

OPPA European Social Fund Prague & EU: We invest in your future.

Cluster analysis – formalism, algorithms

Jiří Kléma

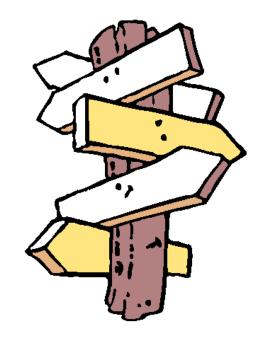
Department of Cybernetics, Czech Technical University in Prague



http://ida.felk.cvut.cz

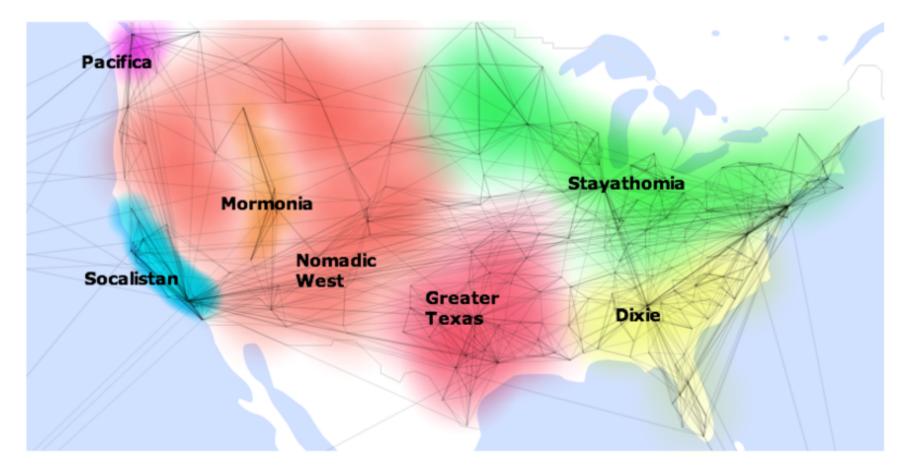
Outline

- motivation
 - why clustering? applications,
- clustering as an optimization task
 - complexity,
- k-means algorithm
 - direct greedy search,
 - (dis)advantages,
- generalization
 - k-means as an instance of EM algorithm,
 - soft clustering,
 - EM algorithm and Gaussian distribution mixture,
- hierarchical clustering
 - motivation extras?
 - agglomerative and divisive approach,
- summary, method categorization.



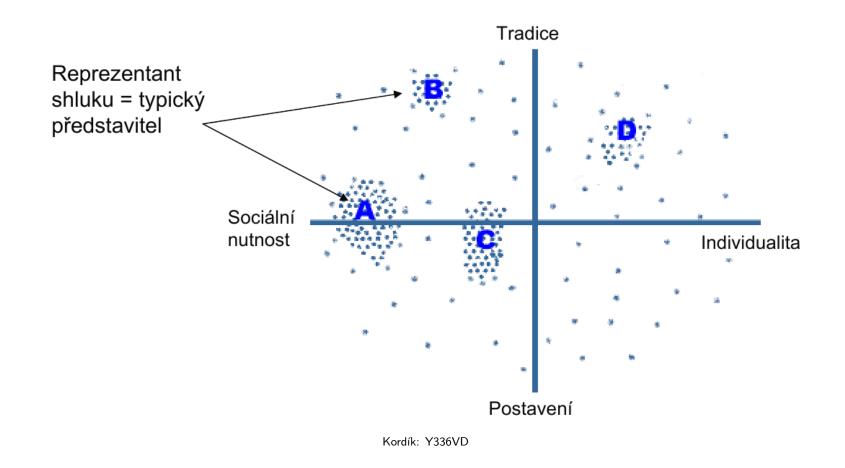
Clustering – example

- clusters and their prototypes bring new domain knowledge,
- interpretation e.g. in connection with geographic data and visualization,
- "clustering" 210 million Facebook profiles based on friendship connections,



Clustering – example

- clusters and their prototypes bring new domain knowledge,
- survey: why do people drink alcohol?
- goal: to find "sample drinkers" and represent condensed people's attitude to alcohol,



Clustering – example

- application for image segmentation,
- features: (coordinates), (a) color components, (b) brightness for b&w image.



