Week 7 - Unsupervised learning 2.

Clustering

Anikó Lovik

a.lovik@fsw.leidenuniv.nl

Statistical learning
2024-03-20

Topics for Week 7

1. Clustering
2. K-means clustering
3. Hierarchical clustering

ISLR2 book:
Clustering: chapter 12, section 12.4 (pp. 516-532)
+ R lab: section 12.5.3 (pp. 538-547)

Different types of unsupervised learning methods

- **Dimension reduction techniques**
  - Principal Component Analysis (PCA)
  - Exploratory Factor Analysis (EFA)
  - Correspondence analysis
  - Canonical Correlation Analysis (CCA)
  - Independent Component Analysis (ICA)
  - Non-negative Matrix Factorisation (NMF)
  - t-distributed Stochastic Neighbor Embedding (t-SNE)
  - Multi-dimensional Scaling (MDS)
- **Clustering techniques**
  - One-mode clustering
    - K-means clustering
    - Hierarchical clustering
    - Gaussian mixture analysis
    - Latent class analysis
  - Two-mode clustering (bi-clustering)

Overview of cluster analysis

- Aim: partition observations into homogeneous subgroups (clusters) based on similarity
- Many different types:
  - Centroid based: e. g., k-means clustering
  - Connectivity based: e.g., hierarchical clustering
  - Distribution based: e.g., Gaussian mixture clustering (Week 8)

  Variations:
  - Each observation belongs to exactly one cluster or can belong to multiple clusters (fuzzy clustering)
  - Some observations do not belong to any cluster

K-means clustering

- group similar objects such that within-cluster variation (WCV of $W(C_k)$ for any cluster C_k) is minimal:

$$\underset{C_1,\ldots,C_K}{\text{minimise}} \left\{ \sum_{k=1}^{K} W(C_k) \right\} \tag{1}$$

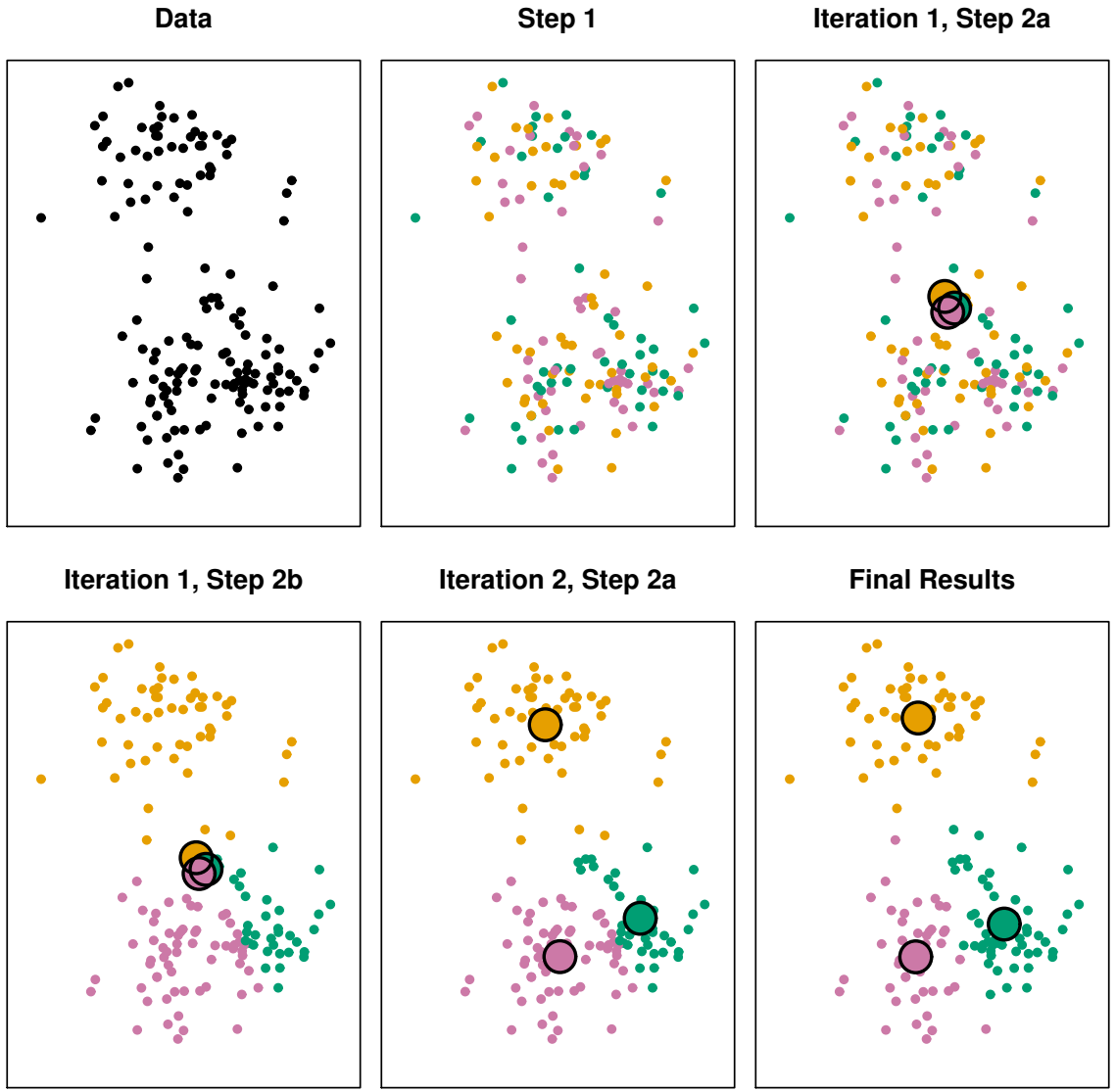
- Using squared Euclidean distances for n observations, p variables and K clusters:

