Анализ данных с использованием языка программирования R

Тема 6
Модели кластеризации данных

Минюкович Екатерина Александровна
к.э.н., доцент

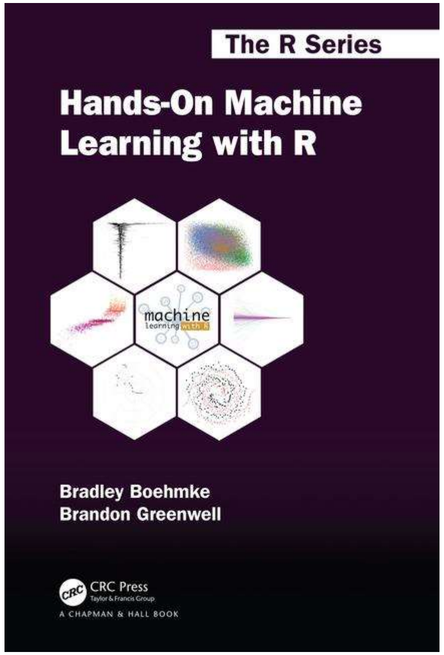
miniukovich@bsu.by

Reference

Hands-On Machine Learning with R

Bradley Boehmke & Brandon Greenwell

https://bradleyboehmke.github.io/HOML/index.html



Reference

Introduction to Machine Learning with R
by Dr. Dimitrios Gouliermis

http://www.mpia.de/homes/dgoulier/MLClasses/Course%20-%20Introduction%20to%20Machine%20Learning%20for%20Scientists%20with%20R.html

Reference

DataCamp

INTERACTIVE COURSE

Introduction to Machine Learning

https://learn.datacamp.com/courses/introduction-to-machine-learning-with-r

Supervised vs. Unsupervised Learning

Supervised	Unsupervised
Data: 1) n observations; 2) p variables X1, X2, . . .,Xp, measured on each observation; 3) response Y measured on same n observations	Data: 1) n observations; 2) p variables X1, X2, . . .,Xp, measured on each observation
Y	Clustering…
Continuous Regression Discrete Classification	

Clustering, what?

- **Cluster**: collection of objects
  - *Similar* within cluster
  - *Dissimilar* between clusters
- **Clustering**: grouping objects in clusters
  - No labels: *unsupervised* classification
  - Plenty possible clusterings

Clustering, why?

- Pattern Analysis
- Visualise Data
- pre-Processing Step
- Outlier Detection
- ...

- Targeted Marketing Programs
- Student Segmentations
- Data Mining
- ...

Clustering methods

- k-means
- Hierarchical *(many variations)*

Compactness and Separation

- **Within Cluster Sums of Squares (WSS):**

$$\mathrm{WSS} = \sum_{i=1}^{N_C} \sum_{x \in C_i} d(\mathbf{x}, \bar{\mathbf{x}}_{C_i})^2$$

$\bar{\mathbf{x}}_{C_i}$	Cluster Centroid
$\mathbf{x}$	Object
C_i	Cluster
N_C	#Clusters

Measure of compactness ← **Minimise WSS**

- **Between Cluster Sums of Squares (BSS):**

$$\mathrm{BSS} = \sum_{i=1}^{N_C} |C_i| \cdot d(\bar{\mathbf{x}}_{C_i}, \bar{\mathbf{x}})^2$$

$\bar{\mathbf{x}}_{C_i}$	Cluster Centroid
N_C	#Clusters
$\lvert C_i \rvert$	#Objects in Cluster
$\bar{\mathbf{x}}$	Sample Mean

