STATISTICAL LEARNING SYSTEMS

LECTURES 10 and 11: CLUSTER ANALYSIS (CA)

Jacek Koronacki

Institute of Computer Science, Polish Academy of Sciences

Ph. D. Program 2013/2014

The project is co-financed by the European Union within the framework of European Social Fund

Introduction

Given n p-dimensional data points, i.e. we are given a cloud of n points in p-dimensional space.

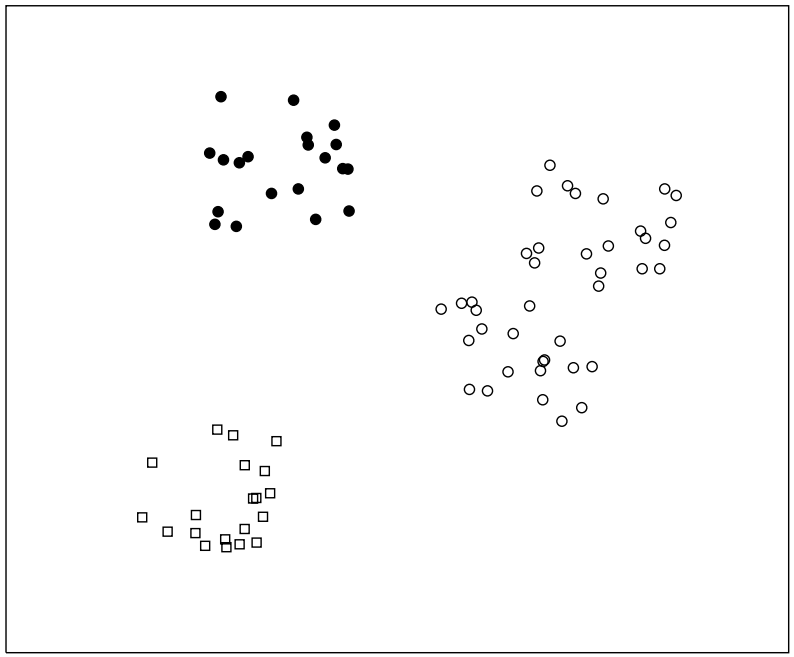
Aim: Divide data points into groups (clusters) such that dissimilarity between points belonging to the same group is on the average smaller than dissimilarity between points belonging to different groups.

Let us first focus on the case when objects are given as points in R^p and dissimilarity is related to Euclidean distance.

Assume that we want to divide the data into K groups with K given (for the time being).

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Introduction



$K = 3$ or $K = 4$ in this case ?
Define

$$C(i) = k \text{ when } \mathbf{x}_i \text{ belongs to } k^{th} \text{ cluster}$$

and $d(\mathbf{x}_i, \mathbf{x}_j)$ - *square* of Euclidean distance between $\mathbf{x}_i$ and $\mathbf{x}_j$

Let

$$T = \frac{1}{2}\sum_{i=1}^{n}\sum_{i'=1}^{n} d_{ii'} \tag{1}$$

and note that

$$T = W + B, \tag{2}$$

where

$$W = \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(i')=k} d(i,i') \tag{3}$$

and

$$B = \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(i')\neq k} d(i,i'). \tag{4}$$

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

It is known that there are (this is one of the Bell numbers)

$$\frac{1}{K!}\sum_{k=1}^{K}(-1)^{K-k}\left(\begin{array}{c} K \\ k \end{array}\right)k^n$$

different groupings of n observations into K groups, which is a forbiddingly large number even for modest n and K ($K << n$)!

PhD STUDIES

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

The project is co-financed by the European Union within the framework of European Social Fund

It is easy to show that

$$W = \sum_{k=1}^{K} \sum_{C(i)=k} d(\mathbf{x}_i, \mathbf{m}_k) n_k, \tag{5}$$

where $\mathbf{m}_k$, $k = 1, \ldots, K$, is the mean of the observations in the k-th cluster,

$$\mathbf{m}_k = \frac{1}{n_k} \sum_{C(i)=k} \mathbf{x}_i, \tag{6}$$

with n_k being the size of the k-th cluster.
Write

$$\tilde{W} = \sum_{k=1}^{K} \sum_{C(i)=k} d(\mathbf{x}_i, \mathbf{m}_k) = \sum_{i=1}^{n} d(\mathbf{x}_i, \mathbf{m}_{C(i)}) \tag{7}$$