$$\underset{C_1,\ldots,C_K}{\text{minimise}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\} \tag{2}$$

K-means clustering algorithm

Algorithm

1. Randomly assign a number, from 1 to K to each of the observations. These serve as initial cluster assignments for the observations.
2. Iterate until the cluster assignments stop changing:
    1. For each of the K clusters, compute the cluster centroid. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.
    2. Assign each observation to the cluster whose centroid is closest (where closest is defined using Euclidean distance).



Source: Page 389, Introduction to Statistical Learning by James, Witten, Hastie and Tibshirani

Number of clusters

How to determine K:

No agreed upon method, but there are several options, including:

1. Scree plot
   - K against total WCV
   - K against percentage explained $\rightarrow$ ratio of (total variance in data total WCV) to (total variance in data)
2. Calinski and Harabasz's method (pseudo-F): maximise $CH(c) = \frac{B(c)/(c-1)}{W(c)/(n-c)}$ where $B(c)$, is the sum of squares between the clusters and $W(c)$ is the sum of squares within the clusters
3. Hartigan's method: $HAR(c) = \left[\frac{W(c)}{W(c+1)} - 1\right]/(n-c-1)$

   Start with c =1 and add a cluster if $HAR(c) > 10$
4. Silhouette plot

Silhouette

For data point i in cluster C_i, calculate the mean distance within the cluster:

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j) \tag{3}$$

and the mean distance to the nearest other cluster C_j (the cluster with the smallest avarage distance):