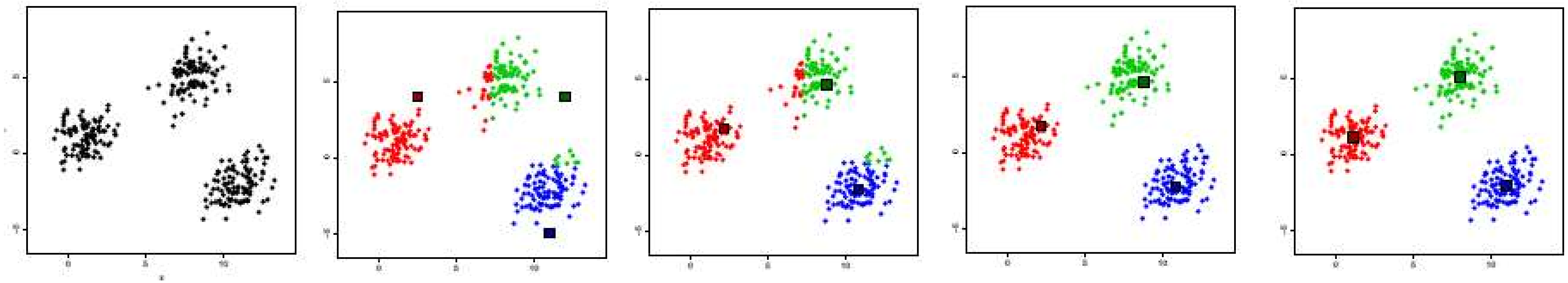
Measure of separation ← **Maximise BSS**

K-means algorithm

Goal: Partition data in k *disjoint* subsets

1. Randomly assign k *centroids*
2. Assign data to *closest* centroid
3. Moves centroids to *average* location
4. Repeat step 2 and 3

Let's take k = 3



Choosing k

- Goal: Find k that minimizes WSS
- Problem: WSS keeps decreasing as k increases!
- Solution: WSS starts decreasing slowly

$$\left.\begin{array}{r}\text{WSS starts decreasing slowly}\\ \text{WSS / TSS} < 0.2\end{array}\right\}\ \text{Fix k}$$

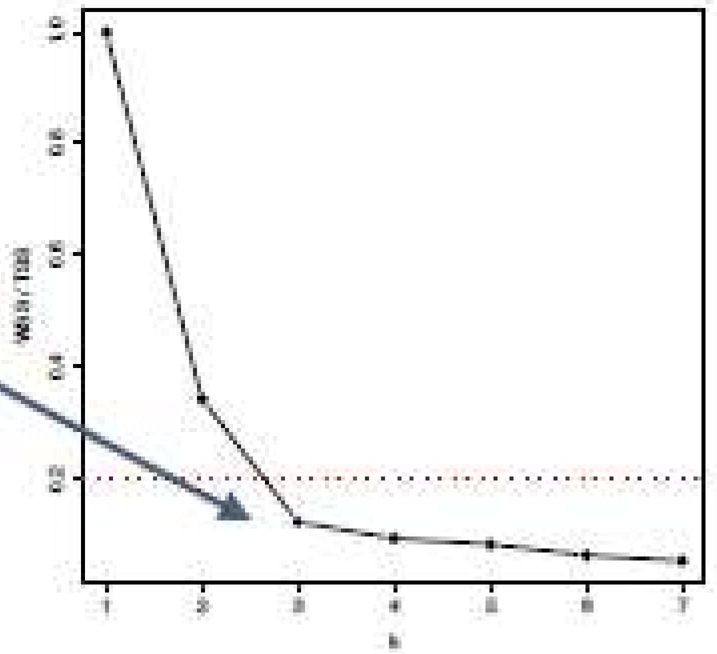
$$TSS = WSS + BSS$$

Choosing k: Scree Plot

Scree Plot: Visualizing the ratio WSS / TSS as function of k

Look for the *elbow* in the plot

Choose k = 3



K-means in R

```
> my_km <- kmeans(data, centers, nstart)
```

- **centers:** Starting centroid or #clusters
- **nstart:** #times R restarts with different centroids

Distance: Euclidean metric

```
> my_km$tot.withinss   ←   WSS
> my_km$betweenss      ←   BSS
```

Cluster evaluation

Not trivial! There is no truth

- No true labels
- No true response

Evaluation methods? Depends on the goal

Goal: Compact and Separated ⟵ Measurable!