Xiao Zhang: Image Segmentation.

Clustering – utilization, applications

- clustering for learning
 - class discovery in (unannotated) data,
 - unsupervised learning,
- data understanding, their structured representation
 - taxonomies (biology organisms, genes),
 - rapid access to pieces of information (web search engine output organization),
 - outlier detection,
- usage of prototypes
 - summarization (original objects completely forgotten),
 - compression (vector quantization),
 - efficient nearest neighbor search.

Clustering – formalization

goal

- split unclassified objects into mutually disjoint subsets, clusters,
- we divide so that the objects
 - 1. are similar inside a cluster,
 - 2. are dissimilar when lying in different clusters,
- disjoint partition of an object set defined in an input space (usually \mathbb{R}^n) into k > 1 classes $\mathcal{X} \dots$ a set of m objects, $\Omega = \{C_1, \dots, C_k\} \dots$ partition of the set \mathcal{X} , $\forall i, j \leq k, i \neq j \ C_i \neq \emptyset, \ C_i \cap C_j = \emptyset, \ C_1 \cup C_2 \cup \dots \cup C_k = \mathcal{X}$,
- (i,j) = (i,i) / j = (i,j) + (i,j) + (i,j) = (i,j) + (i,j) +
- we solve an optimization problem
 - inputs
 - * training data,
 - * distance function (dissimilarity function),
 - * (optimization criterion).
 - unknown
 - * the number of clusters,
 - * cluster-object links partition,
 - * (prototypes cluster ethalons, typical examples).

Clustering – complexity

variant of a Bayesian decision-making task

develop a strategy $Q : \mathcal{X} \to D$ (D stands for decisions) minimizing $\underset{q}{\operatorname{argmin}} \sum_{x \in \mathcal{X}} p(x) W(x, q(x))$ (W is a loss function),

- how large space to be searched?
 - the number of different disjoint partitions: Stirling number of the second kind

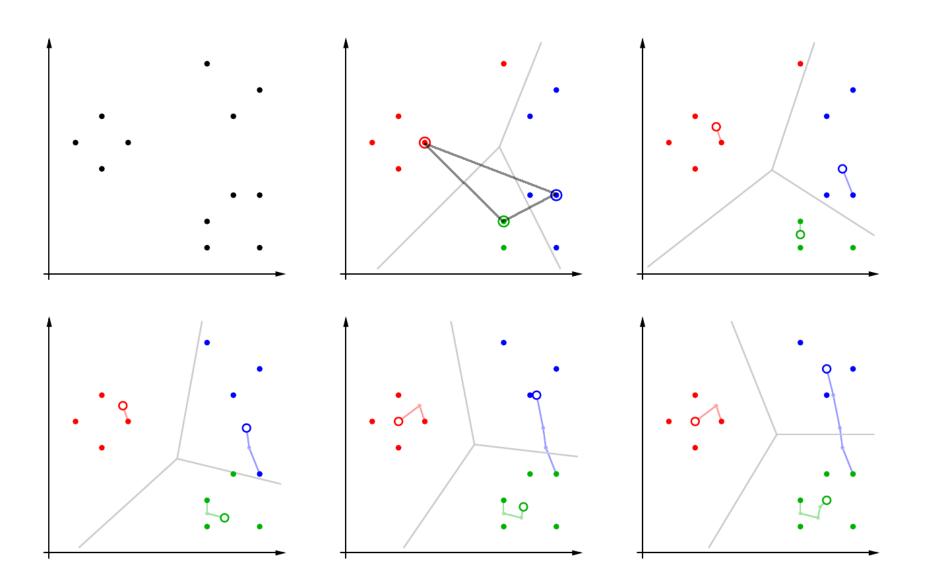
$$S(m,k) = {m \\ k} = \frac{1}{k!} \sum_{j=0}^{k} (-1)^{k-j} {k \choose j} j^{m}, \text{ among other } S(m,2) = {m \\ 2} = 2^{m-1} - 1$$

$m \backslash k$	1	2	3	4	5	6	7	8
2	1	1						
3	1	3	1					
4			6					
5	1	15	25	10	1			
				65		1		
7				350		21		
8	1	127	966	1701	1050	266	28	1

- the optimization criterion cannot be applied in a naïve way (exhaustive search),

• NP-hard problem, heuristic solutions.

