Course. Introduction to Machine Learning

Work 1. Clustering Exercise

Session 1

Course 2021-2022

Dr. Maria Salamó Llorente

Dept. Mathematics and Informatics,

Faculty of Mathematics and Informatics,

University of Barcelona

UNIVERSITAT DE BARCELONA

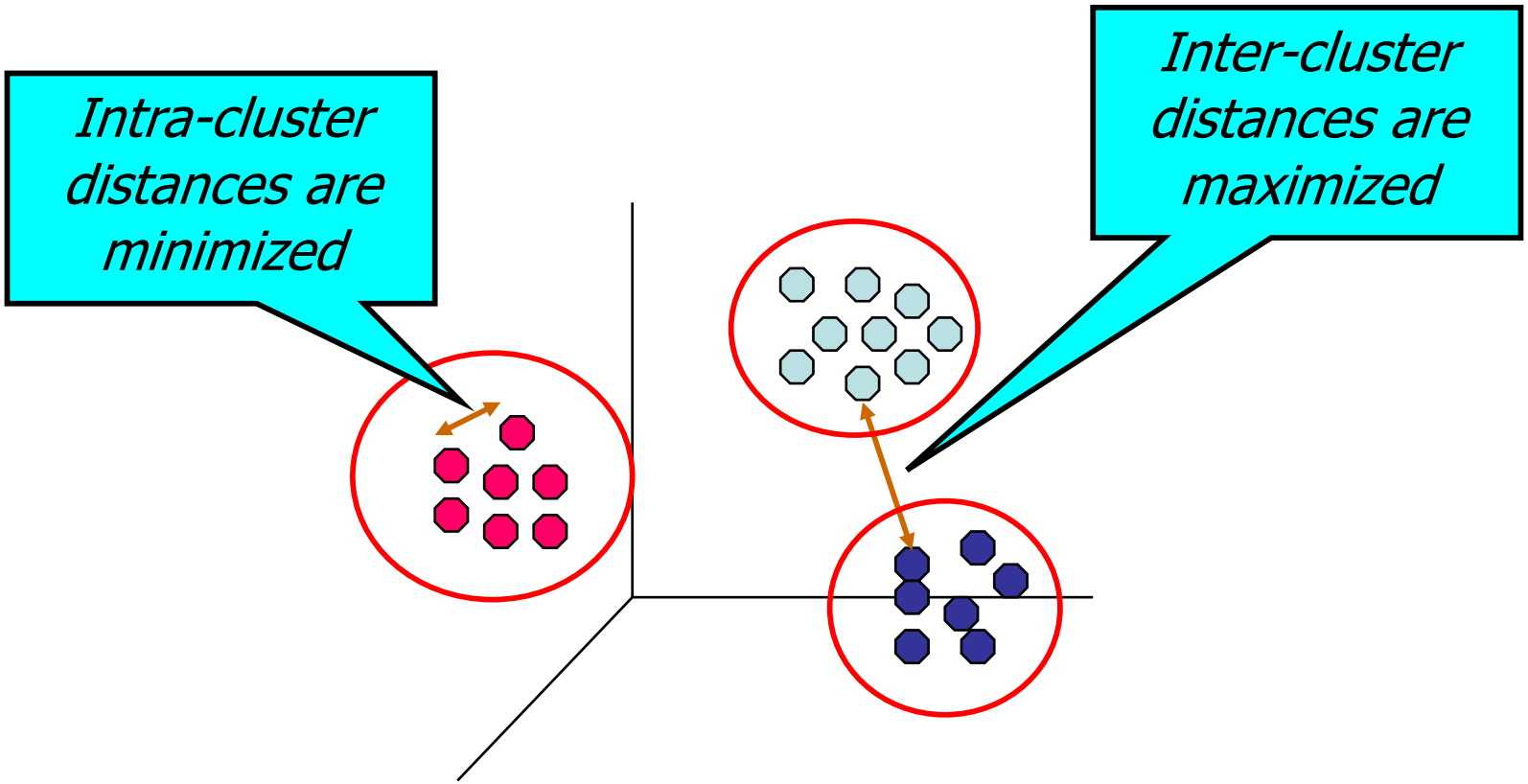
Contents

1. Introduction (session 1)
2. Preprocess the data (session 1)
3. OPTICS with sklearn (session 2)
4. K-Means (your own code) (session 2)
5. K-Modes, K-Medoids or K-Prototypes (your own code) (session 2)
6. Fuzzy clustering (your own code) (session 3)
7. Validation techniques (using sklearn validation metrics) (session 3)