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

K-means criterion

Find a partition into K groups such that

$$\tilde{W} = \sum_{i=1}^{n} d(\mathbf{x}_i, \mathbf{m}_{C(i)})$$

is minimal (i.e. we want to minimize the within-groups sum of squares). This is a combinatorial optimization problem but optimization by direct enumeration is usually not feasible.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

K-means algorithm

1. For a given cluster assignment C the total cluster variance

$$\tilde{W} = \sum_{k=1}^{K} \sum_{C(i)=k} d(\mathbf{x}_i, \mathbf{m}_k)$$

is minimized over (m_k), $k = 1, \ldots, K$ yielding the means of currently assigned clusters;

2. Given a current set of means $\{m_1, m_2, \ldots, m_k\}$ for each $\mathbf{x}_i$, find the closest current cluster mean and assign $\mathbf{x}_i$ to this cluster;

3. Steps 1 and 2 are iterated until the assignments do not change.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

K-means algorithm

Note: The versions of the algorihtm differ depending on:

- the moment the centers are modified:
  - batch version: center is modified after the whole batch $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is assigned in 2;
  - sequential version: center is modified each time a new element is assigned.
- initial cluster assignment.

The algorithm converges as $\tilde{W}$ decreases at each step but convergence to local minimum is possible.

Many modifications and generalizations of the K-means algorithm are known, such as K-medoids algorithm, fuzzy K-means, and self-organizing maps (or SOM).

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Choice of the number of clusters

Usually difficult, no universal algorithm exists.

One possible general heuristics:

Calculate

$$\tilde{W}_K = \sum_{k=1}^{K} \sum_{C(i)=k} d(\mathbf{x}_i, \mathbf{m}_k)$$

for different K. As in the case of a scree-plot choose as a true value of K the value K^* for which the plot of W_K against K levels off. (One can use differences $\tilde{W}_K - \tilde{W}_{K+1}$ instead of the $\tilde{W}_K$.)

Tibshirani, Walther and Hastie (2001) have proposed a gap statistics which is based on comparing the $\ln \tilde{W}_K$ for the given data with that for data of the same size but distributed uniformly over a rectangle containing the original data.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

An intro to CA in a feature space and spectral methods

Let ϕ be a projection function into a feature space F, in which the kernel $\mathcal{K}$ computes the inner product

$$\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j).$$

In general, then, we are interested in minimizing in F

$$W = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(j)=k} \|\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)\|^2.$$

(Clearly, the decomposition $T = W + B$ holds in F as well.)

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Kernel cluster analysis (for convenience for only 2 clusters of equal size and data normalized in a feature space):

It is easy to see that what we need is to minimize the cut cost:

$$2\sum_{y_i \neq y_j} \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{i,j=1}^{n} \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) - \sum_{i,j=1}^{n} y_i y_j \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j), \tag{8}$$

$$\text{subject to } \mathbf{y} \in \{-1, +1\}^n, \ \sum_{i=1}^{n} y_i = 0. \tag{9}$$

Note that (8) is equivalent to maximization of the quadratic form

$$\max \mathbf{y}'\mathbf{G}\mathbf{y} \tag{10}$$

w.r.t. $\mathbf{y}$ under constraints (9), where $\mathbf{G}$ is a suitable Gram matrix. Problem (9)-(10), if conveniently relaxed, reduces to maximizing the Rayleigh quotient:

$$\max \frac{\mathbf{y}'\mathbf{G}\mathbf{y}}{\mathbf{y}'\mathbf{y}}. \tag{11}$$

Spectral CA and their graph-theoretical based predecessors

We shall begin with a brief discussion of the algorithm by Ng, Jordan and Weiss (*Proc. Adv. Neural Info. Processing Systems* 2002

and shall later focus our attention on T. Shi's, M. Belkin's and B. Yu's *Data Spectroscopy: Eigenspaces of Convolution Operators and Clustering*, Ann. Statist. 2009.