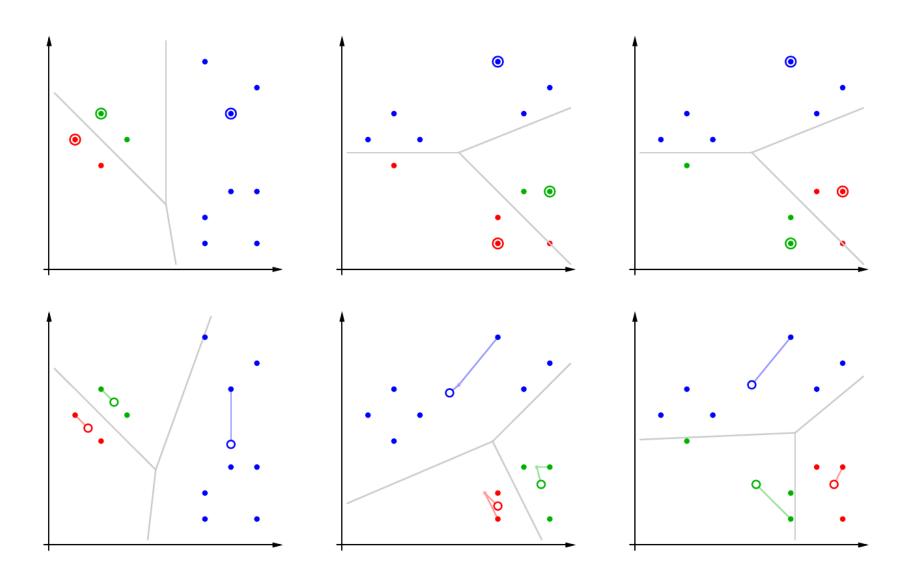
K-means – strategy, an ideal run (Borgelt: IDA slides)



K-means algorithm

- global homogeneity criterion: $W(k) = \underset{O}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{x_j \in C_i} d(x_j, \mu_i)$,
- inputs: $\mathcal{X} = \{x_1, \dots, x_m\} \subset \mathbb{R}^n$, $k \in \mathbb{N}$, $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$,
- 1. randomly **initialize** cluster centroids μ_j (e.g. select k objects),
- 2. each object $x_i \in \mathcal{X}$ assign to the nearest centroid $\forall i \underset{i=1...k}{\operatorname{argmin}} d(x_i, \mu_j)$,
- 3. recompute cluster centroids centroid is a mean vector of objects assigned to the cluster,
- 4. repeat steps 2 and 3 until cluster centroids change.
- greedy algorithm
 - guaranteed convergence, typically fast,
 - finds a locally optimal solution,
 - initialization sensitive,
- illustrative demo applet
 - http://www.kovan.ceng.metu.edu.tr/~maya/kmeans/index.html.

K-means – stucked in local optima (Borgelt: IDA slides)



Distance function

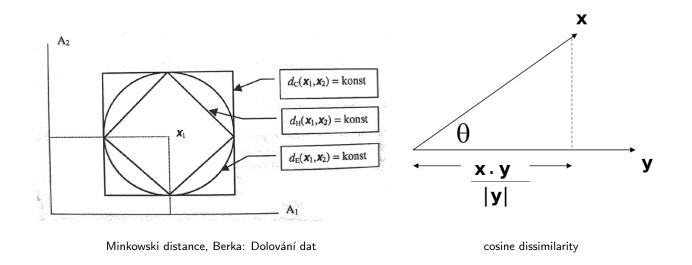
• typically metric on \mathcal{X} , $\forall x, y, z \in \mathcal{X}$:

