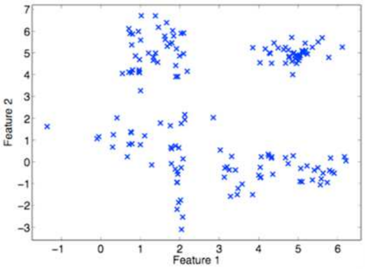
G. gallus

Clustering [and Some Dissimilarities]

APR Course, Delft, The Netherlands

Marco Loog

Clustering



- What salient structures exist in the data?
- How many clusters?

Cluster Analysis

- Grouping observations based on [dis]similarity
  - Data mining [exploration, searching for concepts]
    - Relating species based on genetic similarity
    - Reducing amount of data to be analysed, helps defining concept [class]
  - Selecting typical class examples
    - Multi-modal classes may be represented using typical examples
    - Interpretation is not a goal here!
  - Image presegmentation / oversegmentation

Dissimilarity Measures

- Let d(r,s) be dissimilarity between objects r, s
- Formally, dissimilarity measures should satisfy
  - d(r,s)≥0
  - d(r,r)=0
  - d(r,s)=d(s,r)
- If triangle inequality holds measure is a metric
  - d(r,t)+d(t,s)≥d(r,s)

E.g. Measures Between Distributions

•Histogram intersection

$$D_{hist}(H,K) = 1 - \frac{\sum_i \min(h_i, k_i)}{\sum_i k_i}$$

•Kullback-Leibler divergence

•Efficiency coding distribution using other as code-book

$$D_{KL}(H,K) = \sum_i h_i \log \frac{h_i}{k_i}$$

•Kolmogorov-Smirnov

•Maximum difference between cumulative distributions

$$D_{KS}(H,K) = \max_i(|\hat{h}_i - \hat{k}_i|)$$

•Chi squared statistic

•Likelihood of one distribution drawn from the other

$$D_{\chi^2}(H,K) = \sum_i \frac{(h_i - m_i)^2}{m_i}$$

$$m_i = \frac{h_i + k_i}{2}$$

Perceptually-Inspired Measures

- Earth-mover's distance
  - Transforms one object into another by shifting "evidence" in a feature space
  - Compare to L1 metric
- Tversky counting similarity
  - Large set of "predicates" [detectors] is defined [e.g. is the object round?]
  - Similarity increases with increasing number of matching predicates
- Dynamic partial function
  - Large number of features is computed for both objects
  - Compare m smallest feature differences with Minkowski metric

Data-Specific Measures

- Measures defined for binary data

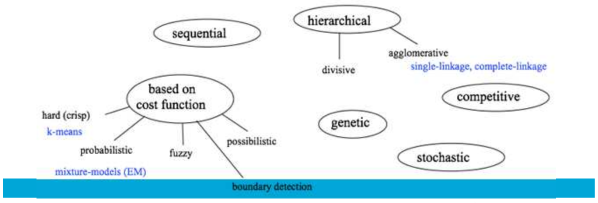
		object s	
		1	0
object r	1	a	b
	0	c	d

Similarity measure	Metric	Euclidean	Similarity	Dissimilarity
Jaccard	Yes	Yes	$S_{rs} = \frac{a}{a+b+c}$	$D_{rs} = \sqrt{1 - S_{rs}}$
Simple matching	Yes	No	$S_{rs} = \frac{a+d}{a+b+c+d}$	$D_{rs} = 1 - S_{rs}$
Yule	No	No	$S_{rs} = \frac{ad-bc}{ad+bc}$	$D_{rs} = 1 - S_{rs}$

- Dissimilarity measures for spectra
  - Spectral angle mapper $D_{SAM}(H,K) = \text{acos}\left(\frac{\langle H,K\rangle}{||H|| \cdot ||K||}\right)$
  - Derivative-based distances
    - Using derivatives of spectra, emphasizing particular shape differences

Clustering Clustering Algorithms

- Very large field, huge number of methods
  - See for example latest Theodoridis and Koutroumbas, Pattern Recognition
    - More than 300 page overview of cluster analysis



k-Means [ISODATA]

- Clustering N observations into m clusters
- Representing clusters by prototypes / concepts
- Dissimilarity : squared Euclidean distance
- Minimize the criterion :

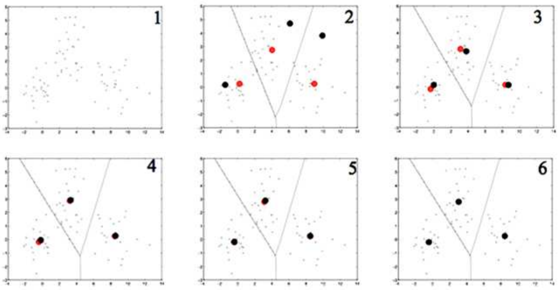
$$J(\theta, U) = \sum_{i=1}^{N}\sum_{j=1}^{m} u_{ij} ||\mathbf{x}_i - \theta_j||^2$$