Cluster measures

WSS and BSS: Good indication

Underlying idea:

- Variance within clusters
- Separation between clusters

Compare

Alternative:

- Diameter
- Intercluster Distance

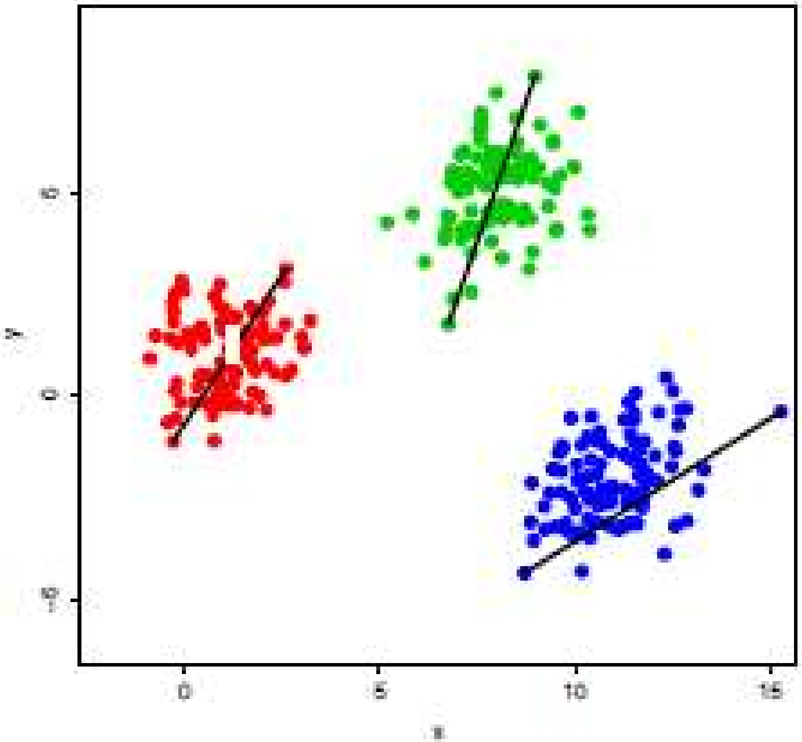
Diameter

$$\mathrm{Dia}_i = \max_{x,y \in C_i} d(x, y)$$

x, y : Objects

C_i : Cluster

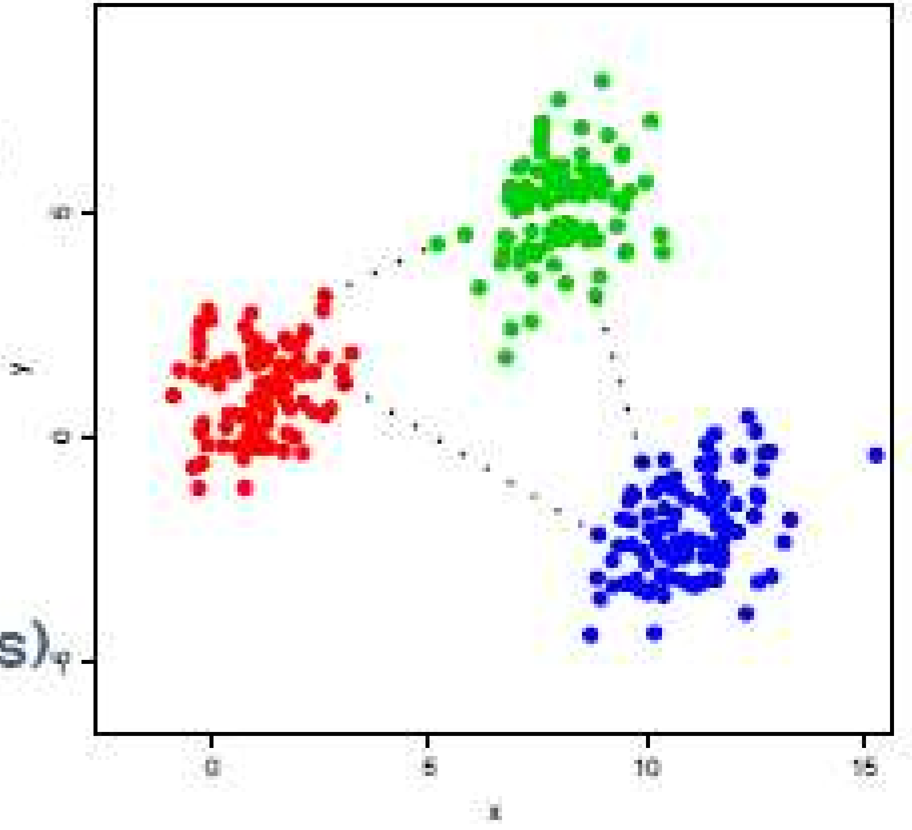
d : Distance (objects)



