Business Intelligence
SS 2017

Cross-sectional Analysis 3
Clustering Methods

W. Grossmann

Content

- Problem Formulation
- Hierarchical Clustering
- Partitioning Methods
- Other Methods

Problem Formulation

- In case of clustering the data contain no output variable and we want to find a group structure in the data, such that the observations in the groups are rather homogeneous with respect to the variables
- Data N observations in p variables

Variables: $\vec{X} = (X_1, X_2, \ldots, X_p)$

Observations: $\vec{x} = (x_1, x_2, \ldots, x_p)$

Problem Formulation

- Main prerequisite for clustering is a distance between observations
- Most important distance for quantitative variables is the Euclidean distance:

$$d^2(\vec{x}, \vec{z}) = \sum_{i=1}^{p} (x_i - z_i)^2$$

- Alternatives:
  - Absolute deviation $d_1(\vec{x}, \vec{z}) = \sum_{i=1}^{p} | x_i - z_i |$
  - Maximum distance: $d_s(\vec{x}, \vec{z}) = \max_{p} | x_i - z_i |$

Problem Formulation

- In case of binary variables the Hamming distance is frequently used, which is defined by the number of different values in the variables
  - The Hamming distance is equivalent to the Euclidean distance

Problem Formulation

- Another measure frequently used is the Cosine similarity defined by the angle between two vectors
  - The cosine defines a similarity of two vectors; if the have the same direction, the similarity is 1, if the vectors are orthogonal the similarity is 0, if the vectors have opposite direction the similarity is -1
  - Cosine similarity is frequently used if the vectors are binary vectors

Problem Formulation

- In case qualitative variables on can reduce the problem to a problem for binary variables by defining indicator variables for each attribute value
  - q different values lead to $\ln_2(q)$ binary variables
- Combination of qualitative and quantitative variables: procedure `daisy` in R

Problem Formulation

- For more complex structures like string variables (text) or graphs the distance calculation can be based on kernels
  - String kernels: based on counting the simultaneous occurrence of substrings of certain length

Problem Formulation

- An important issue is many times standardization of the variables
  - All variables are standardized with zero and variance 1
  - All variables are standardized such that the values are in the interval [0 , 1] , or [-1 , 1]

Problem Formulation

- Main goals in clustering:
  - Find homogeneous groups, i.e. , variability of observation within groups is explanation of observation, distance between groups measures separation of groups
- Main analysis tasks
  - Determine the number of clusters
  - Assign observations to clusters
  - Find representative objects in the groups (vector quantization)

Problem Formulation

- Evaluation of clusters
  - Homogeneity measures for groups
  - Validation with a test sample
  - Validity of the solution with respect to a subject matter explanation
- Types of cluster algorithms
  - Hierarchical methods
  - Partitioning Methods
  - Combination of the methods
  - Model based clustering

Hierarchical Clustering

- Most frequently used is agglomerative clustering
- Basic outline of an agglomerative cluster algorithm:

Algorithm 4: Agglomerative clustering

Define clusters $C_k, 1 \leq k \leq N$ by the observations, $N_{cl} = N$;
for $k = 1$ *to* $N - 1$ **do**
  Merge clusters C_r and C_s for which $d(C_r, C_s) = \min_{(l,k)} D((C_l, C_k))$;
  $N_{cl} = N_{cl} - 1$;
end