When discussing the task of minimizing the cut cost one should notice close connection between analyzing an affinity matrix and the corresponding graph Laplacian matrix (actuallly the very term cut cost comes from partitioning a graph).

In fact, already a long time ago it was noticed that graph-theoretical methods are well suited to solving the problem of CA, in particular when detecting clusters with irregular boundaries is of interest (cf. C. Zahn, *Graph-Theoretical Methods for Detecting and Describing Gestalt Clusters*, IEEE Trans. Computer, 1971, and O. Grygorash, Y. Zhou and Z. Jorgensen, *Minimum Spanning Tree based Clustering Algorithm*, Int. Conf. on Tools with AI, 2006).

Graph-theoretical based and information theoretic clustering

An interesting alternative to graph-theoretical arguments are those based on information theory. While E. Gockay and J. C. Principe (by far) were not the first to use entropy measure to cluster data (they used the Renyi entropy; see their *Information Theoretic Clustering*, IEEE Trans. Pattern Anal. and ML, 2002), their paper is one of those published in recent years which influenced further development of nonparametric information theoretic approaches to CA.

During the lecture, we shall discuss briefly L. Faivishevsky's and J. Goldberger's *A Nonparametric Information Theoretic Clustering Algorithm*, Int. Conf. on ML, 2010, and A. C. Müller's, S. Nowozin's and C. H. Lampert's *Information Theoretic Clustering using Minimum Spanning Trees*, 2012.

Theoretically, and not only theoretically, much of the above material comes within Clustering with Bregman Divergences (the title of the paper by A. Banerjee, S. Merugu, I. S. Dhillon and J. Ghosh, *J. ML Research*, 2005).

The algorithm of Ng, Jordan and Weiss

We are given n observations $\mathbf{x}_1, \ldots, \mathbf{x}_n \in R^p$ that we want to cluster into K clusters:

- Form the affinity matrix $\mathbf{A}$ with terms $A_{ij} = \exp(-||\mathbf{x}_i - \mathbf{x}_j||^2/2\sigma^2)$ for $i \neq j$, and zeros when $i = j$.
- Define diagonal matrix $\mathbf{D}$ with $D_{ii} = \sum_{j=1}^{n} A_{ij}$ and construct matrix $\mathbf{M} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$.
- Find K (orthogonal) eigenvectors of $\mathbf{M}$, $\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_K$, which correspond to K largest eigenvalues of $\mathbf{M}$, and form matrix $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_K]$ of dimension $n \times K$.
- Form matrix $\mathbf{Y}$ from $\mathbf{T}$ by renormalizing each of $\mathbf{T}$'s rows to have unit length (i.e. $y_{ij} = t_{ij}/(\sum_j t_{ij}^2)^{1/2}$).
- Treating each row of $\mathbf{Y}$ as a point in R^K, cluster them into K clusters via K-means.
- Assign $\mathbf{x}_i$ to cluster j if and only if row i of the matrix $\mathbf{Y}$ was assigned to cluster j.

Parameter σ^2, which controls how rapidly the affinity A_{ij} falls off with the distance between the observations $\mathbf{x}_i$ and $\mathbf{x}_j$, is chosen automatically.

The algorithm of Ng, Jordan and Weiss

Informal discussion of the "ideal" case (for $K = 3$:

Let $\mathbf{A}$ be the following affinity matrix (with zeros on the main diagonal):

$$\mathbf{A} = \begin{bmatrix} A^{(11)} & 0 & 0 \\ 0 & A^{(22)} & 0 \\ 0 & 0 & A^{(33)} \end{bmatrix}$$

and

$$\mathbf{M} = \begin{bmatrix} M^{(11)} & 0 & 0 \\ 0 & M^{(22)} & 0 \\ 0 & 0 & M^{(33)} \end{bmatrix}$$

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

The algorithm of Ng, Jordan and Weiss

To construct $\mathbf{T}$, we find first $K = 3$ eigenvectors of $\mathbf{M}$ (each of them can be shown to correspond to eigenvalue 1, while all other eigenvalues are strictly less than 1):

$$\mathbf{T} = \begin{bmatrix} \mathbf{t}_1^{(1)} & 0 & 0 \\ 0 & \mathbf{t}_1^{(2)} & 0 \\ 0 & 0 & \mathbf{t}_1^{(3)} \end{bmatrix}$$