Measure of Compactness

Intercluster distance

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



x, y : Objects

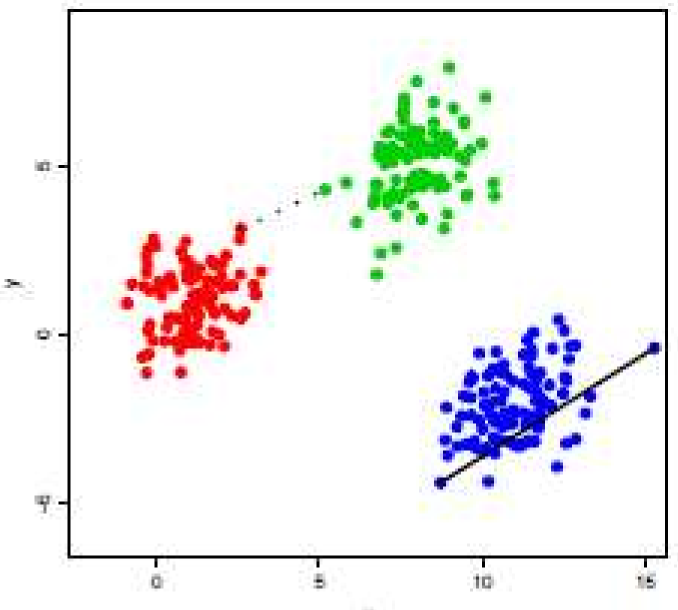
C_i, C_j : Clusters

d : Distance (objects)

Measure of Separation

Dunn’s index

$$\frac{\min\limits_{1 \leq i < j \leq k} \delta(C_i, C_j)}{\max\limits_{1 \leq m \leq k} \mathrm{Dia}_m}$$



Higher Dunn ⟶ Better separated / more compact

Notes:

- High computational cost
- Worst case indicator

Evaluating in R

Libraries: cluster and clValid

Dunn's Index:

```
> dunn(clusters = my_km, Data = ...)
```

- clusters: cluster partitioning vector
- Data: original dataset

Scale issues

Metrics are often scale dependent!

Which pair is most similar? (Age, Income, IQ)

- $X_1 = (28, 72000, 120)$
- $X_2 = (56, 73000, 80)$
- $X_3 = (29, 74500, 118)$

→

- Intuition: (X_1, X_3)
- Euclidean: (X_1, X_2)

Solution: Rescale income / 1000$

Standardizing

Problem: Multiple variables on different scales

Solution: Standardize your data

1. Subtract the mean
2. Divide by the standard deviation

```
> scale(data)
```

Note: Standardizing → Different interpretation

Hierarchical clustering

Hierarchy:

- Which objects cluster first?
- Which cluster pairs merge? When?

