	12	212	48	13	1.5523	0.5525
Total	354	555	341	943	273	738	1.1450	0.7203

Cluster

Class distribution

- Entropy of a cluster *j*: how pure in terms of the classes a cluster is: $e_j = \sum_{i=1}^{L} p_{ij} \log_2 p_{ij}$
  - p_{ij}: the probability of observing class *i* in cluster *j*. $p_{ij} = m_{ij}/m_j$
- Entropy of a clustering: $e = \sum_{i=1}^{K} \frac{m_i}{m} e_j$

External measures of cluster validity

- Purity focuses on the most likely class in the cluster

Table 5.9. K-means Clustering Results for LA Document Data Set

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Entropy	Purity
1	3	5	40	506	96	27	1.2270	0.7474
2	4	7	280	29	39	2	1.1472	0.7756
3	1	1	1	7	4	671	0.1813	0.9796
4	10	162	3	119	73	2	1.7487	0.4390
5	331	22	5	70	13	23	1.3976	0.7134
6	5	358	12	212	48	13	1.5523	0.5525
Total	354	555	341	943	273	738	1.1450	0.7203

Cluster **Class distribution**

- Purity of cluster *j*: $purity_j = \max p_{ij}$
- Purity of the clustering: $purity = \sum_{i=1}^{K} \frac{m_i}{m} purity_j$