



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

NOVA
IMS
Information Management School

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;

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

3

NOVA
IMS
Information Management School

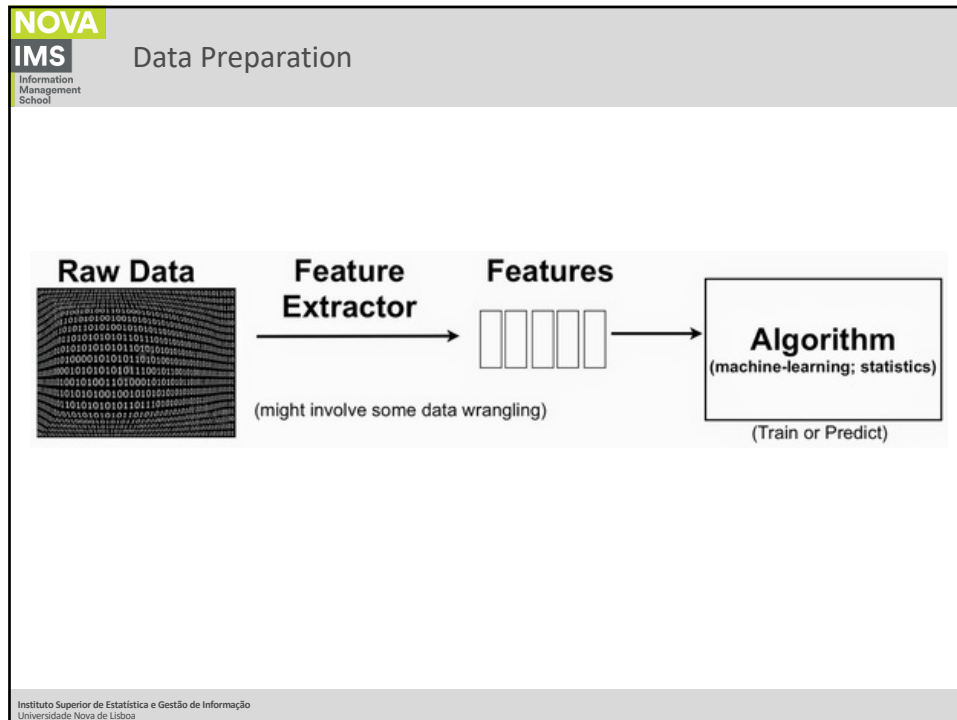
Data Preparation

What we observe can be divided into:

The diagram shows a box labeled "what we see" with an upward arrow pointing to a jagged, noisy waveform. From this waveform, a red arrow points to the right, where the waveform is decomposed into two parts: a smooth, slightly downward-sloping line labeled "signal" and a high-frequency, high-amplitude oscillation labeled "noise".

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

4



5





6

NOVA
IMS
Information Management School

Reducing Input Space

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



7

NOVA
IMS
Information Management School

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.

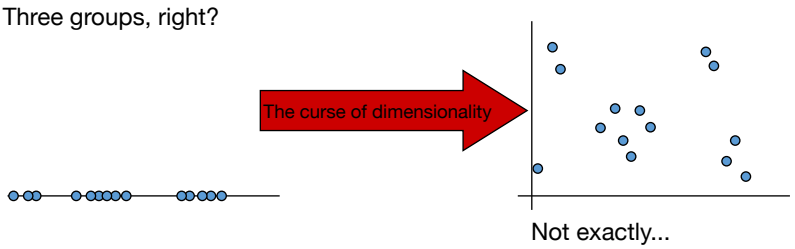
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

8

NOVA
IMS
Information Management School

Data Preprocessing

Three groups, right?



Not exactly...

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

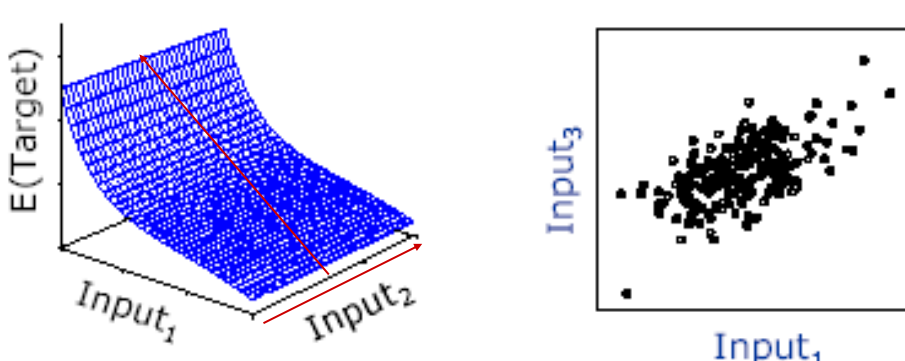
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

9

NOVA
IMS
Information Management School

Data Preprocessing

- **Size Reduction of the Input Space** (or feature selection):
Two major principles:
Relevance and Redundancy



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

10

NOVA
IMS
Information Management School

Reducing Input Space

Feature Engineering

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

UNIGES, A3ES, Schools, eaduniversal

11

NOVA
IMS
Information Management School

Data Preprocessing

- **Size Reduction of the Input Space:**
 - To create input combinations
 - Height²/weight (obesity index)
 - Population/area (density)
 - Euros spent/n° of purchases (average buy)
 - Euros spent/time as customer
 - Debt/income
 - Average number of different products purchased per transaction
 - Relative spend on each product

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

12

NOVA
IMS
Information Management School

Data Preprocessing

- **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))


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

13


NOVA
IMS
Information Management School

Reducing Input Space

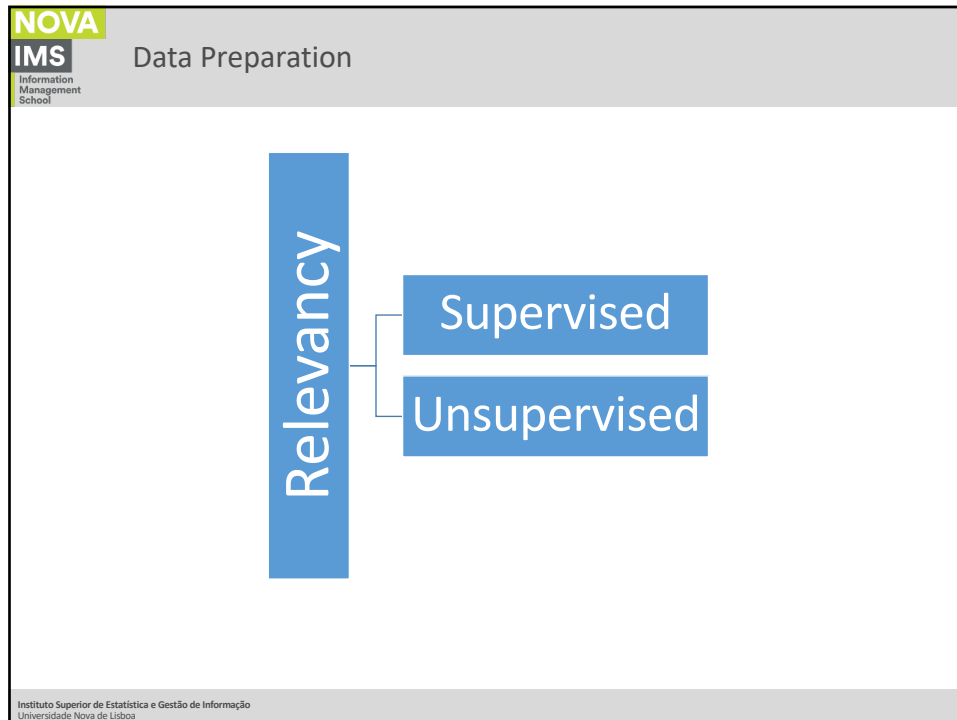
Relevancy



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



14



15

NOVA
IMS
Information Management School

Data Preprocessing

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

A cartoon illustration shows a man in a white lab coat holding a large ruler horizontally. He is measuring the width of two overlapping bell curves on a graph. The graph has a vertical axis with tick marks. The man is looking at the ruler with a focused expression.

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

16

NOVA
IMS
Information Management School

Data Preprocessing

- Size Reduction of the Input Space:**

Variável 1

Bar	Black	Red
1	75	0
2	125	0
3	142	0
4	117	0
5	106	0
6	116	0
7	132	0
8	113	0
9	121	0
10	104	0
11	124	0
12	95	0
13	125	0
14	97	0
15	99	0
16	105	0
17	115	0
18	85	0

Variável 2

Bar	Black	Red
1	50	0
2	100	0
3	99	0
4	99	0
5	97	0
6	96	0
7	102	0
8	101	0
9	81	0
10	94	0
11	99	0
12	110	0
13	114	0
14	92	0
15	92	0
16	101	0
17	106	0
18	104	0
19	101	0
20	108	0
21	56	0

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

17

NOVA
IMS
Information Management School

Data Preprocessing

- 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

```

graph TD
    A[40 60] --> B[28 42]
    A --> C[12 18]
  
```

Perfect Split

```

graph TD
    A[40 60] --> B[40 0]
    A --> C[0 60]
  
```

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

18

NOVA
IMS
Information Management School

Reducing Input Space

Redundancy

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

UNIGES, A3ES, Schools, eduniversal

19

NOVA
IMS
Information Management School

Data Preprocessing

- Size Reduction of the Input Space:**

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

20

NOVA
IMS
Information Management School

Data Preprocessing

- **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.

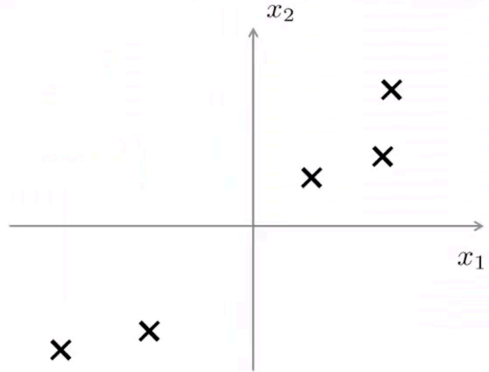
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

21

NOVA
IMS
Information Management School

Data Preprocessing

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



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

22

NOVA
IMS
Information Management School

Data Preprocessing

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

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

23

NOVA
IMS
Information Management School

Data Preprocessing

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

Algebra: orthonormal transform
Geometry: axis rotation

Column vector x_2
Column vector x_1
Principal comp. z_2
Principal comp. z_1
N-dimensional space

Only needed direction

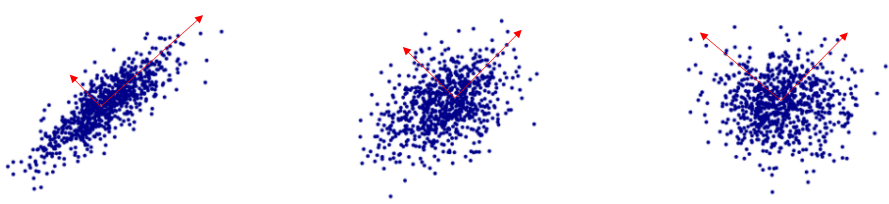
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

24

NOVA
IMS
Information Management School

Data Preprocessing

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



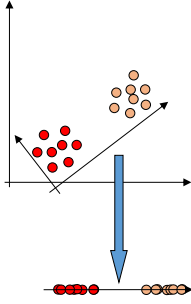
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

25

NOVA
IMS
Information Management School

Data Preprocessing

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



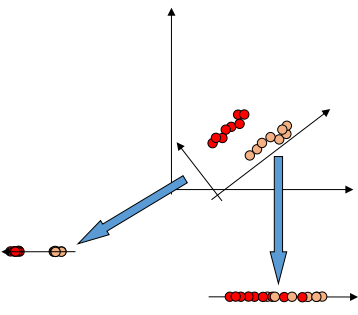
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

26

NOVA
IMS
Information Management School

Data Preprocessing

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



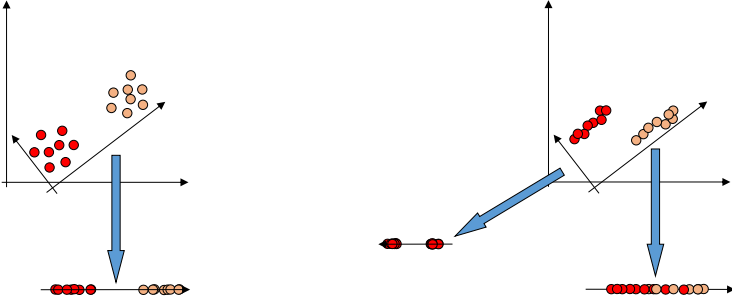
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

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

NOVA

IMS

Information
Management
School

Data Standardization

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

29

NOVA

IMS

Information
Management
School

Data Preprocessing

- **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

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

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
$$y' = \frac{y - \mu}{std}$$

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

31

NOVA
IMS
Information Management School

Data Preprocessing

- Normalization:**

Min Max

Dados Originais

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

32

