UNIVERSITAT DE BARCELONA

Introduction to Machine Learning

Master in Artificial Intelligence

UPC, UB, URV

Course. Introduction to Machine Learning

Work 2. Classification with Lazy Learning and SVM

Session 1

Dr. Maria Salamó Llorente

Dept. Mathematics and Informatics,

Faculty of Mathematics and Informatics,
University of Barcelona

Introduction

The **goal** of Work 2 is to…

1. Implement a parser to read data (Week 1)
2. Implement a kNN algorithm (Week 2)
3. Implement kNN parameters
    1. Distance metrics (Week 2)
    2. Voting schemes (Week 2)
    3. Weighting approaches (Week 3)
4. Analyze Best kNN algorithm (Week 3)
5. Use an SVM (Week 4)
6. Implement instance reduction techniques (Week 4)
7. Compare kNN with and without reduction techniques using different metrics: accuracy, efficiency and storage (Week 4)
8. Perform statistical analysis and write the report (Week 5)

Contents

1. Introduction to pre-process data

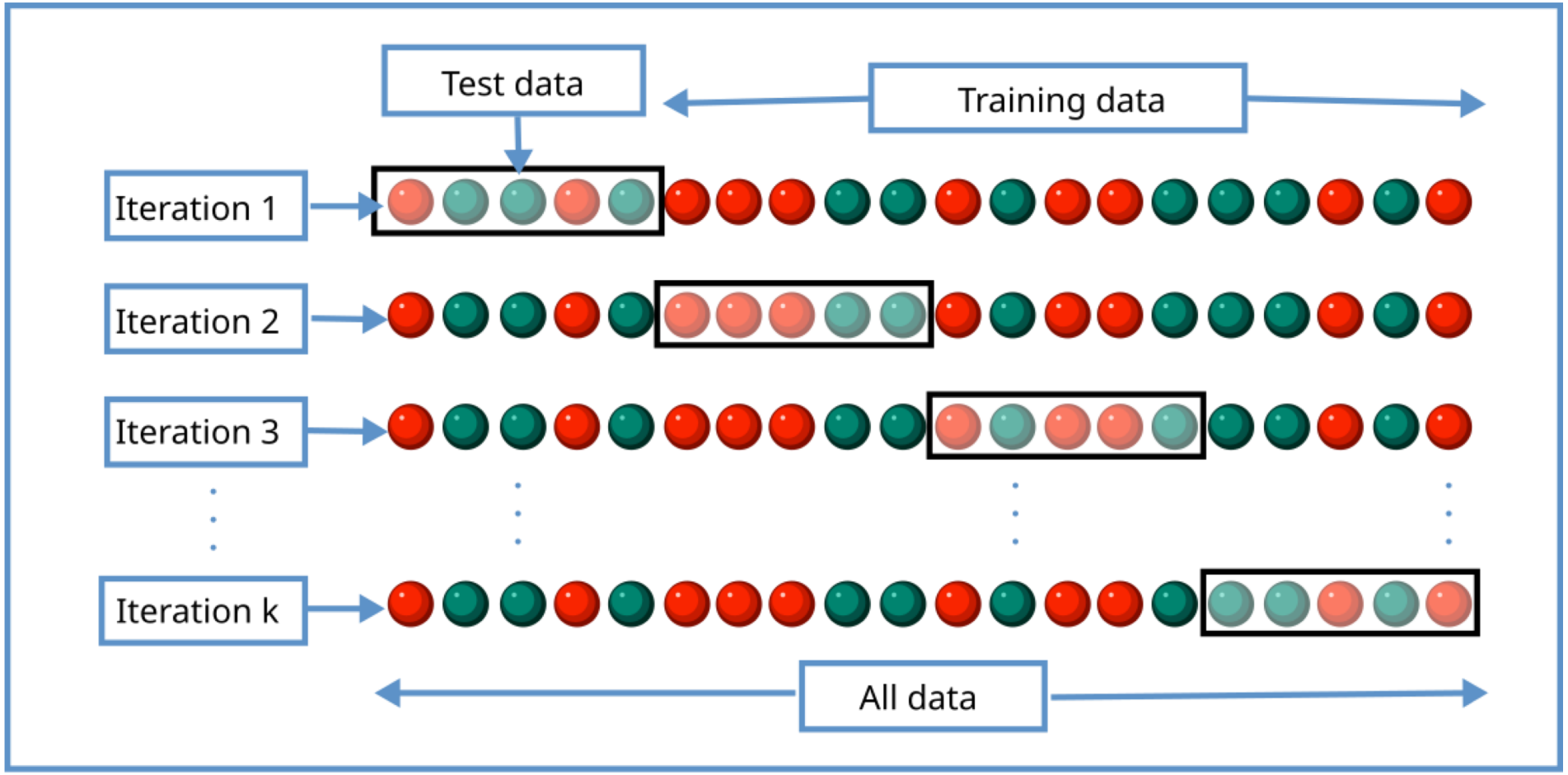
Code and Packages

- You need to implement the code using Python 3.9 and Pycharm IDE

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

Data

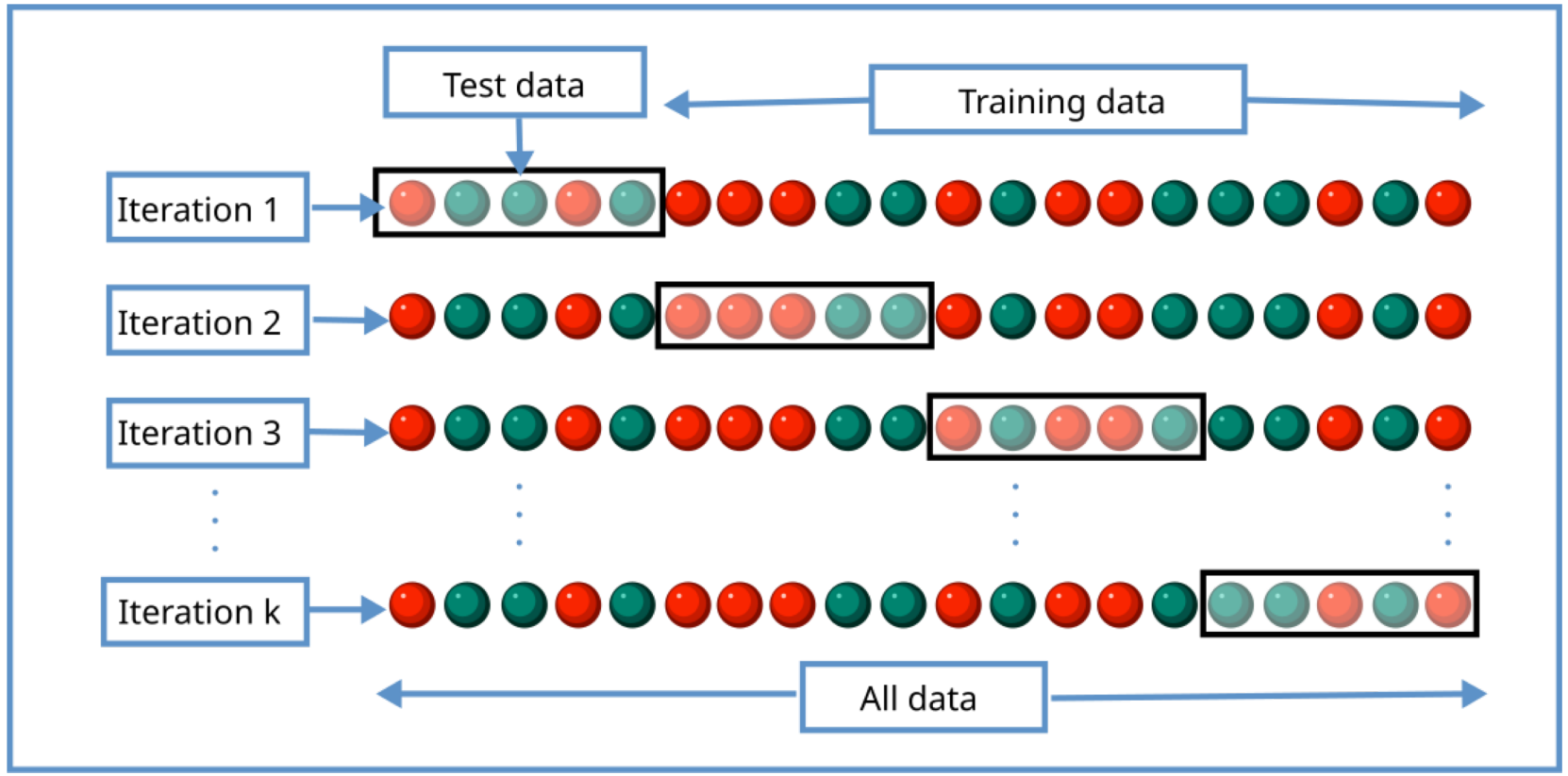
You will use a **predefined stratified 10-fold cross-validation** sets



Fold description

You will use a **predefined stratified 10-fold cross-validation** sets

`DatasetName.fold.number.train.arff`
`DatasetName.fold.number.test.arff`




```
vote.fold.000000.train.arff
vote.fold.000000.test.arff

vote.fold.000001.train.arff
vote.fold.000001.test.arff

.
.
.

vote.fold.000009.train.arff
vote.fold.000009.test.arff
```