Introduction

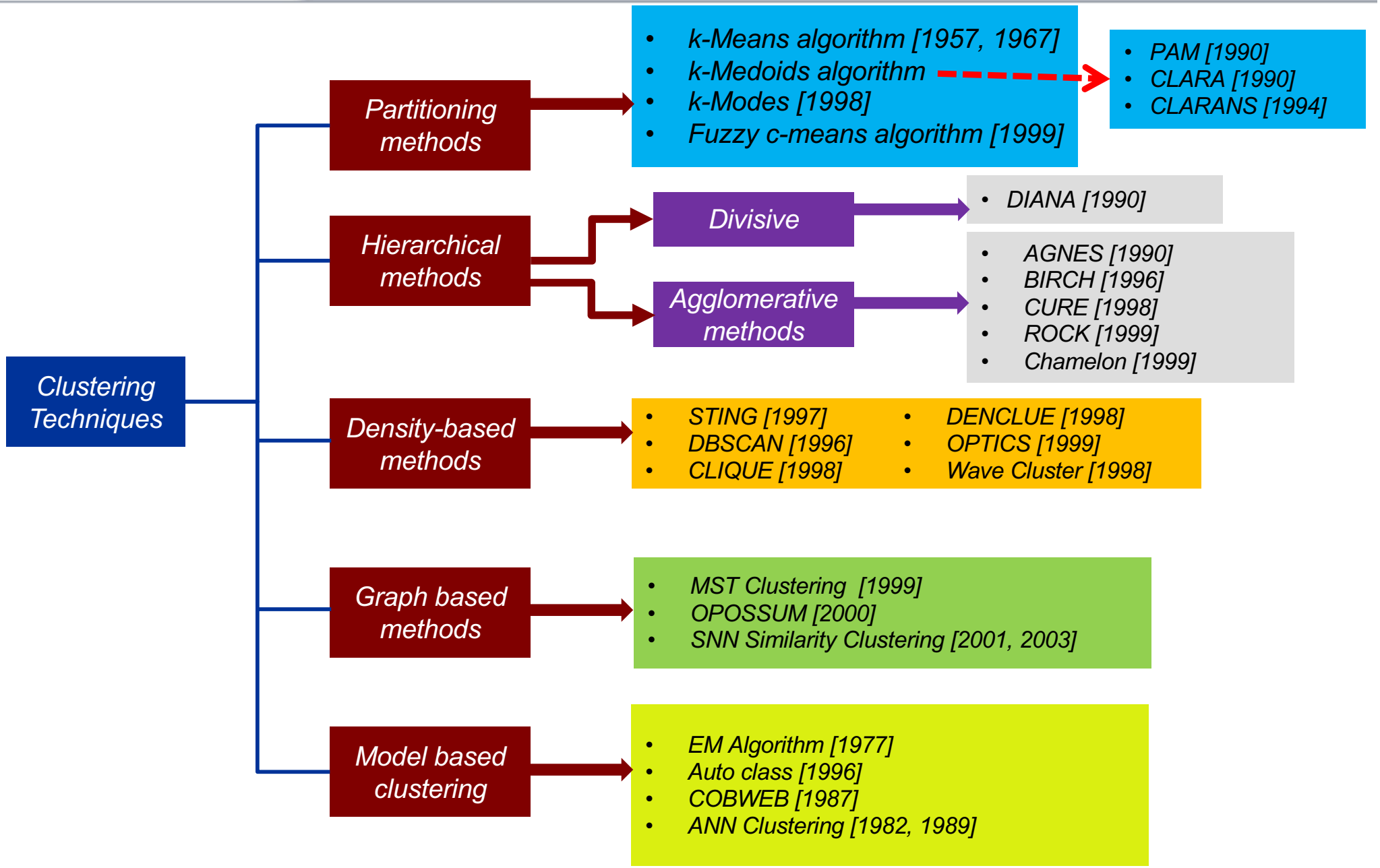
What is a Clustering?

- In general a **grouping** of objects such that the objects in a group (cluster) are similar (or related) to one another and different from (or unrelated to) the objects in other groups



UNIVERSITAT DE BARCELONA

Taxonomy of Clustering Algorithms



UNIVERSITAT DE BARCELONA

Code and Packages

- You need to implement the code using Python 3.6 and Pycharm IDE

- Packages allowed in this exercise:
  - arff_loader
  - numpy
  - pandas
  - scipy
  - sklearn (only for some parts)
  - matplotlib
  - seaborn

Preprocess the data

Data in

- You need to read the .arff file
  - You can implement your own code or use scipy.io.arff.loadarff
- Data needs pre-processing
  - Features may contain **different ranges**
    - Normalize or Standarize the machine learning data
  - Features may have **different types**
    - Categorical, Numerical, and mix-type data
  - Data may contain **missing values**
    - Use the median (for example)