 $- \ d(x,y) \geq 0 \text{, } d(x,y) = 0 \Leftrightarrow x = y \text{, } d(x,y) = d(y,x) \text{, } d(x,z) \leq d(x,y) + d(y,z)$

common functions

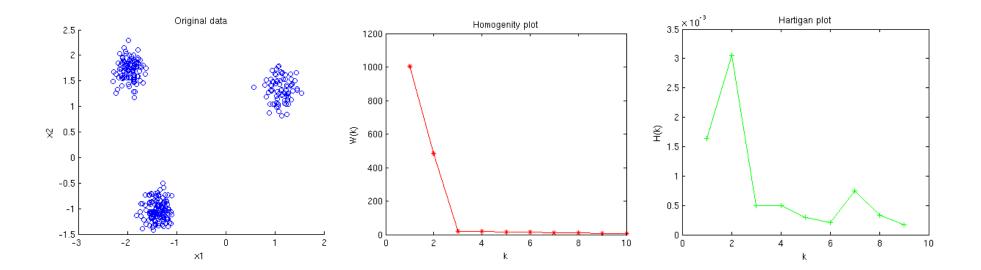
- Minkowski metric: $d(x, y) = \left(\sum_{i=1}^{n} (x_i y_i)^k\right)^{\frac{1}{k}}$ * selection of k: $d_H(k = 1)$ (Manhattan, Hamming, taxi), $d_E(k = 2)$ (Euclid), $d_C(k = \infty)$ (Chebyshev),
- cosine dissimilarity (documents): $d(x, y) = 1 cos(\theta) = 1 \frac{x \cdot y}{|x||y|}$
- edit (Levenshtein) distance (words, strings, sequences)

* minimum number of edits (change, insert, delete) to transform one string into the other.



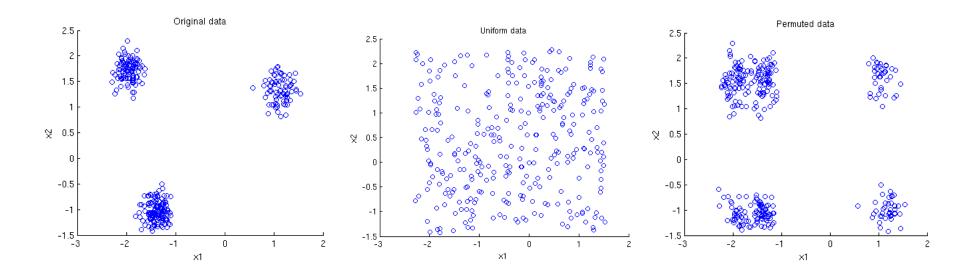
K-means: the number of clusters choice

- k known a priori,
- k based on the object number only: $k \sim \sqrt{\frac{m}{2}}$,
- homogeneity W necessarily monotonously increases with increasing k, a heuristic "elbow" method:
 - run k-means algorithm repeatedly with increasing k,
 - a proper k is in the point of sudden non-homogeneity decrease or in a curve elbow,
 - Hartigan criterion: $H(k) = \frac{W(k) W(k+1)}{W(k+1)(m-k-1)}$ choose the smallest $k \ge 1$ with H(k) small enough.