Bottom-up:

- Starts from the objects
- Builds a hierarchy of clusters

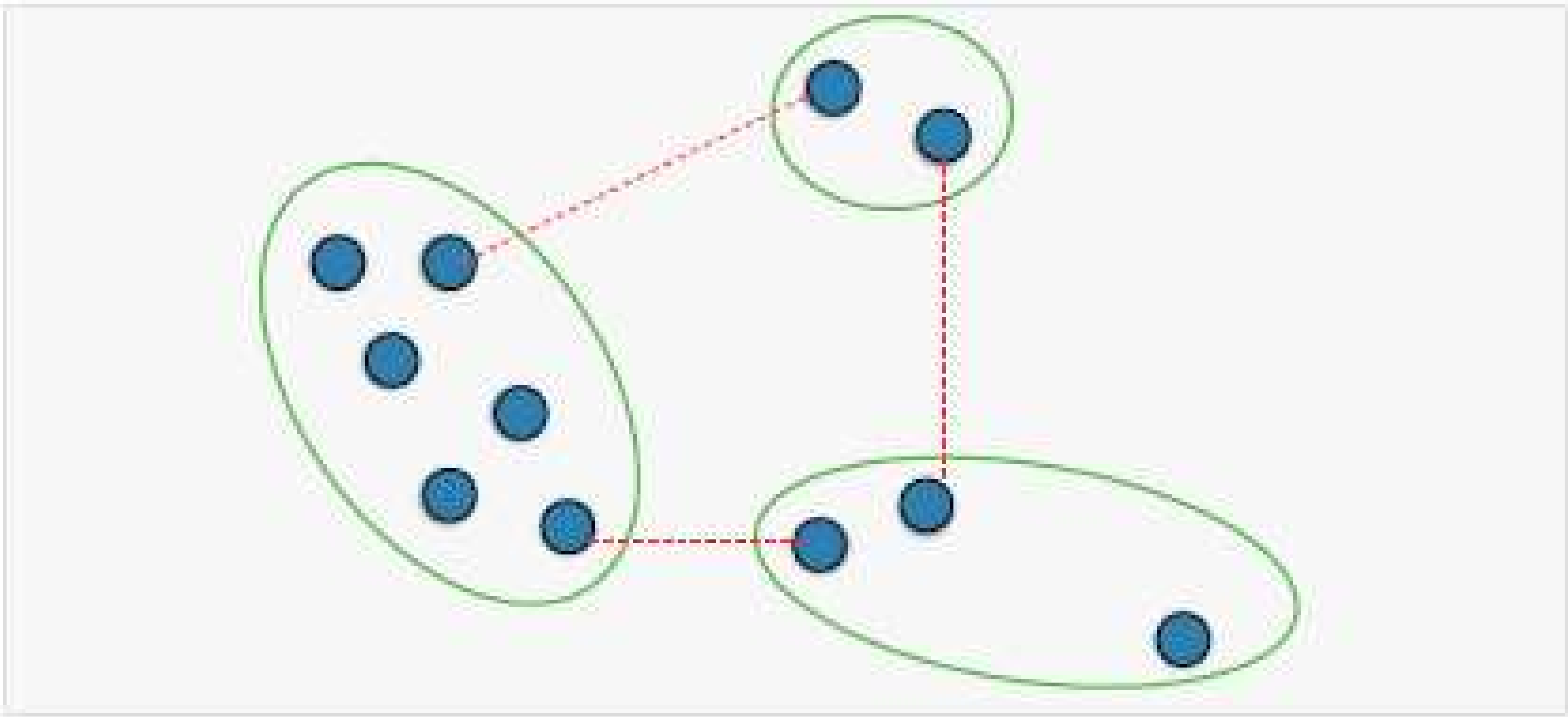
Linkage - methods

- Simple-Linkage: *minimal* distance between clusters
- Complete-Linkage: *maximal* distance between clusters
- Average-Linkage: *average* distance between clusters