After renormalization of each of $\mathbf{T}$'s to have unit length, we obtain:

$$\mathbf{Y} = \begin{bmatrix} \mathbf{1} & 0 & 0 \\ 0 & \mathbf{1} & 0 \\ 0 & 0 & \mathbf{1} \end{bmatrix} \mathbf{R},$$

where $\mathbf{R}$ is an orthogonal 3×3 matrix.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Data spectroscopy or DaSpec by Shi, Belkin and Yu: An informal introduction

We are given n observations $\mathbf{x}_1, \ldots, \mathbf{x}_n \in R^p$ that we want to cluster accordingly. Let the data come from a mixture distribution

$$P = \sum_{k=1}^{K} \pi^{(k)} P^{(k)}, \tag{12}$$

and assume that each mixture component, $P^{(k)}$, $k = 1, \ldots, K$, corresponds to another single cluster.

Consider the following (thought) experiment:

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Data spectroscopy or DaSpec by Shi, Belkin and Yu: An informal introduction

- Our task is to deal with Kn data from (12), where the mixture components $P^{(k)}$, $k = 1, \ldots, K$, are sufficiently well separated and n is large enough to bring results close to those following from theoretical analysis. Let the affinity matrix $\mathbf{G}$ be given by a nonnegative definite kernel $\mathcal{K}(\cdot, \cdot)$ with the tails that have sufficiently fast decay. Generate K data sets from $P^{(k)}$, $k = 1, \ldots, K$, respectively, each of them of size n. Given $\mathcal{K}(\cdot, \cdot)$, construct affinity matrices $\mathbf{G}^{(k)}$, $k = 1, \ldots, K$, for these data sets. For each $\mathbf{G}^{(k)}$, compute the eigenvalues $\lambda_i^{(k)}$, $i = 1, \ldots, n$, from the largest to the smallest, and corresponding eigenvectors.
- It follows that the eigenvectors of $\mathbf{G}$, corresponding to the eigenvalues of $\mathbf{G}$ ordered from the largest to the smallest, are approximately equal to the eigenvectors of the $\mathbf{G}^{(k)}$, $k = 1, \ldots, K$, ordered according to the mixture magnitudes $\pi^{(k)} \lambda_i^{(k)}$.
- For each (k-th) mixture component, the eigenvector which corresponds to the largest eigenvalue for this component ($\lambda_1^{(k)}$) is the only eigenvector that has no sign changes (up to some precision).

DaSpec Algorithm

- Construct the Gaussian kernel matrix $\mathbf{G}$ with terms $G_{ij} = \frac{1}{n}\exp(-||\mathbf{x}_i - \mathbf{x}_j||^2/2\sigma^2)$ and compute its eigenvalues $\lambda_1, \ldots \lambda_n$ and eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$.
- Identify all eigenvectors $\mathbf{v}_j$ that have no sign changes up to precision ε_j; estimate the number of clusters, K, as equal to the number of such eigenvectors; denote these eigenvectors by $\mathbf{v}_0^{(1)}, \ldots, \mathbf{v}_0^{(K)}$.
- Assign a cluster label to each data point $\mathbf{x}_i$, $i = 1, \ldots, n$, as

$$L(\mathbf{x}_i) = \arg\max_k \{|v_{0i}^{(k)}| : k = 1, \ldots, K\},$$

where $v_{0i}^{(k)}$ denotes i-th coordinate of $\mathbf{v}_0^{(k)}$.

The authors discuss some heuristics for choosing proper values of the algorithm's parameters.