$$b(i) = \min \frac{1}{|C_j|} \sum_{j \in C_j} d(i, j) \tag{4}$$

Now define te silhouette score of point i as

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{5}$$

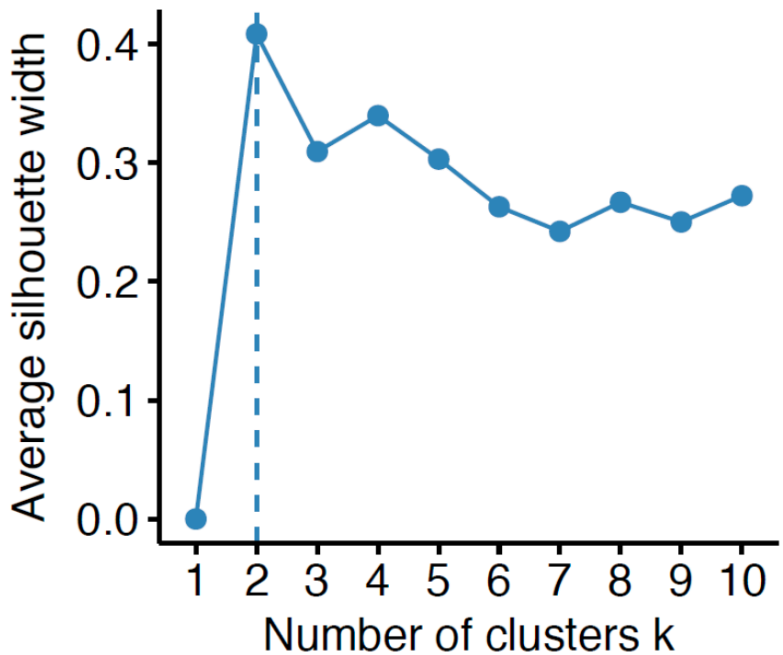
if $|C_i| > 1$ and

$$s(i) = 0 \tag{6}$$

if $|C_i| = 1$

Silhouette plot

1. Perform K-means clustering for different values of k (e.g., 1 to 10)
2. For each k, calculate the average silhouette (AS) of observations
3. Plot AS according to the number of clusters k.
4. Choose k with the maximum AS.



Issues

- Problem of local optima (non-convex optimisation problem): always use a multi-start procedure
- Normalisation (and/or centering) of the variables
- Robustness. Try a subset of the data and see whether you obtain the same clusters
- Outliers. Use median instead of mean for centroid (k-median clustering)

Selecting best solution after multi-start procedure

Source: Page 521, Introduction to Statistical Learning by James, Witten, Hastie and Tibshirani

Strengths and limitations of K-means clustering

Strengths:

- Simple and runs fast on large datasets
- Efficient: Time complexity: O(tkn)
  where n is the number of observations, k is the number of clusters, and t is the number of iterations.
- Since both k and t are small $\rightarrow$ k-means is considered a linear algorithm

Limitations:

- Difficulty handling clusters are of differing sizes
- Non-globular shapes
- Applicable only to numerical data
- All of the issues already mentioned

Hierarchical clustering

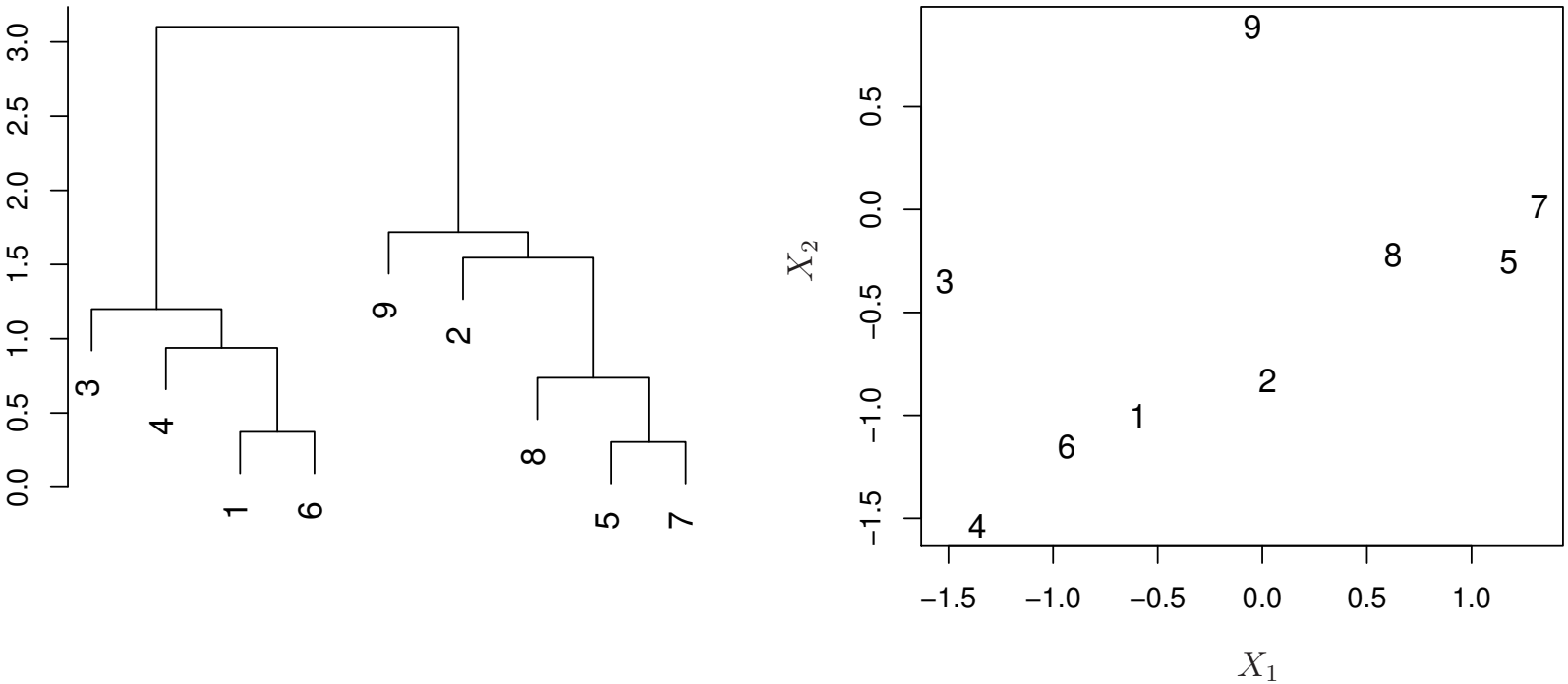
- Objects that belong to a child cluster also belong to the parent cluster (see dendrogram)
- Does not require to set the number of clusters in advance
- Agglomerative (bottom-up) or divisive (top-down)

Hierarchical agglomerative clustering algorithm

Algorithm

1. Begin with n observations and a (distance/(dis)similarity) measure (e.g., Euclidean distance) of all pairwise dissimilarities. Treat each observation as its own cluster.
2. For $i = n, n-1, ..., 2$:
   1. Examine all pairwise inter-cluster dissimilarities among the i clusters and identify the pair of clusters that are the least dissimilar (= the most similar). Fuse these two clusters. The dissimilariy of these two clusters indicates the height in the dendrogram at which the fusion should be placed.
   2. Compute the new pairwise inter-cluster dissimilarities among the $i-1$ remaining clusters

Example of hierarchical clustering



Source: Page 524, Introduction to Statistical Learning by James, Witten, Hastie and Tibshirani

Distance/(dis)similarity measures

- Euclidean distance $||a - b||_2 = \sqrt{\sum_i (a_i - b_i)^2}$
- Manhattan distance $||a - b||_1 = \sum_i |a_i - b_i|$
- Minkowski distance $||a - b||_p = (\sum_i (a_i - b_i)^p)^{1/p}$
- Maximum $\max_i |a_i - b_i|$
- Correlation-based distance: two observations are similar if their features are highly correlated (focus on shape)
- Any other measure of distance or dissimilarity (e.g., city-block, Chebysev, Mahalanobis, etc.)

The choice of the dissimilarity measure will impact the result

Distance/(dis)similarity measures

”An appropriate dissimilarity measure is far more important in obtaining success with clustering than choice of clustering algorithm. This aspect of the problem ... depends on domain specific knowledge and is less amenable to general research.”

Page 506, Elements of Statistical Learning by Hastie, Tibshirani and Friedman

Linkage types

Complete - - **Maximal inter-cluster dissimilarity**
Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the largest of these dissimilarities.

Single - - **Minimal inter-cluster dissimilarity**.
Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the smallest of these dissimilarities.

Average - - **Mean inter-cluster dissimilarity**
Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the average of these dissimilarities.

Centroid
Dissimilarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B. Centroid linkage can result in undesirable inversions.