Hierarchical Clustering

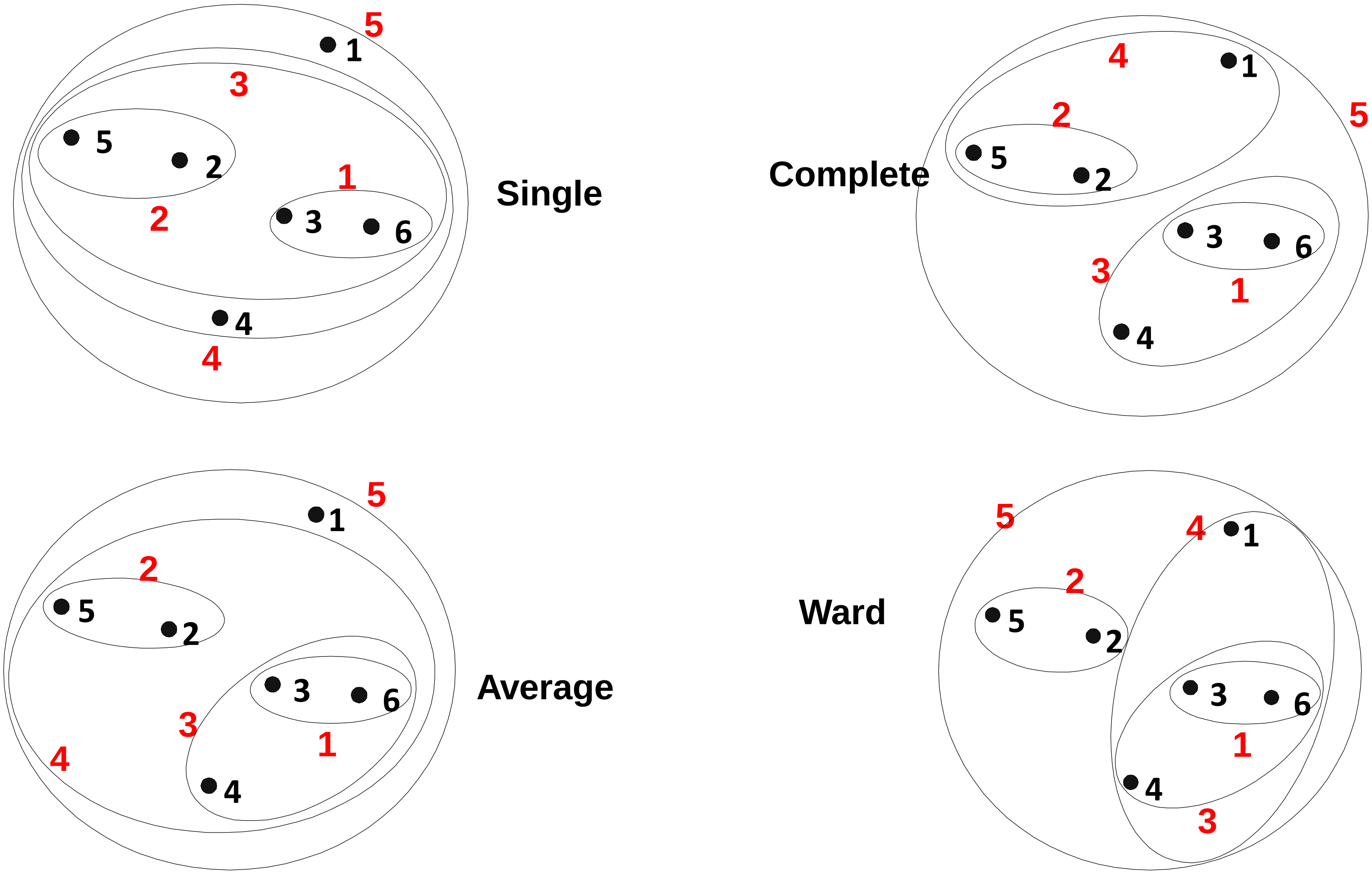
- Main problem is the definition of the distance between the clusters based on the distance between the objects, and determination of number of clusters
- The distance is called the linkage of the clusters
- Different specifications are possible
  - Single linkage: Distance between clusters is the distance of the closest points (minimum spanning tree)
  - Complete linkage: Distance between clusters is the distance of the farthest points

Hierarchical Clustering

- Average linkage: mean distance between all the point sin the two clusters
- Ward distance: difference between the total within cluster sum of squares for the two clusters separately, and the within cluster sum of squares resulting from merging the two clusters in cluster