A nonparametric information theoretic CA: An informal introduction

Faivishevsky and Goldberger:
We are given n observations $\mathbf{x}_1, \ldots, \mathbf{x}_n \in R^p$ that we want to cluster into K clusters; i.e., we want to find a clustering function $C : \mathbf{X} \rightarrow \{1, 2, \ldots, K\}$. Actually, then, our task is to maximize (w.r.t. C) mutual information

$$I(\mathbf{X}; C) = H(\mathbf{X}) - H(\mathbf{X}|C)$$

or, equivalently, to minimize

$$H(\mathbf{X}|C).$$

Note that (with obvious notations)

$$H(\mathbf{X}|C) = \sum_{j=1}^{K} \frac{n_j}{n} H(\mathbf{X}|C = j). \tag{13}$$

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

A nonparametric information theoretic CA: An informal introduction

One can start with a k-NN estimator of $H(\mathbf{X}|C)$, but they propose to use mean NN estimator

$$H_{\text{mean}} = \frac{1}{n-1} \sum_{k=1}^{n-1} H_k$$

where (after Kozachenko and Leonenko (1987))

$$H_k = \frac{p}{n} \sum_{i=1}^{n} \log \varepsilon_{ik} + \text{const}(k)$$

and ε_{ik} is the Euclidean distance from $\mathbf{x}_i$ to its k-th NN. Accordingly,

$$H_{\text{mean}} = \frac{p}{n(n-1)} \sum_{i \neq \ell} \log \parallel x_i - x_\ell \parallel + \text{const}$$

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

A nonparametric information theoretic CA: An informal introduction

Thus, for (13) we get (with obvious notations)

$$H(\mathbf{X}|C=j) \approx \frac{p}{n_j(n_j-1)} \sum_{i \neq \ell | c_i = c_\ell = j} \log \parallel x_i - x_\ell \parallel$$

and

$$H(\mathbf{X}|C) \approx \sum_j \frac{p}{(n_j-1)} \sum_{i \neq \ell | c_i = c_\ell = j} \log \parallel x_i - x_\ell \parallel .$$

Müller, Nowozin and Lampert, whose method is based on Rényi entropy,

$$\frac{1}{1-\nu} \log \int f^\nu(\mathbf{x}) d(\mathbf{x}), \quad \nu \in (0,1),$$

claim to obtain comparable or better results, while they can use a simpler estimate of this entropy.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Another approaches to CA - hierarchical methods

We skip discussion of density-based CA (cf. lectures by S. Jaroszewicz) and turn to hierarchical methods.

It is more than worth noting that already in 1969 J. C. Gower and G. J. S. Ross published a paper in which they provided an algorithm to build a hierarchical classifier (based on single linkage dissimilarity, a type of dissimilarity to be discussed) using minimum spanning trees. The paper's title: *Minimum Spanning Trees and Single Linkage Cluster Analysis*, Applied Statistics.

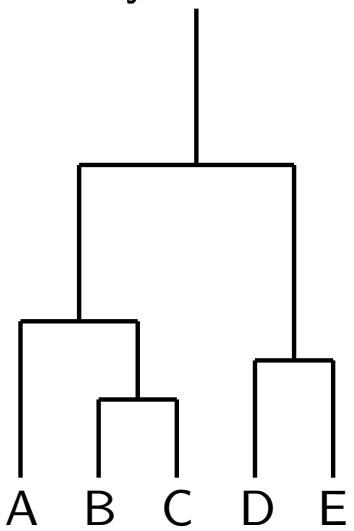
PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Another approaches to CA - hierarchical methods

We skip discussion of density-based CA (cf. lectures by S. Jaroszewicz) and turn to hierarchical methods:

- agglomerative methods
- divisive methods

An agglomerative method begins with n subclusters, each containing one data point and at each step merges the two most similar groups to form a new cluster. The algorithm proceeds until forming a single cluster. This is usually visualised in terms of a dendrogram.



The merging process is:

$$A, B, C, D, E \rightarrow A, \{B, C\}, D, E$$
$$A, \{B, C\}, D, E \rightarrow A, \{B, C\}, \{D, E\}$$
$$A, \{B, C\}, \{D, E\} \rightarrow \{A, B, C\}, \{D, E\}$$
$$\{A, B, C\}, \{D, E\} \rightarrow \{A, B, C, D, E\}$$

Hierarchical methods

A divisive method starts from a single cluster and uses division instead of merging.
Each method requires definition of clusters' dissimilarity which is based on dissimilarities between members of the clusters.
Clusters' dissimilarities:
Consider two clusters i and j. Let their dissimilarity be denoted by D_{ij}.
Single-linkage dissimilarity

$$D_{ij} = \min d_{kk'},$$

where k ranges over cluster i and k' ranges over cluster j. It is also called closest nearest neighbor.
Complete-linkage dissimilarity

$$D_{ij} = \max d_{kk'},$$

where k ranges over cluster i and k' ranges over cluster j. It is also called furthest nearest neighbor.