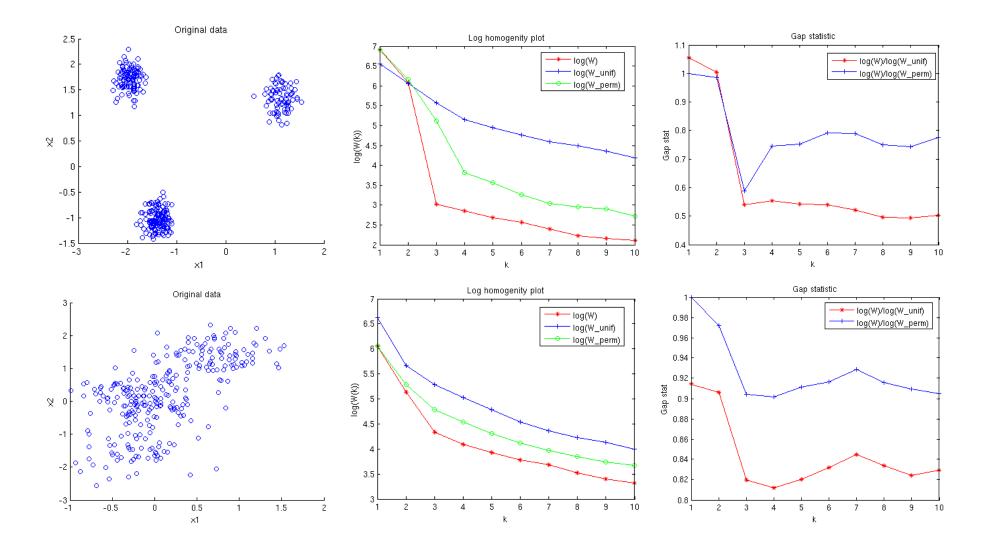


K-means: the number of clusters choice

- Tibshirani (2001): gap statistic
 - compares development of W(k), resp log(W(k)), with the referential curve $W_{ref}(k)$,
 - instead of log(W(k)) searches minimum in $\frac{log(W(k))}{log(W_{ref}(k))}$,
 - $-W_{ref}(k)$ can be obtained in two ways
 - * uniform distribution homogeneity "without clusters" ($W_{unif}(k)$),
 - * permuted distribution homogeneity feature values randomly shuffled ($W_{perm}(k)$),
 - \ast the domain is kept in both,
 - the method originated in statistics.



K-means: the number of clusters choice



• EM with theoretically well-founded AIC or BIC criteria.

Expectation Maximization (EM) algorithm

- k-means is an EM algorithm specialization,
- maximizes likelihood $Pr(\mathcal{X}|\theta)$

$$\theta^* = \underset{\theta}{\operatorname{argmax}} Pr(\mathcal{X}|\theta) = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^m Pr(x_i|\theta)$$

• introduces a latent variable Q, which simplifies maximization of $Pr(\mathcal{X}|\theta)$

- E-step:

* estimate latent variable (distribution) for the given data and current param values θ ,

– M-step:

* modify parameters θ so that likelihood is maximized wrt given Q,

- k-means specification
 - -Q gives binary cluster membership,
 - E-step: assign objects and centroids,
 - M-step: recalculate cluster centroids.

Soft (probabilistic) clustering

- "hard" object membership in a single cluster not needed,
- membership function $Pr(C_j|x_i)$ is understood as probability
 - it must hold: $\forall i = 1, \dots, m : \sum_{j=1,\dots,k} Pr(C_j|x_i) = 1$
- a soft clustering algorithm "soft" k-means
 - EM principle,
 - a model with parameters heta used to calculate $Pr(C_j|x_i)$,
 - θ most often defines a Gaussian Mixture Model (GMM),

$$* Pr(x_i|\theta) = \sum_{j=1}^k \alpha_j \frac{1}{(2\pi)^{n/2} |\Sigma_j|^{1/2}} e^{-\frac{1}{2}(x_i - \mu_j)^t \sum_j^{-1} (x_i - \mu_j)}$$

* $\theta = \{\alpha_1, \dots, \alpha_k, \mu_1, \dots, \mu_k, \Sigma_1, \dots, \Sigma_k\}, \sum_{j=1}^n \alpha_j = 1$

 $* \alpha_i \ldots$ a mixture element weight, $\mu_i \ldots$ centroid vector, $\Sigma_i \ldots$ covariance matrix,