Hierarchical clustering with different linkages

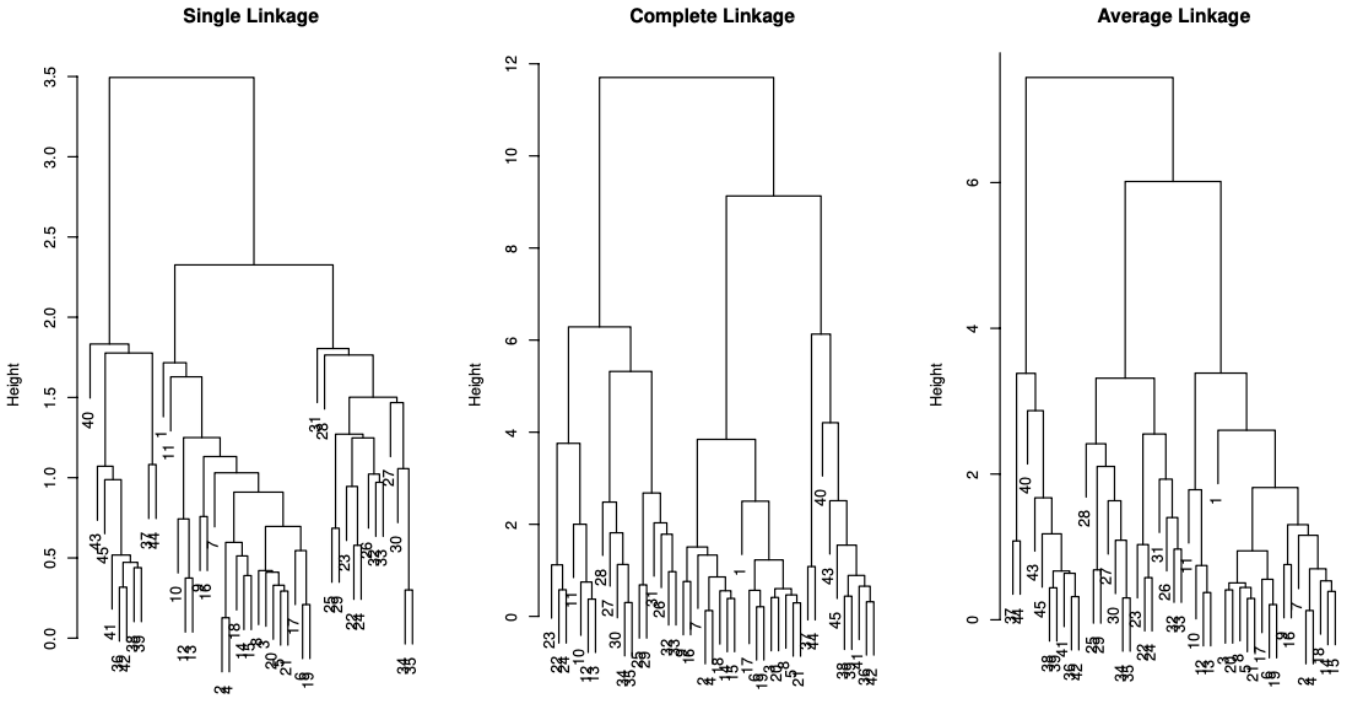


Figure 21.2 Hierarchical clustering of `pottery` data and resulting dendrograms.

Source: Page 5, A Handbook of Statistical Analyses Using R — 3rd Edition by Hothorn and Everitt

Strengths and limitations of hierarchical clustering

Strengths:

- Flexible
- Logical and easy to interpret
- No need to make assumptions about K

Limitations:

- Slow: Time complexity: $O(n^3)$
- Relatively unstable and unreliable: first combination of observations, which may be based on a small difference in the criterion, will constrain the rest of the analysis

Combination approach

A combination approach using a hierarchical approach followed by a non-hierarchical approach is often applied to combine the strengths and overcome the limitations

Process:

- First, hierarchical clustering is used to select the number of clusters and profile cluster centers that serve as initial cluster seeds in the non-hierarchical procedure
- Second, a centroid-based method can be employed to cluster all observations to provide more accurate cluster memberships

This way, the advantages of hierarchical methods are complemented by the ability of switching of cluster membership

Group Exercise

Please form groups of 2-4 students.
Look at the paper on Brightspace.

Discuss the following questions with your group:

- How does clustering fit into addressing the aim of the study?
- How was clustering used in this study?
- What information has been given about the clustering method in the paper?
- How are the (clustering) results presented in the study?
- Do you think the clustering method in the paper was applied correctly? Why (not)?

Preparation for workgroup on Friday

Please read in ISLR2 book:

- Clustering: chapter 12, section 12.4 (pp. 516-532) + relevant R lab (second part of sections 12.5)
- R Lab is also available as html file on Brightspace $\rightarrow$ look under "Between Lecture + Workgroup"

There will be no lecture or workgroup next week!
Next lecture willl be on Wednesday 3 April 2024.

You may submit the weekly assignment for this week until Tuesday 2 April.