Group-average linkage dissimilarity

$$D_{ij} = \frac{1}{n_i n_j} \sum d_{kk'},$$

where k ranges over cluster i and k' ranges over cluster j and n_i is the number of observations in cluster i.

It is possible to calculate dissimilarity between merged cluster i and j and cluster k using D_{ik} and D_{jk}. E.g., in the case of single-linkage algorithm

$$D_{k.ij} = \min(D_{ki}, D_{kj}).$$

For a fixed number of clusters K we stop hierarchical algorithm when exactly K clusters are obtained.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

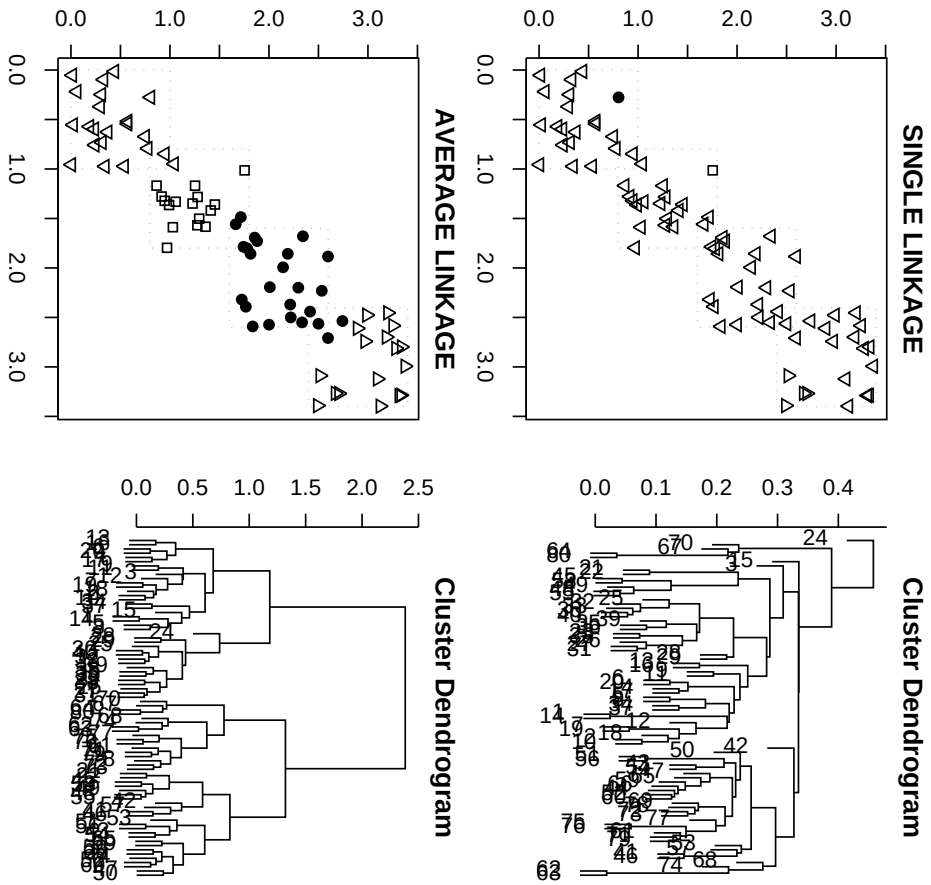
Hierarchical methods

Main characteristics of various methods:

- single linkage: tends to produce "long" groups with large diameters (effect of chaining)
- complete-linkage: *relatively* compact clusters *relatively* far apart
- average-linkage: usually compromise between these two methods

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Hierarchical methods



$K = 4$ was chosen.