- - heta can also define a naïve bayes model etc.,
- EM GMM clustering

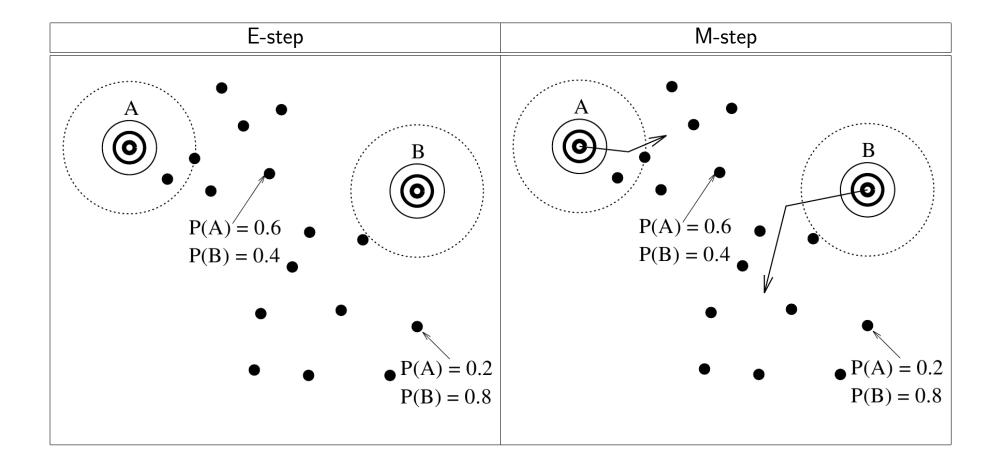
-Q determines probability that an object was generated by a particular gaussian distribution,

soft clustering is a special case of fuzzy clustering

- membership $Pr(C_j|x_i)$ without constraints needed for probability.

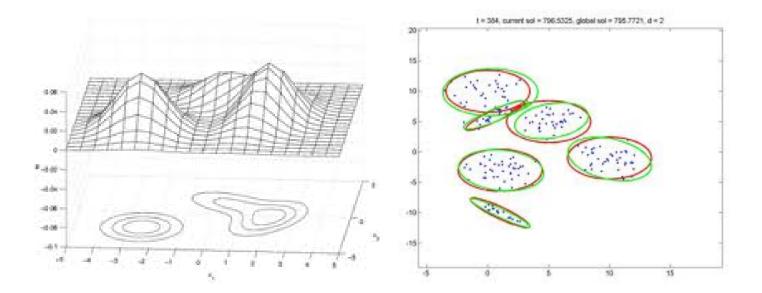
EM for GMM clustering

- EM is an iterative algorithm,
- illustration of one step after random initialization.

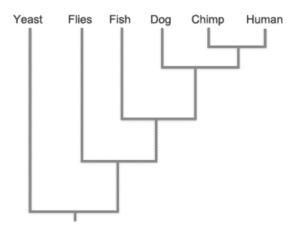


EM clustering – k-means comparison

- clustering defined as GM optimization in n dimensions,
- the number of elements (distributions) k (can be a part of likelihood maximization resp. AIC),
- partition: object belongs to the distribution with the highest a posteriori prob $Pr(C_j|x_i)$,
- assumes a normal object distribution within a cluster,
- more robust, but slower than k-means,
- demo: http://staff.aist.go.jp/s.akaho/MixtureEM.html.

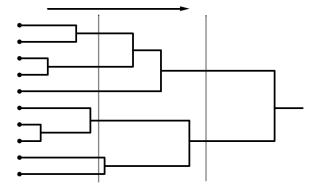


- taxonomy is more informative than partition
 - analyzes on various granularity levels,
 - binary tree = **dendrogram**,
- a reasonable decomposition of the clustering problem to subproblems
 - a straightforward and computationally efficient solution.



Hierarchical clustering – algorithm

- recursive application of the standard clustering step,
- agglomerative approach (bottom-up)
 - at the beginning each object makes a cluster,
 - iterate with merging the most similar clusters, typically pairs,
- divisive approach (top-down)
 - split the object set into clusters, typically two of them,
 - iterate with splitting the clusters,
 - more difficult to implement needs an internal clustering algorithm,
 - more efficient than agglomerative, namely when the complete dendrogram not needed,
- needs no prior k, constructs a hierarchy.
- a partition results from a dendrogram cut.