- Iterative procedure started from random prototypes
- Produces crisp assignment [binary sample weights]

k-Means Algorithm

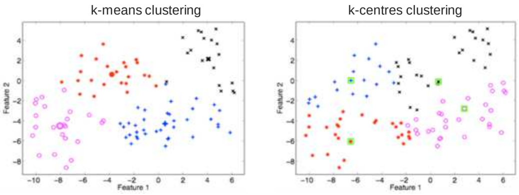
- Input : dataset, desired number of clusters m
- Output : sample labels
- Choose arbitrary initial prototypes $\theta_j, j = 1, \ldots, m$
- Loop :
  - Determine the closest prototype for each observation [label the observations]
  - Compute new prototypes as cluster means
- Repeat the loop until there is no change in prototypes

k-Means, Some Iterations of



k-Centers / k-Medoids

- Minimizes maximum distance within objects in the cluster
- Selects existing objects as prototypes



Probabilistic Mixture Model

- Probabilistic mixture model $p(\mathbf{x}|\Theta) = \sum_{j=1}^{m} u_j p(\mathbf{x}|\theta_j)$
- Mixing proportions $u_j \geq 0, \sum_{j=1}^{m} u_j = 1$
- Often Gaussian mixture is used $\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
  - Probabilistic clustering allows for overlapping clusters
  - Model parameters are usually estimated by maximum-likelihood approach using Expectation-Maximization [EM] algorithm

EM Algorithm

- Expectation step computes an expectation of the likelihood by including the unknown labels as if they were observed

- Maximization step computes maximum likelihood estimates of parameters by maximizing expected likelihood found in the E step

- This process is iterated

EM Algorithm

- E step $\gamma(z_{nk}) = \dfrac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$
- M step

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$$

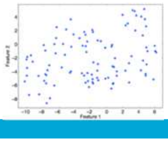
$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right) \left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right)^{\mathrm{T}}$$

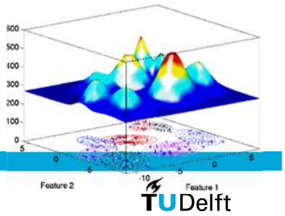
$$\pi_k^{\text{new}} = \frac{N_k}{N}$$

- This process is iterated [but how do we start?]

EM for Mixture Models

- EM clustering
  - Assumes apriori known number of clusters K
  - Guarantees finding of [only] local optimum
  - May converge slowly
  - Is dependent on initialization
- An example :





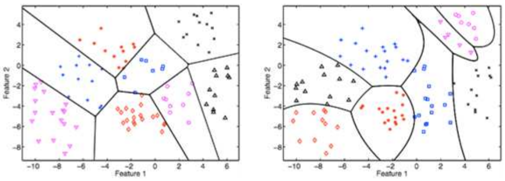
"Generalized" EM Clustering

- Replacing probability model by an arbitrary classifier
- E step : assign each observation x by classifier S to one of the classes
- M step : use the labels to train new classifier S
- Stopping criterion : Labels do not change
- Note that
  - `emclust` provides a final trained classifier which may be applied to new data
  - Move from soft to hard sample assignments
  - Some classifiers allow for soft labels [see dataset labtype]

“Generalized” EM Clustering

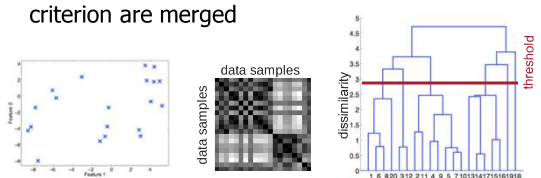
- `nmc` : assuming Gaussian densities with equal covariances [= k-means]

•`qdc` : assuming Gaussian densities with full covariance matrices



Agglomerative Hierarchical Clustering

- Agglomerative algorithms : starting from individual observations, produce a sequence of clusterings of increasing cluster size
- At each level, two clusters chosen by a criterion are merged



2D scatter plot of data dissimilarity matrix dendrogram

Different Combining Rules

- Two nearest objects in the clusters : single linkage

$$g(R,S) = \min_{ij}\{d(\mathbf{x}_i,\mathbf{x}_j) : \mathbf{x}_i \in R, \mathbf{x}_j \in S\}$$