- In general average linkage and Ward's method are recommended
- Single linkage is not useful in most cases because the clusters are forming chains

Hierarchical Clustering, Comparison of Linkages

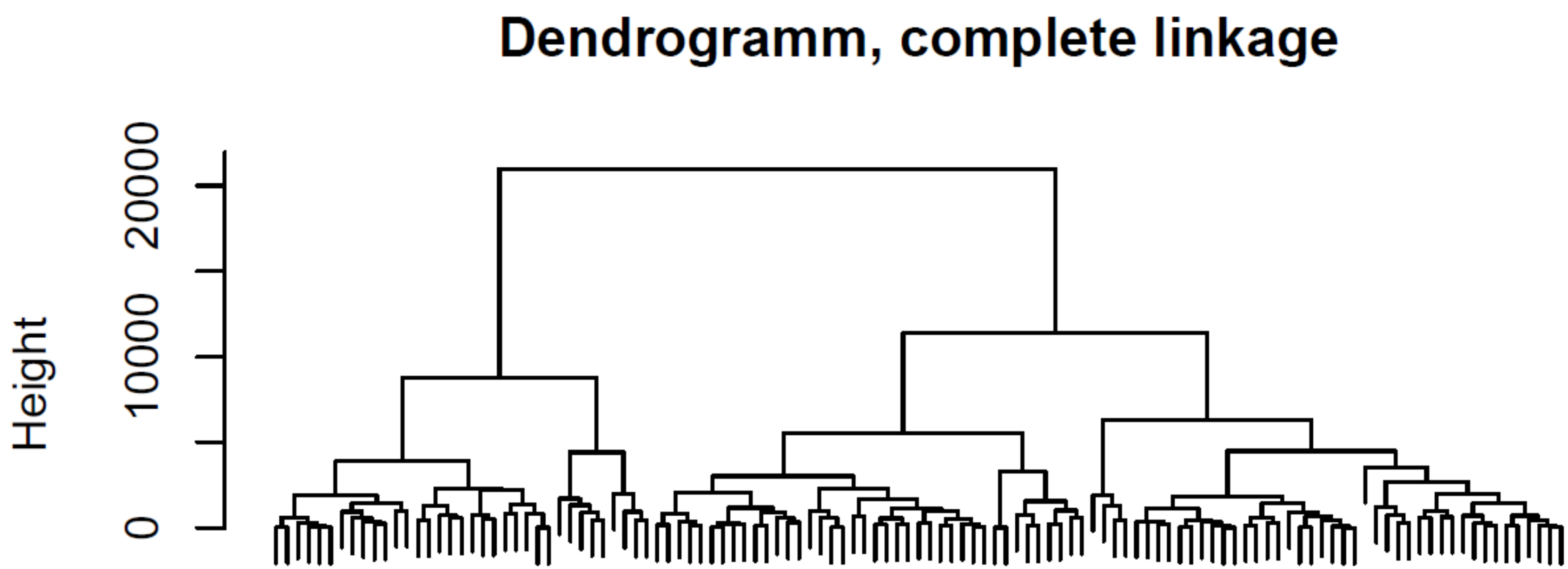


Hierarchical Clustering

- Determination of the number of clusters is done by visual inspection of the distance of the clusters which are merged
- Most popular method is using the dendrogram
  - A dendrogram is a visual representation of the aggregation process as a tree
  - The leaves of the tree are defined by the objects
  - Other nodes are formed according to the aggregation process
  - The heights of branches is given by the distance

Hierarchical Clustering

- Example of a dendrogram

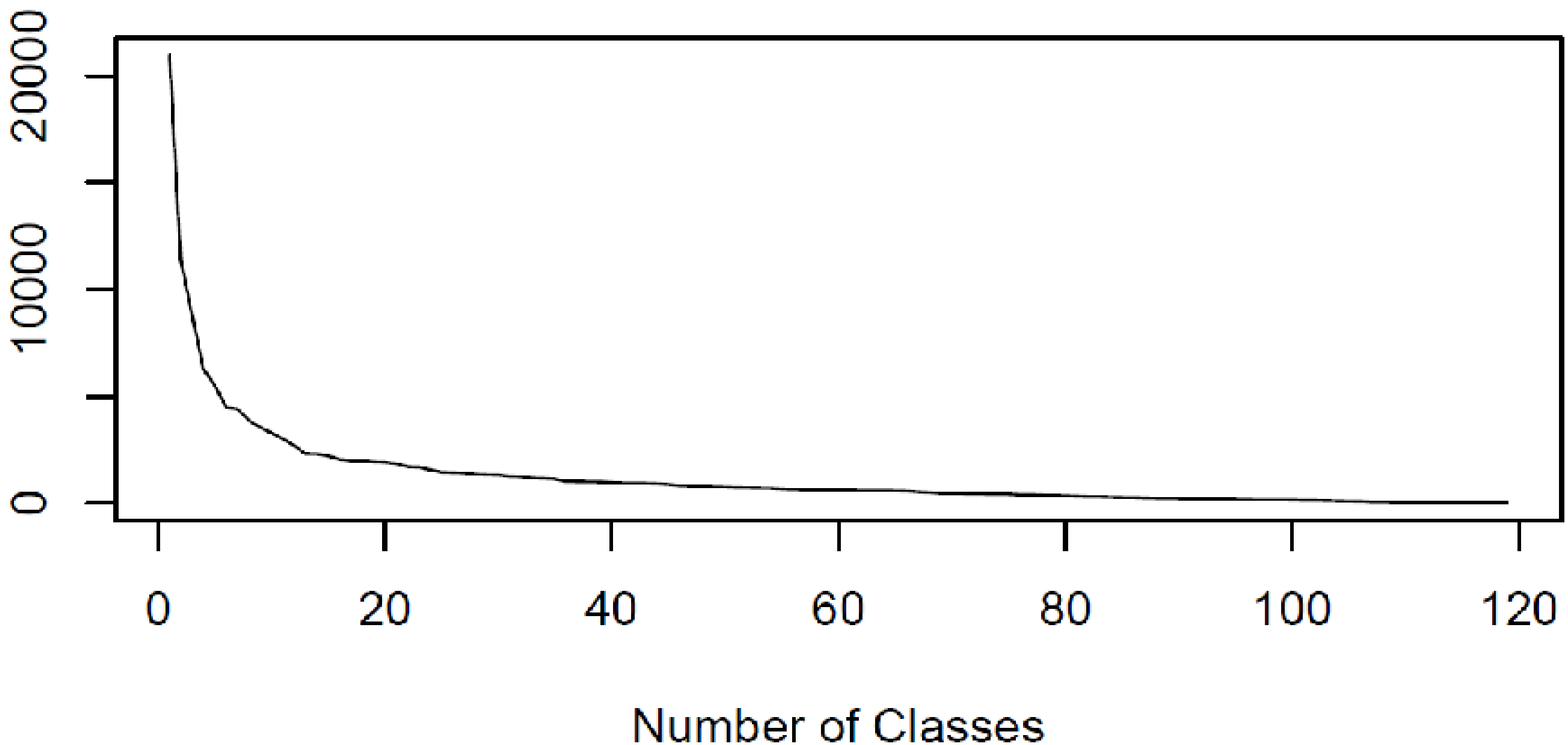


Hierarchical Clustering

- An alternative to the dendrogram is a scree plot of the distance between the merged classes in dependence of the number of classes
- The decision about the number of classes is defined by the elbow of the scree plot