Hierarchical Clustering via Joint Between-Within Distance

Such methods have been of interest to statisticians for decades now. Recently, their valuable version has been proposed by G.J. Székely and M.L. Rizzo. The method can be recommended as applicable in general and useful in particular when standard dendrograms fail to give satisfactory results.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Hierarchical Clustering via Joint Between-Within Distance

Let $A = \{\mathbf{a}_1, \ldots, \mathbf{a}_{n_1}\}$ and $B = \{\mathbf{b}_1, \ldots, \mathbf{b}_{n_2}\}$ be two sets in R^p and let

$$e(A,B) = \frac{n_1 n_2}{n_1 + n_2}\Big(\frac{2}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\|\mathbf{a}_i - \mathbf{b}_j\| \tag{14}$$

$$-\frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\|\mathbf{a}_i - \mathbf{a}_j\| - \frac{1}{n_2^2}\sum_{i=1}^{n_2}\sum_{j=1}^{n_2}\|\mathbf{b}_i - \mathbf{b}_j\| \tag{15}$$

Theorem. Suppose $\mathbf{X}, \mathbf{X}' \in \mathbf{R^p}$ are i.i.d. random vectors with distribution F, $\mathbf{Y}, \mathbf{Y}' \in \mathbf{R^p}$ are i.i.d. random vectors with distribution G, independent of $\mathbf{X}, \mathbf{X}'$. Suppose $E||\mathbf{X}|| < \infty$ and $E||\mathbf{Y}|| < \infty$. Then

$$2E\|\mathbf{X} - \mathbf{Y}\| - E\|\mathbf{X} - \mathbf{X}'\| - E\|\mathbf{Y} - \mathbf{Y}'\| \geqslant 0,$$

and equality holds if and only if $F = G$.

Hierarchical Clustering via Joint Between-Within Distance

Corollary. For all finite nonempty sets $A, B \in R^p$, $e(A, B) \geqslant 0$ and equality holds if and only if $A = B$.

The Authors have developed a hierarchical algorithm that merges the pair of clusters with minimum e-distance at each level.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Clustering on subsets of attributes - COSA (by Friedman and Meulman)

Let

$$W(C) = \sum_{k=1}^{K} \frac{W_k}{n_k^2} \sum_{C(i)=k} \sum_{C(i')=k} d(i, i'). \tag{16}$$

(In particular, by setting $W_k = n_k^2$ we assign the same weight to all pairs of objects, what amounts to aiming at clusters of possibly similar sizes.) More generally, let

$$W(C, \{\mathbf{w}_k\}_1^K) = \sum_{k=1}^{K} \frac{W_k}{n_k^2} \sum_{C(i)=k} \sum_{C(i')=k} \left(\sum_{j=1}^{p} w_{j,k} d(i, i')_j + \lambda w_{j,k} \log w_{j,k} \right), \tag{17}$$

where $\lambda \geqslant 0$, $d(i, i')_j$ is the squared distance on j-th attribute for objects i and i', $\mathbf{w}_k = \{w_{j,k}\}_{j=1}^p$, $k = 1, \ldots, K$,

$$\{w_{j,k} \geqslant 0\}_{j=1}^p, \quad \sum_{j=1}^{p} w_{j,k} = 1, \quad k = 1, \ldots, K.$$

Now, (9) is minimized wrt clusters C and weights $\{\mathbf{w}_k\}_1^K$.

COSA as a preliminary step to hierarchical clustering

Let us replace minimization of

$$W(C,\{\mathbf{w}_k\}_1^K)=\sum_{k=1}^{K}\frac{W_k}{n_k^2}\sum_{C(i)=k}\sum_{C(i')=k}\left(\sum_{j=1}^{p}w_{j,k}d(i,i')_j+\lambda w_{j,k}\log w_{j,k}\right), \tag{18}$$

by that of

$$W(\mathbf{W})=\sum_{i=1}^{n}\frac{1}{K}\sum_{i'\in KNN(i)}\left(\sum_{j=1}^{p}w_{j,i}d(i,i')_j+\lambda\sum_{j=1}^{p}w_{j,i}\log w_{j,i}\right), \tag{19}$$

where $KNN(i)$ denotes K nearest neighbors of object i and $\mathbf{W}$ is a $p\times n$ matrix. (K is chosen experimentally, say, $K\approx\sqrt{n}$.) In this way, the following distances are defined

$$D(i,i')=\sum_{j=1}^{p}w_{j,i}d(i,i').$$

Clusterability assessment

A few comments on the subject will be given during the lecture.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund