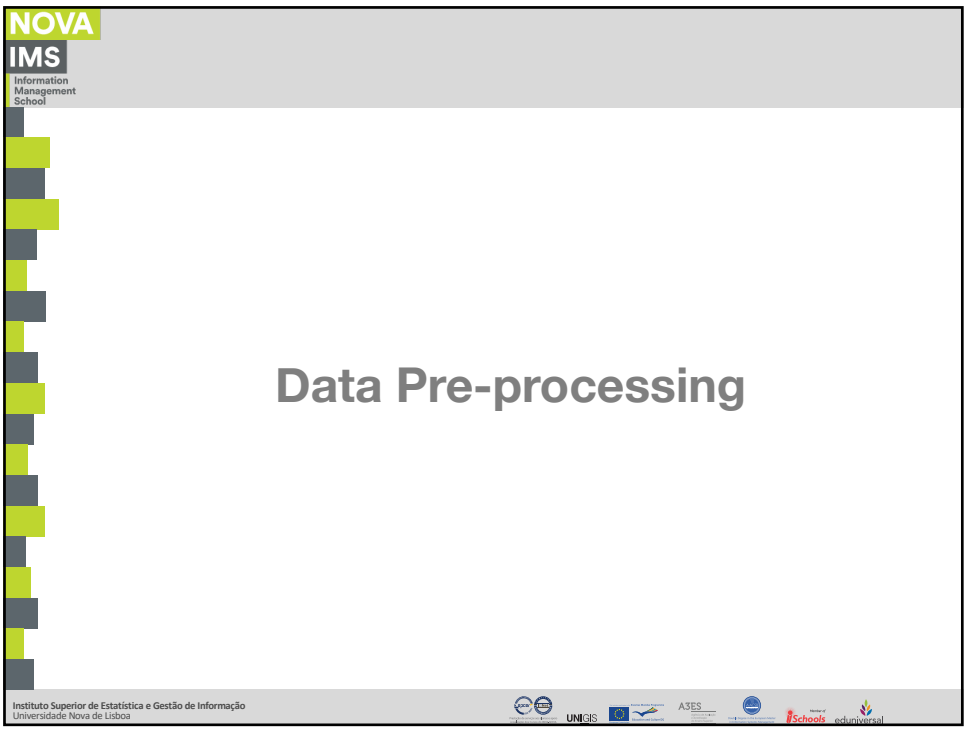

NOVA
IMS
Information
Management
School
Data Mining
211020
NOVA-IMS 2019/2020
Fernando Lucas Bação
bacao@isegi.unl.pt
http://www.isegi.unl.pt/fbacao
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

1


NOVA
IMS
Information
Management
School
Data Pre-processing
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

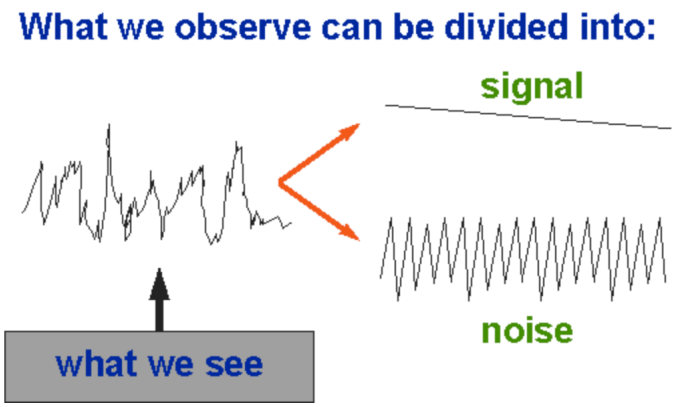
2

Data Preprocessing

- **Reasons:**
  - Noise Reduction;
  - Signal amplification;
- **Tasks:**
  - Domain-specific knowledge application;
    - Constructing ratios and derived variables
  - Size Reduction of the Input Space;
    - Remove correlated variables
    - Remove irrelevant variables
  - Normalization;

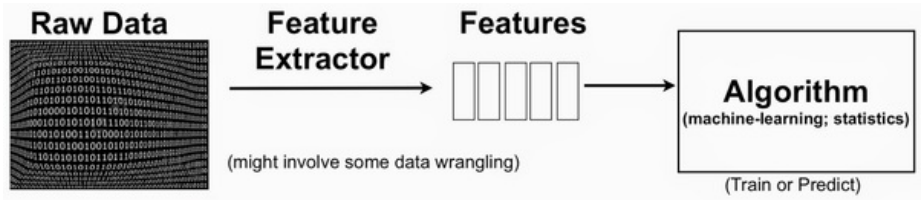
3

Data Preparation



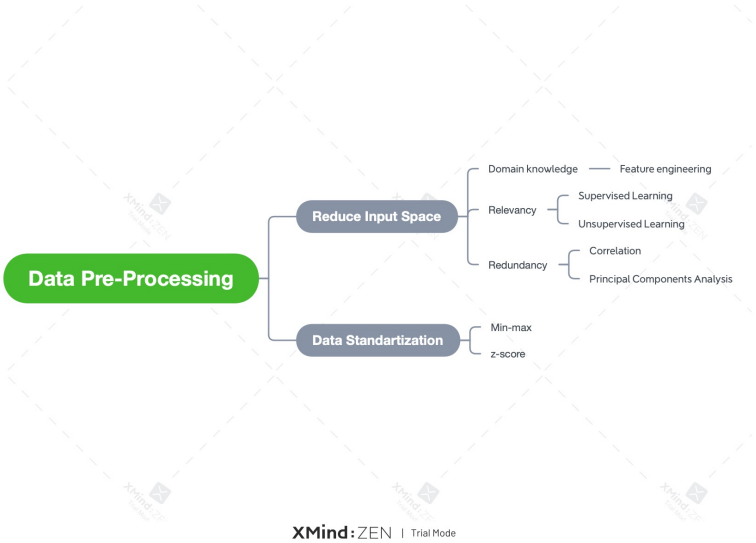
4

Data Preparation



5

Data Preparation



6

Reducing Input Space

7

Data Preprocessing

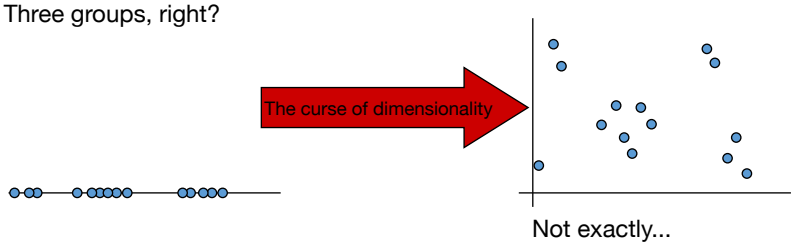
- **Additional considerations about data:**
  - Curse of dimensionality – the input space grows exponentially with the number of input variables;
  - The larger the input space, the more data and computing power we need.

8

NOVA IMS Information Management School

Data Preprocessing

Three groups, right?



When the dimensionality increases, the space becomes more sparse and it becomes more difficult to find groups

9

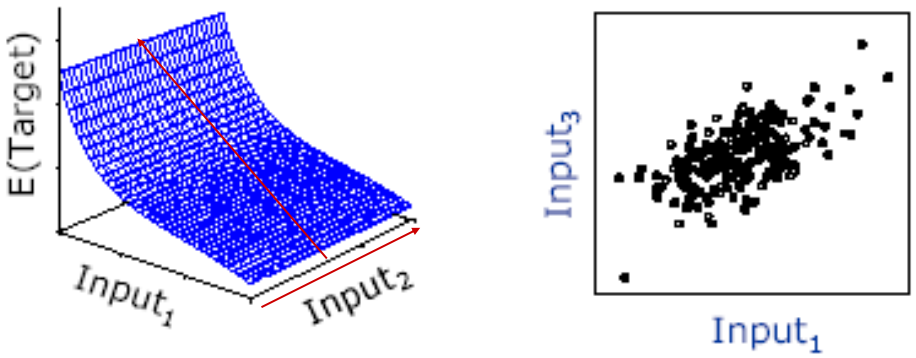
NOVA IMS Information Management School

Data Preprocessing

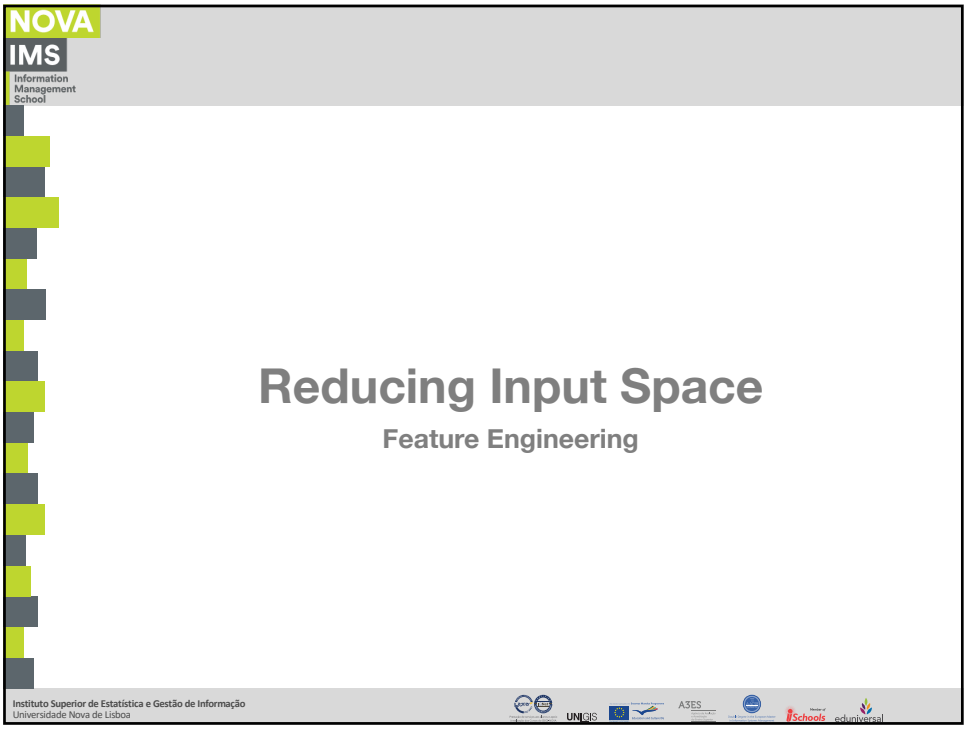
- **Size Reduction of the Input Space** (or feature selection):

Two major principles:

Relevance and Redundancy



10



11

Data Preprocessing

- **Size Reduction of the Input Space:**
  - To create input combinations
    - Height2/weight (obesity index)
    - Population/area (density)
    - Euros spent/nº of purchases (average buy)
    - Euros spent/time as customer
    - Debt/income
    - Average number of different products purchased per transaction
    - Relative spend on each product

12

- **Size Reduction of the Input Space:**
  - To create input combinations
    1. Average time between transactions (transaction interval)
    2. Variance of transaction interval
    3. Customer stability index (ratio of (2)/(1))

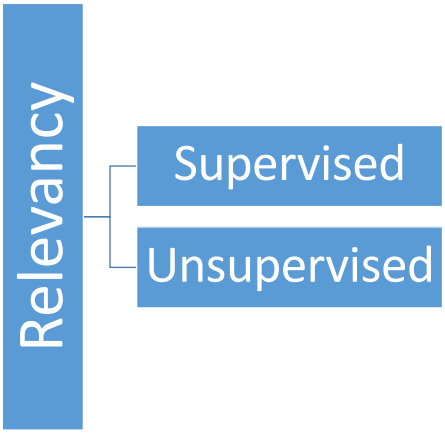
13

Reducing Input Space

Relevancy

14

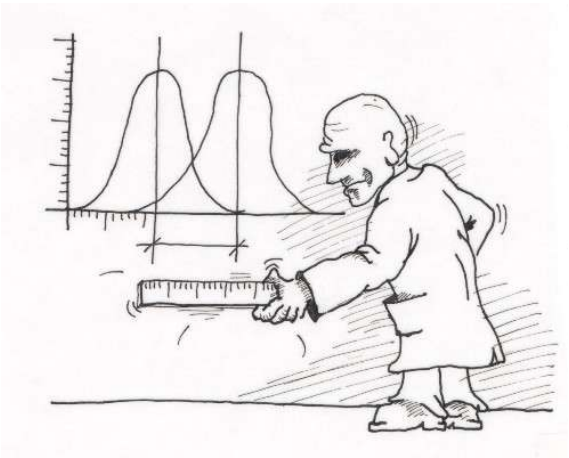
Data Preparation



15

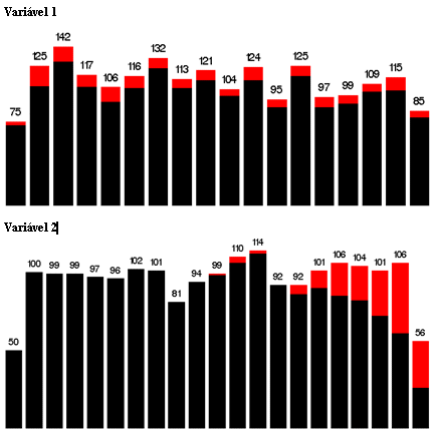
Data Preprocessing

- **Size Reduction of the Input Space:**



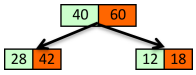
16

- **Size Reduction of the Input Space:**

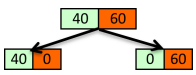


17

- **Size Reduction of the Input Space:**
  - Heuristic feature selection methods:
    - Best single features
      - Choose by information gain measures (e.g. entropy)
      - A feature is interesting if it reduces uncertainty



No improvement



Perfect Split

18

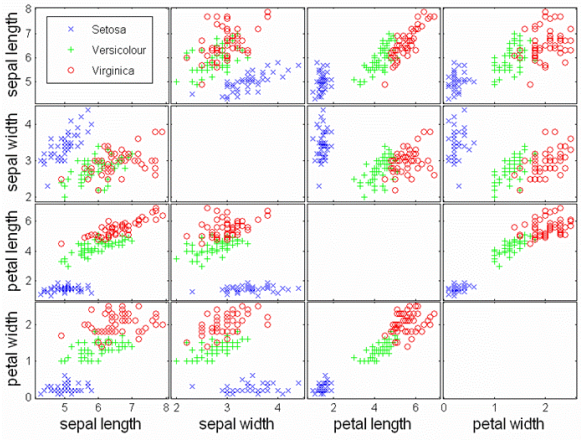
Reducing Input Space

Redundancy

19

Data Preprocessing

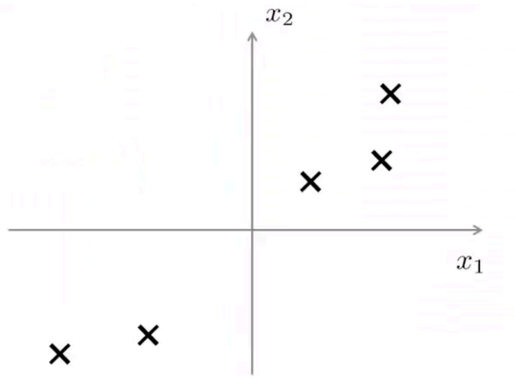
- **Size Reduction of the Input Space:**



20

- **Size Reduction of the Input Space:**
  - Principal Component Analysis
    - A procedure that uses an **orthogonal transformation** to convert a set of observations of possibly correlated variables into a set of values of **linearly uncorrelated variables called principal components**.
    - The number of principal components is **equal to the number of original variables**.
    - This transformation is defined in such a way that the first principal component has the **largest possible variance** (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance.

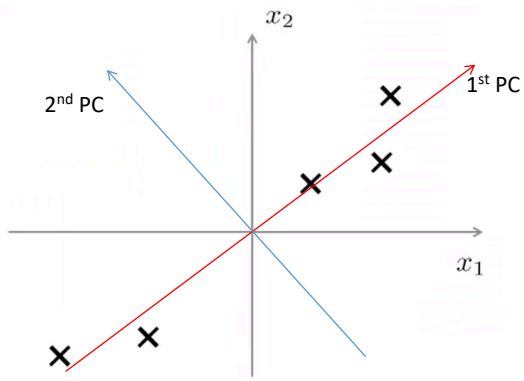
- **Size Reduction of the Input Space:**
  - Principal Component Analysis



NOVA IMS Information Management School

Data Preprocessing

- **Size Reduction of the Input Space:**
  - Principal Component Analysis

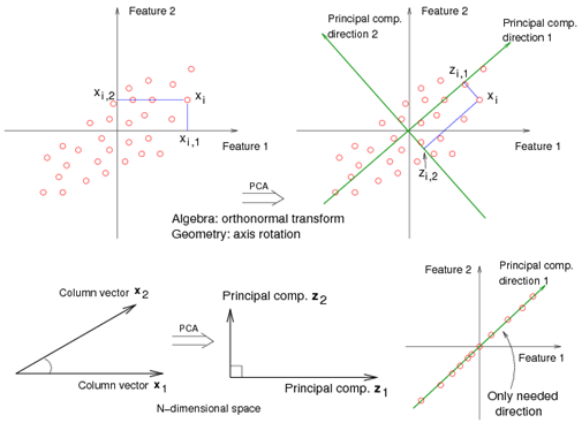


23

NOVA IMS Information Management School

Data Preprocessing

- **Size Reduction of the Input Space:**
  - Principal Component Analysis

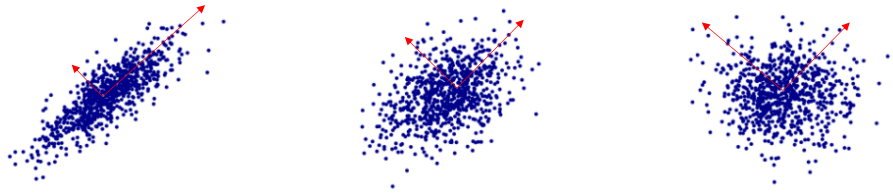


24

NOVA IMS Information Management School

Data Preprocessing

- **Size Reduction of the Input Space:**
  - Principal Component Analysis

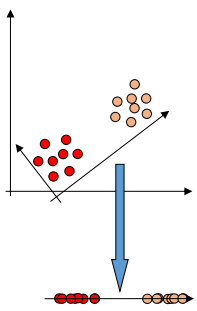


25

NOVA IMS Information Management School

Data Preprocessing

- **Size Reduction of the Input Space:**
  - Principal Component Analysis (careful)

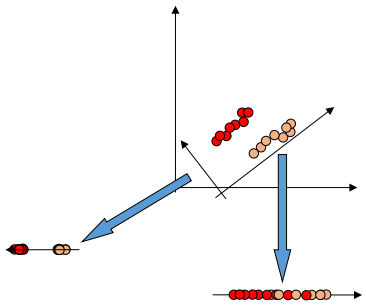


26

NOVA IMS Information Management School

Data Preprocessing

- **Size Reduction of the Input Space:**
  - Principal Component Analysis (careful)

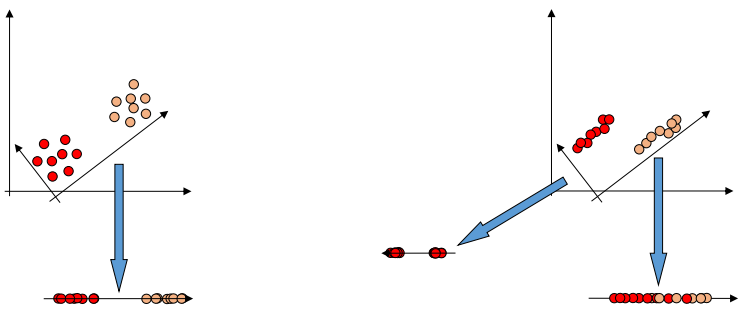


27

NOVA IMS Information Management School

Data Preprocessing

- **Size Reduction of the Input Space:**
  - Principal Component Analysis (careful)



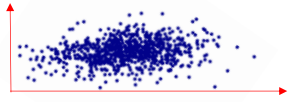
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

28

Data Standardization

29

- **Normalization:**
  - Models assume that the distances in different directions of the input space have the same importance.
  - Variables come in many **different scales** (percentages, euros, kilos, meters, days…)
  - Normalization: is about adjusting values measured on different scales to a common scale



30

NOVA IMS Information Management School

Data Preprocessing

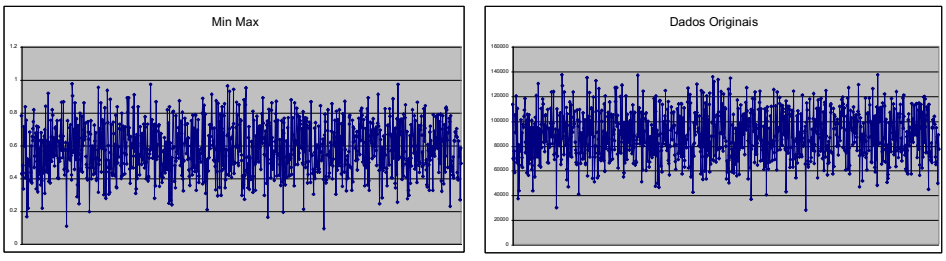
- **Normalization:**
  - Min-Max $$y' = \left(\frac{y - \min 1}{\max 1 - \min 1} \right) \underbrace{(\max 2 - \min 2) + \min 2}_{\text{optional}}$$
  - Zscore $$y' = \frac{y - \mu}{std}$$

31

NOVA IMS Information Management School

Data Preprocessing

- **Normalization:**



32

33