Hierarchical Clustering

- Example of a scree plot

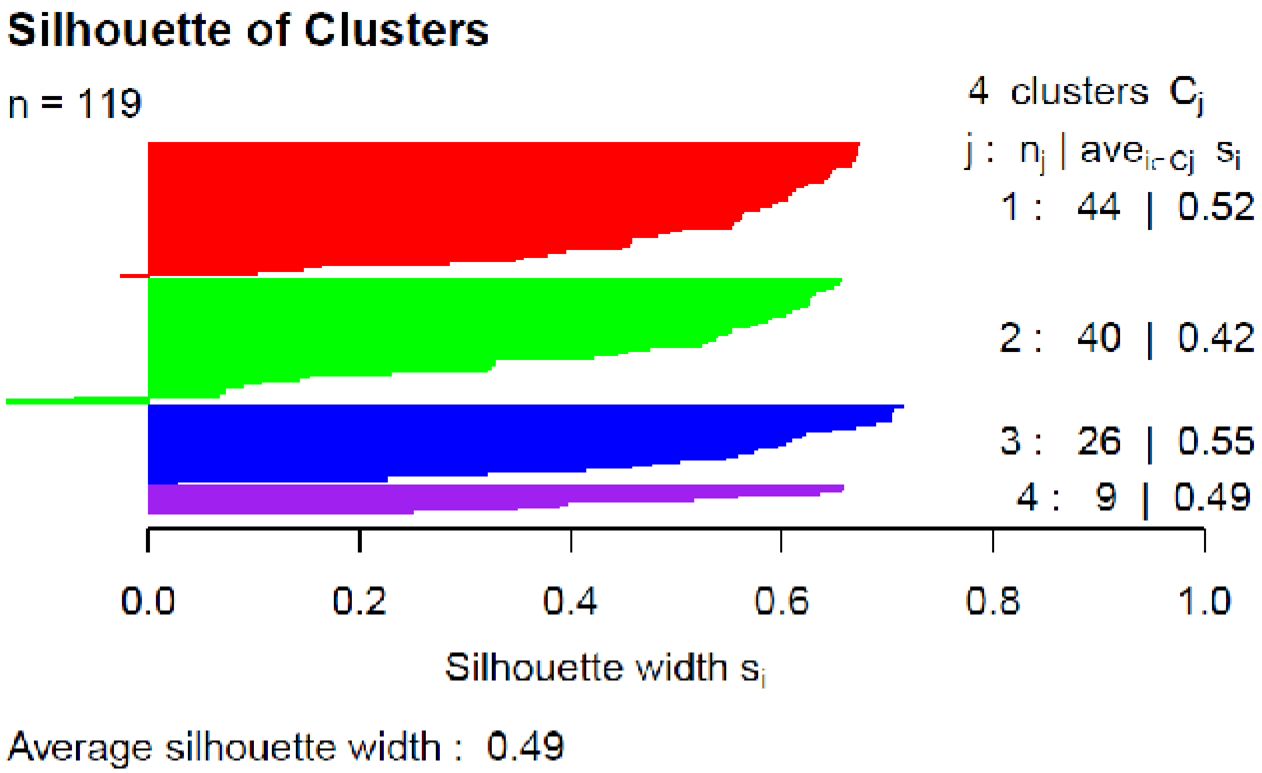


Hierarchical Clustering

- A method for evaluation of the cluster solution is the silhouette plot
  - The silhouette shows for each point how well the point is located in the cluster
  - A value close to 1 shows that the point is well located in the cluster
  - A negative value indicates that the point is not well assigned
- The silhouette can be used also for comparing different cluster solutions

Hierarchical Clustering

- Example of a silhouette



Hierarchical Clustering

- Properties of Hierarchical clustering
  - Good visualization of the solution
  - Decision about the numbers of clusters can be done after analysis
  - Limited to small number of observations
  - The decision about the cluster assignment cannot be changed in the algorithm

Partitioning Methods

- Partitioning methods define in an iterative way a cluster solution for the observations given the number of clusters in advance
- The most popular method is k-means clustering, where k stands for the number of clusters