Data

- 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

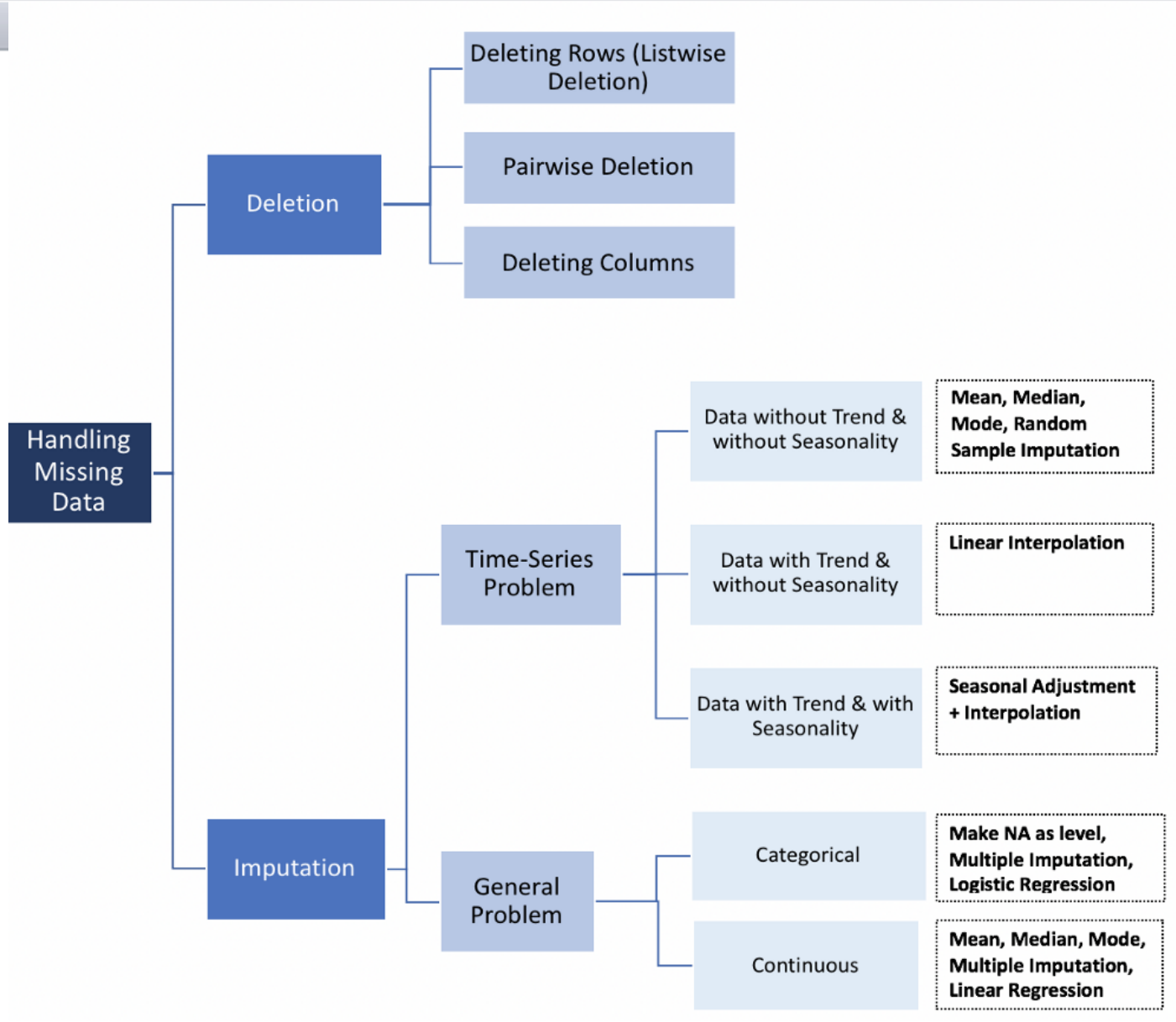
- **To deal with different ranges**
  - Normalize or scale features
- Alternatives
  - **Standardisation**: Standardisation replaces the values by their Z scores. `sklearn.preprocessing.scale`
  - **Mean normalisation**: This distribution will have values between **-1 and 1** with **μ=0**. `sklearn.preprocessing.StandardScaler`
  - **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 2. Classification with Lazy Learning and SVM

Session 1

Dr. Maria Salamó Llorente

Dept. Mathematics and Informatics,

Faculty of Mathematics and Informatics,
University of Barcelona