Data pre-processing

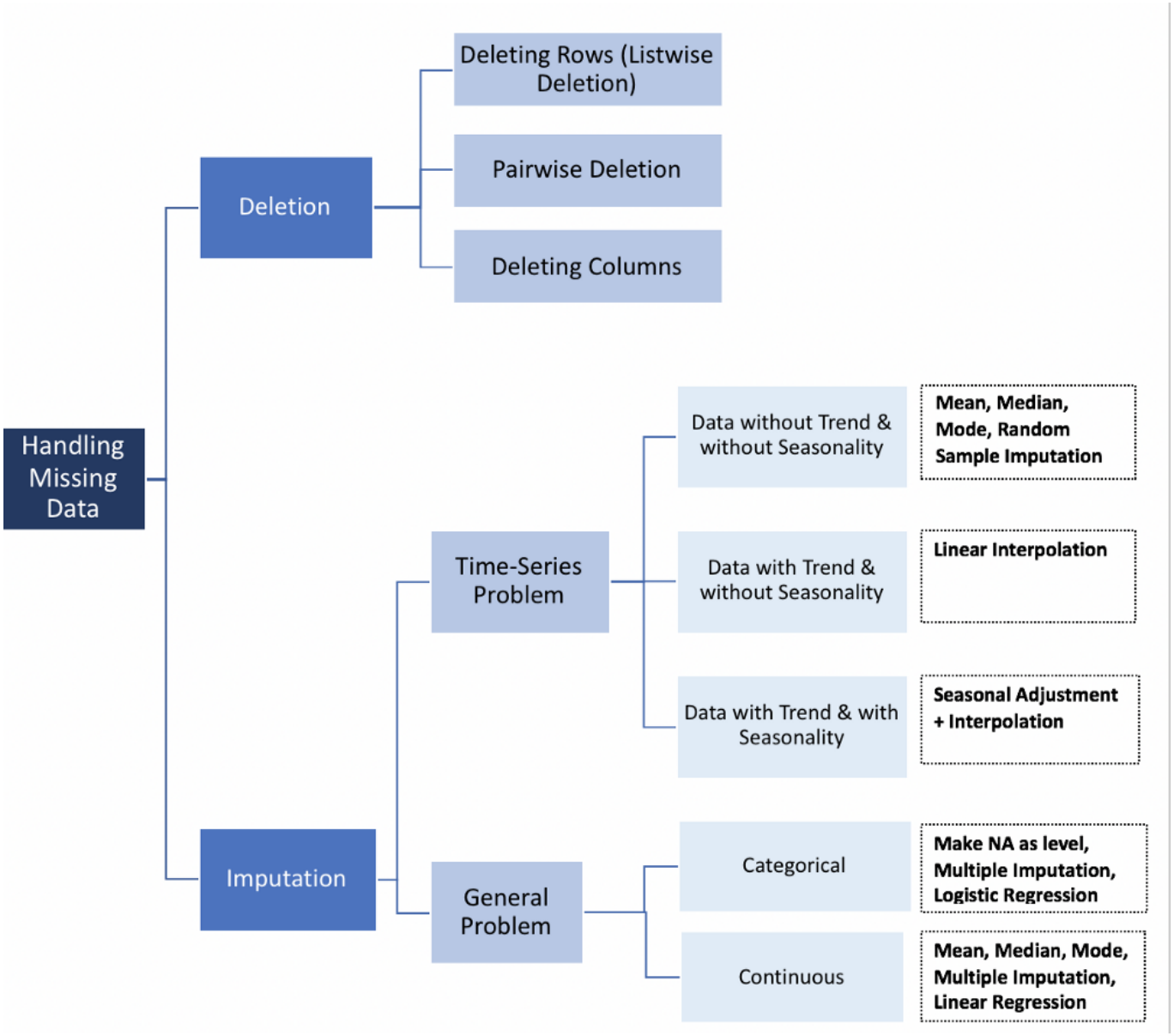
- **To deal with different ranges**
  - Normalize or scale features
- Alternatives
  - **Standardisation**: Standardisation replaces the values by their Z scores. `sklearn.preprocessing.scale`
  - **Mean normalisation**: This distribution will have values between **-1 and 1**with **μ=0**. `sklearn.preprocessing.StandardScaler`
  - **Min-Max scaling**: This scaling brings the value between 0 and 1. `sklearn.preprocessing.MinMaxScaler`
  - **Unit vector**: Scaling is done considering the whole feature vector to be of unit length. `sklearn.preprocessing.Normalizer`

Data pre-processing

- **To deal with different types**
- Alternatives
  - **Label encoding**: convert to a number
    `sklearn.preprocessing.LabelEncoder`
  - **One hot encoding**: where a categorical variable is converted into a binary vector, each possible value of the categorical variable becomes the variable itself with default value of zero and the variable which was the value of the categorical variable will have the value 1.
    `sklearn.preprocessing.OneHotEncoder`

Data pre-processing

- **To deal with missing values**



Course. Introduction to Machine Learning

Work 1. Clustering Exercise

Session 1

Course 2021-2022

Dr. Maria Salamó Llorente

Dept. Mathematics and Informatics,

Faculty of Mathematics and Informatics,

University of Barcelona