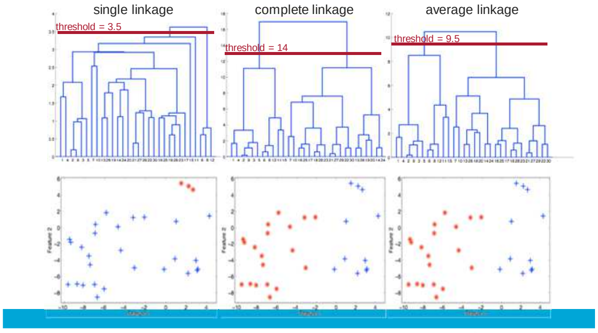
- Two most remote objects in the clusters : complete linkage

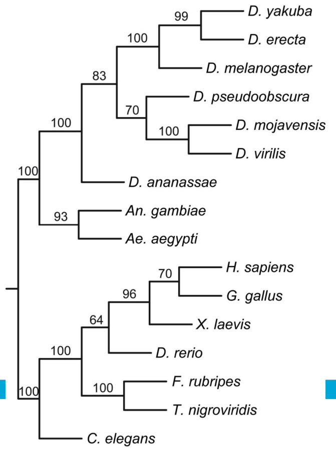
$$g(R,S) = \max_{ij}\{d(\mathbf{x}_i,\mathbf{x}_j) : \mathbf{x}_i \in R, \mathbf{x}_j \in S\}$$

- Cluster centers : average linkage

$$g(R,S) = \frac{1}{|R||S|}\sum_{ij}\{d(\mathbf{x}_i,\mathbf{x}_j) : \mathbf{x}_i \in R, \mathbf{x}_j \in S\}$$

Agglomerative Clustering E.g.





Evaluation of Clustering Validity

- Every clustering algorithm will produce some result, but which one is better?
- Clustering is an ill-posed problem and results should be evaluated

Evaluation Strategies

- Expert judgment : can the identified clusters be interpreted?
- External criterion : if clustering is used to define a set of prototypes for building of a classifier, what is the eventual classification performance?
- Stability : which solution remains unchanged under data perturbation, parameter change or over scales?
- Based on the user-defined "ground-truth" data partitioning [Problematic : if user knows the grouping of data, why not use supervised techniques?]

Number of Clusters?

- Hierarchical clustering : maximum lifetime criterion
  - Problems : noise sensitivity in single linkage
- Based on clustering stability
  - Choose clustering which is the most stable to data perturbation, parameter choice or initialization

Number of Clusters?

- Probabilistic methods : penalized likelihood [– log likelihood + degrees of freedom]
  - AIC, BIC, MDL, etc.
- Validity indices : many methods based on different definitions of cluster compactness and intra-cluster diversity
  - Dunn index, Davies-Bouldin index, SD index, Xie-Beni index, and so on and so forth
  - Problem : often derived on simple artificial problems strongly imposing data structure

Some Conclusions

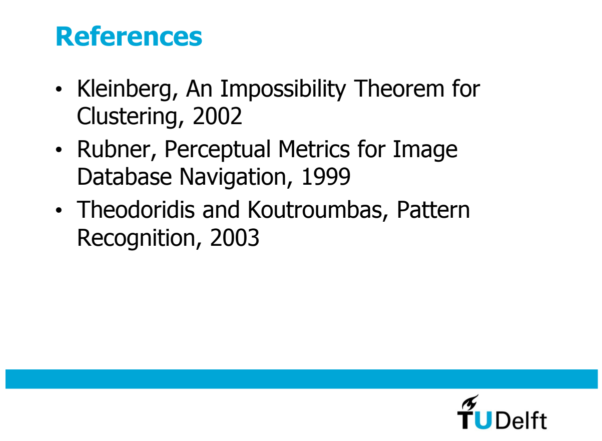
- Many decisions to be made :
  - Measure [dis]similarity between observations
  - What type of structures we search for [blobs, elongated, whatever but stable, …]
  - Choice algorithm parameters [number of clusters, thresholds, scale, …]
  - How to evaluate clustering result? [panel of experts, final classification error, …]
- Clustering is an ill-posed problem
- Axiomatic approach might shed some light

An Impossibility Theorem

- Let f clustering function and S set of objects
- Axioms
  - Scale-Invariance : For any distance function d and any $a > 0$, we have f(d) = f(a d)
  - Richness : Range(f) is equal to the set of all partitions of S
  - Consistency : when we shrink distances between points inside a cluster and expand distances between points in different clusters, we get the same result

An Impossibility Theorem

- Theorem : For every nontrivial set, there is no clustering function f that satisfies scale-invariance, richness, and consistency
  - Implies set of basic trade-offs inherent in clustering problem
  - Possibility to distinguish between clustering methods based on ways to resolve the choices implicit in these trade-off



References

- Kleinberg, An Impossibility Theorem for Clustering, 2002
- Rubner, Perceptual Metrics for Image Database Navigation, 1999
- Theodoridis and Koutroumbas, Pattern Recognition, 2003

TUDelft