Partitioning Methods

- Basic algorithm

Algorithm 5: k-Means Algorithm

Data: Observation matrix $\mathbf{X}$ and distance for the objects; number of clusters K.
Result: Cluster solution for observations

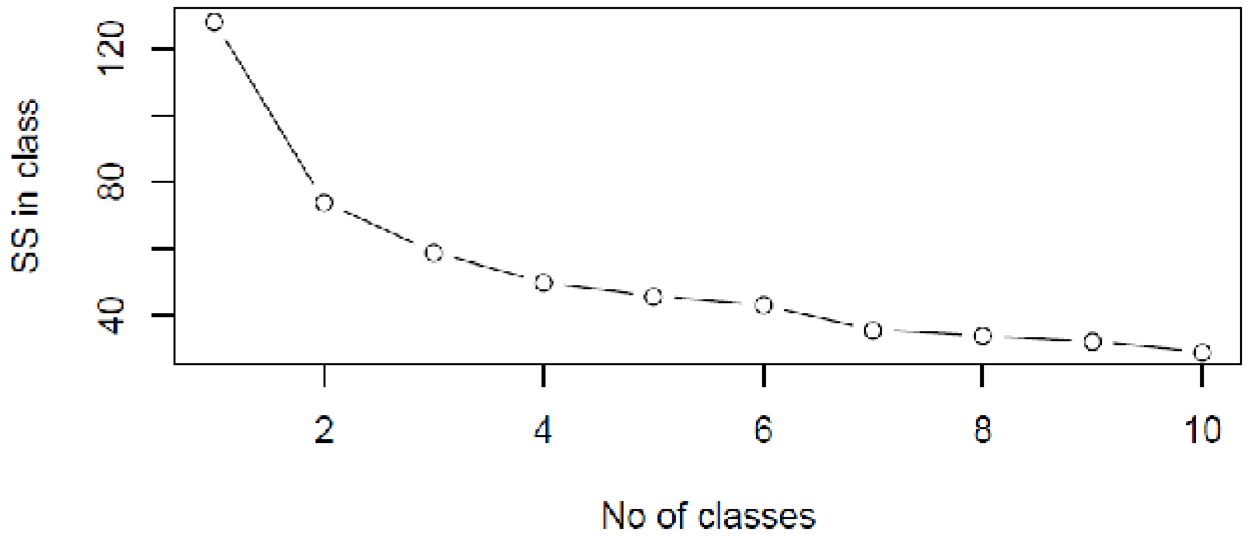
1 **begin**
2 Define an initial solution for the cluster centers $(c_1, c_2, \ldots, c_k)$;
3 Assign each observation x to the cluster which center is closest to the observation;
4 Compute new centers for the clusters as means of the assigned observations;
5 Repeat steps 2 and 3 as long as there is no significant change in the centers;
6 **end**

Partitioning Methods

- Main points in application
  - Decision about the number of clusters
  - Finding initial centers of the clusters
- Determination of the number of clusters can be based on a visualization of the sum of squares within the clusters for a solution in dependence of the number of clusters
  - This is similar to the idea of variance decomposition in case of analysis of variance

Partitioning Methods

- Usually the plot shows a shape similar to a scree plot and the decision is based on the elbow criterion



Partitioning Methods

- With respect to the initial solution the standard procedure is choosing the centers randomly and try different solutions
- Properties of k-means clustering
  - Procedure is fast from computational point of view
  - Applicable for large data sets (parallel implementations exist)
  - A found solution can be applied to new observations (cf. nearest neighbor classification)

Other Methods

- There exist numerous clustering algorithms for specific problems
  - Self organizing maps (SOM) are clustering methods based on the idea of a neural net. They can be understood as a k-means clustering defined on a distorted grid
  - Two stage clustering combine the ideas of hierarchical clustering with the ideas of k-means (IBM/SPSS) by using a cluster-feature tree
  - Model based clustering looks at the problem from a more theoretical perspective and defines a model for the data by a mixture of normal distributions