Hierarchical clustering – cluster distance

- the key point is a generalized cluster distance function
 - makes a step from the object distance towards the object set distance,
 - originally: $d: \mathcal{X} imes \mathcal{X}
 ightarrow \mathbb{R}$,
 - $\text{ now: } \delta: 2^{\mathcal{X}} \times 2^{\mathcal{X}} \to \mathbb{R},$
- ${\scriptstyle \bullet}$ elemental δ definitions based on d
 - concern two most similar objects (single linkage)

$$\delta(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x, y),$$

- concern two most distant objects (complete linkage)

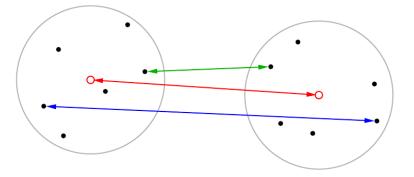
$$\delta(C_i, C_j) = \max_{x \in C_i, y \in C_j} d(x, y),$$

– average pair distance (average linkage)

$$\delta(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i} \sum_{y \in C_j} d(x, y)$$
 ,

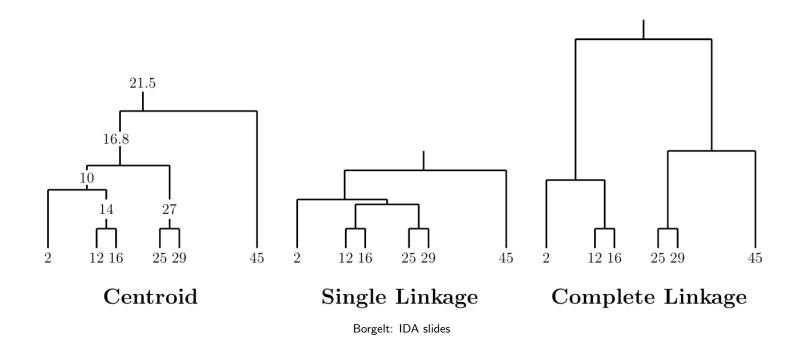
- distance between cluster centroids (centroid)

$$\delta(C_i, C_j) = d(\mu_i, \mu_j),$$



Example: relation between distance function and clustering outcome

- Ex.: 1 dimensional object set 2, 12, 16, 25, 29, 45.
 - the objects can be proportionally positioned on x dendrogram axis,
- different generalized distance functions lead to different dendrograms.

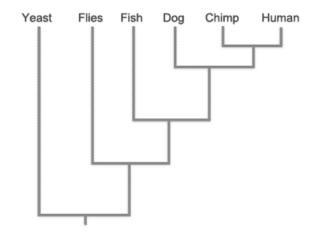


Clustering – summary

- Intuitively comprehensible principle, in many contexts, in many domains
 - in general identification of any frequent event co-occurrence in data,
- combinatorially difficult optimization problem
 - heuristic solutions, local optimality,
- basic steps
 - representation definition,
 - distance function selection,
 - clustering itself,
 - abstract representation of partition,
 - evaluation, iteration.
- clustering algorithm quality
 - scalability no of objects, dimensions,
 - robustness noise, outliers, feature types, distance function,
 - ability to deal with various cluster shapes.

Clustering – method categorization

- nonhierarchical methods
 - aim to deliver the partition that minimizes an optimization criterion,
 - apply a global homogeneity criterion,
 - cluster membership can be hard (crisp) as well as probabilistic,
 - examples: k-means, EM
- hierarchical methods
 - generate a cluster hierarchy
 - * binary tree = dendrogram,
 - apply a local cluster similarity criterion,
 - agglomerative bottom-up,
 - divisive top-down, divide and conquer,
 - examples: AHC (a general principle).



Recommended reading, lecture resources

:: Reading

- Hastie et al.: The Elements of Statistical Learning: DM, Inference and Prediction.
 - Springer book.
- Jain et al.: Data Clustering: A Review.
 - ACM Computing Surveys,
 - http://www.prip.tuwien.ac.at/teaching/ss/einfuhrung-in-die-mustererkennung/download-areaand-links/p264-jain.pdf.
- Borgelt: Intelligent Data Analysis.
 - slides, a detailed intelligent data analysis course, clustering near the end,
 - http://www.borgelt.net/courses.html#ida,



OPPA European Social Fund Prague & EU: We invest in your future.