→ Different Clusterings

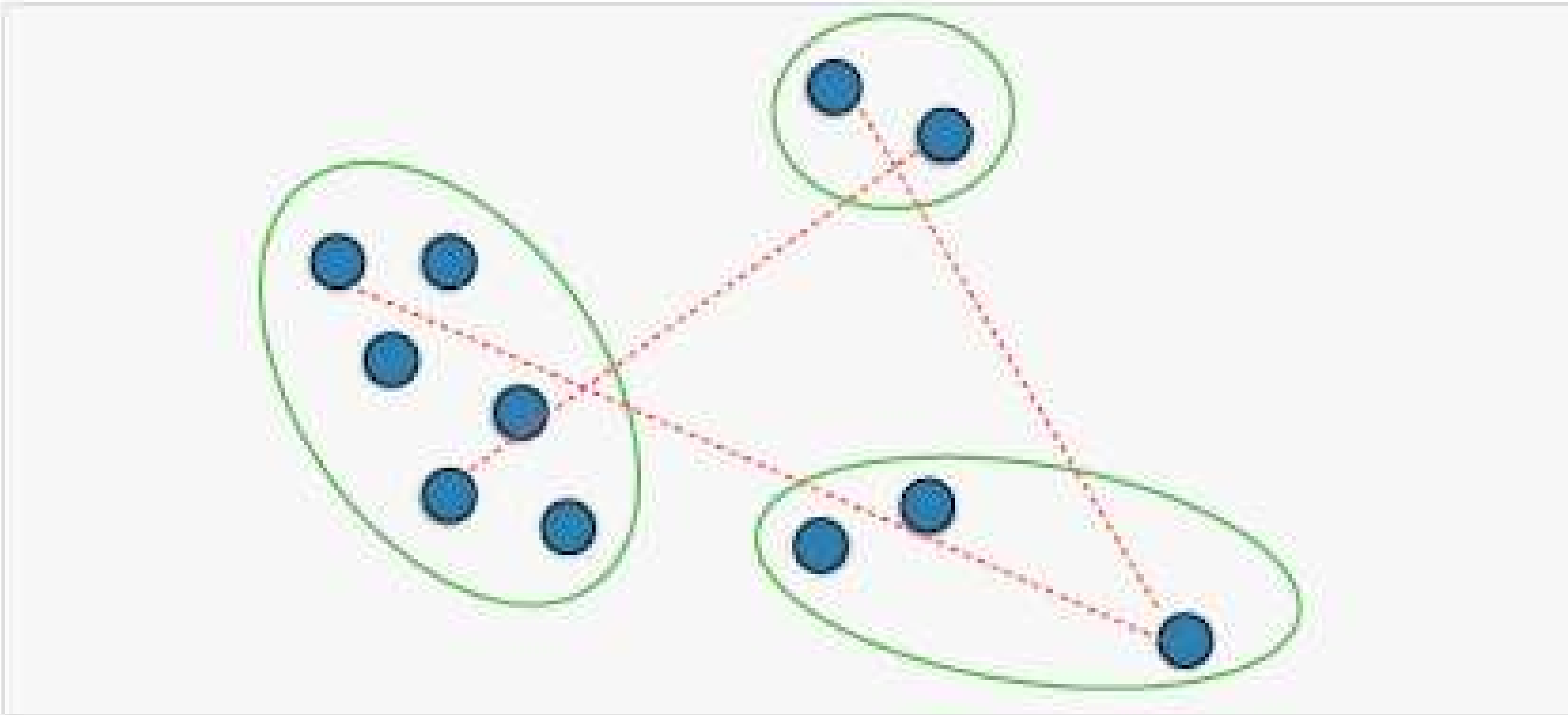
Simple - linkage

Minimal distance between objects in each clusters

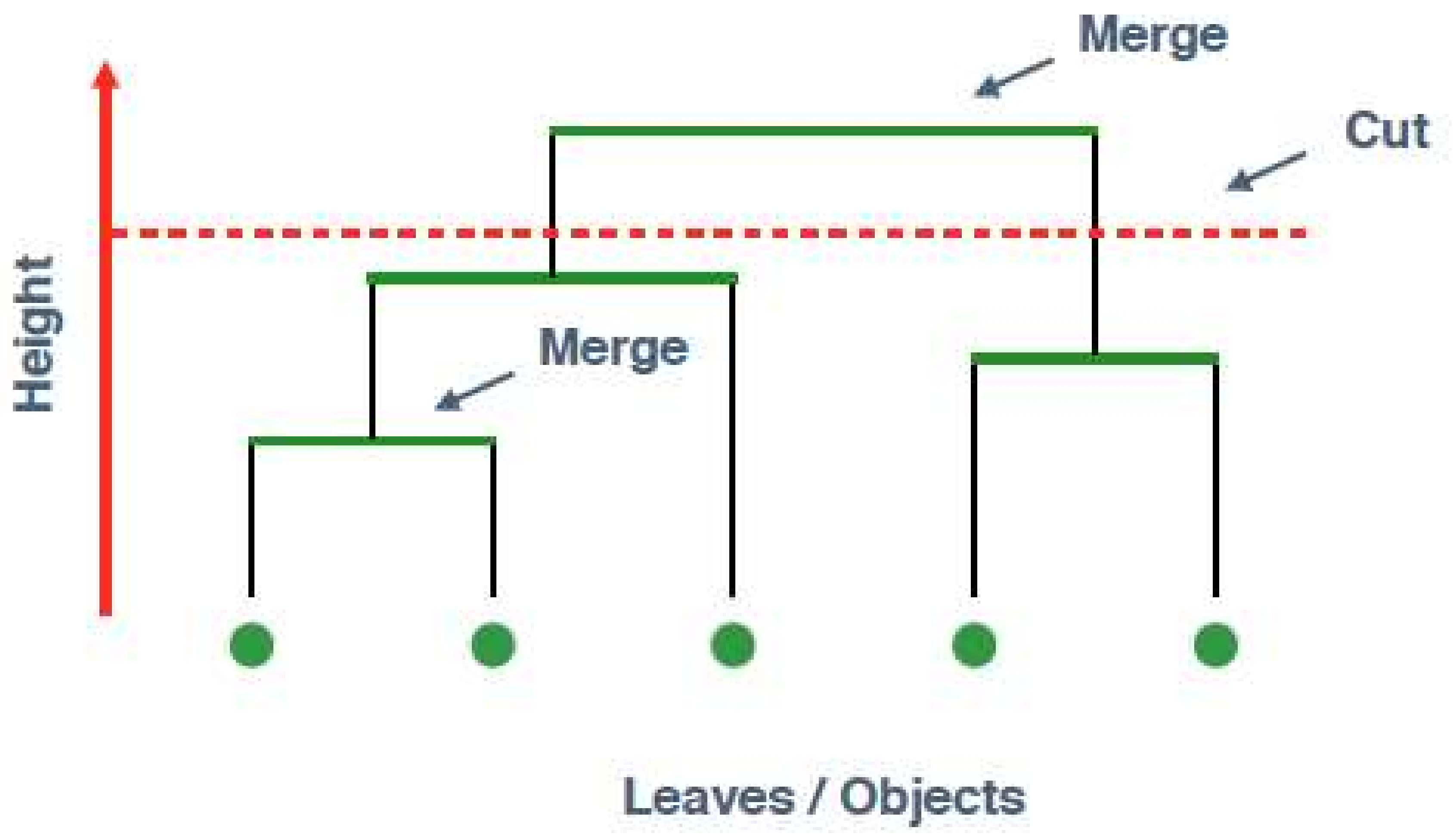


Complete - linkage

Maximal distance between objects in each cluster



Dendrogram



Hierarchical clustering in R

Library: stats

```
> dist(x, method)
```

- x: dataset
- method: distance

```
> hclust(d, method)
```

- d: distance matrix
- method: linkage

Hierarchical: Pro and Cons

- Pros
  - In-depth analysis
  - Linkage-methods → Different pattern
- Cons
  - High computational cost
  - Can never undo merges

k-means: Pro and Cons

- Pros
  - Can undo merges
  - Fast computations
- Cons
  - Fixed #Clusters
  - Dependent on starting centroids

Practise

clustering.R

Managed Independent Work

pr_clustering.R