NOVA IMS Information Management School

Data Mining

S6

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 Preparation

Data Preparation RoadMap

Missing Values → Outliers → Discretization and Encoding → Imbalanced Datasets

Data Pre-processing RoadMap

Feature Selection
- Relevancy analysis
- Redundancy removal

→ Feature Engineering → Feature Scaling and Normalization

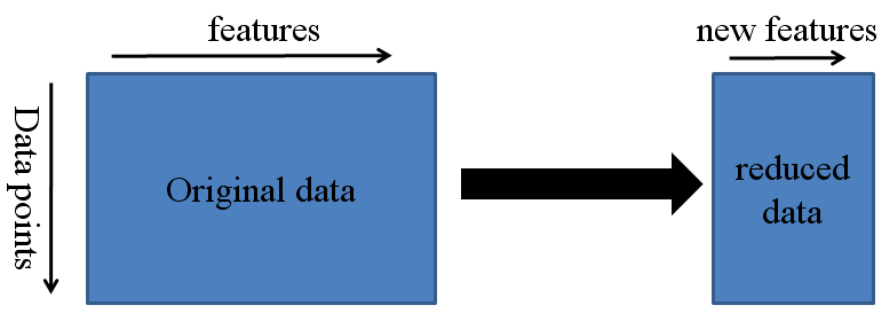
2

Data Preprocessing

Preparing data for mining models

3

Data Preparation



Through preprocessing, we transform original data into a more compact and informative representation.

4

NOVA IMS Information Management School

Data Preprocessing

- **Reasons:**
  - Noise reduction and signal amplification;
  - Size Reduction of the Input Space;
    - Domain-specific knowledge application;
    - Feature selection
      - Remove correlated variables
      - Remove irrelevant variables
    - Feature engineering
      - Constructing ratios and derived variables
  - Scaling (normalization);

5

NOVA IMS Information Management School

Reducing Input Space

Why fewer variables often means better models

6

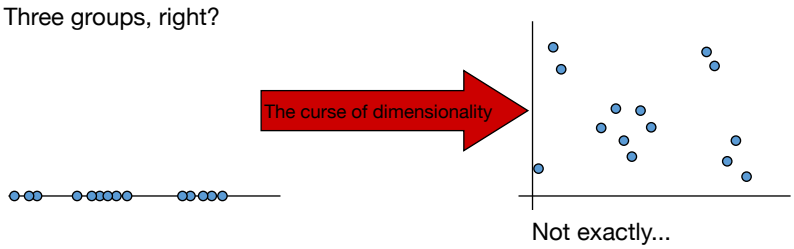
Data Preprocessing

- **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.
  - In high dimensions, distance metrics lose meaning.

7

Data Preprocessing

Three groups, right?



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

As dimensionality increases, apparent clusters disappear.

8

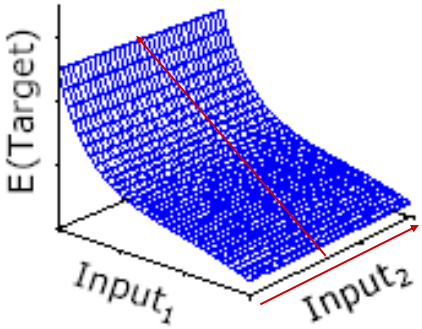
NOVA IMS Information Management School

Data Preprocessing

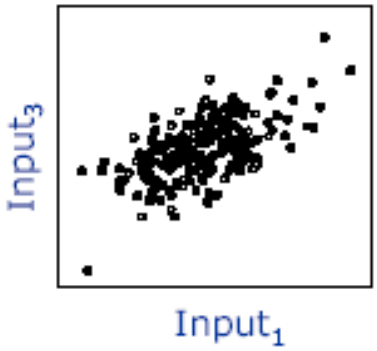
- **Size Reduction of the Input Space** (or feature selection):

Two major principles:

Irrelevance and Redundancy



Irrelevant variable: doesn't affect target.



Redundant variable: highly correlated with another

9

NOVA IMS Information Management School

Reducing Input Space

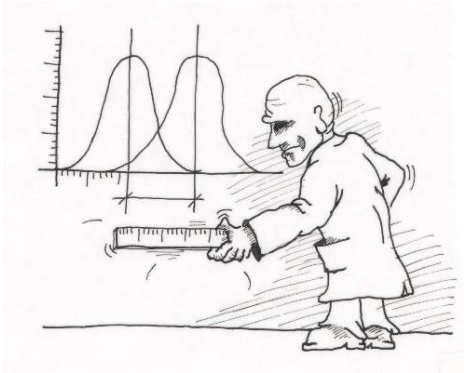
Relevancy

10

NOVA IMS Information Management School

Data Preprocessing

- **Size Reduction of the Input Space:**
  - **Feature selection: measuring which variable best separates the classes**



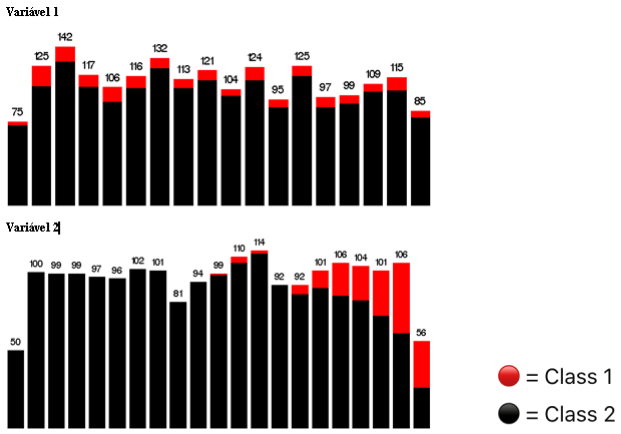
Feature selection helps us find the variables that best separate the target classes. We are, in essence, measuring which 'ruler' (feature) allows us to distinguish one group from another.

11

NOVA IMS Information Management School

Data Preprocessing

- **Size Reduction of the Input Space:**
  - **Feature selection: measuring which variable best separates the classes**



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
    - Height2/weight (Body Mass Index, BMI)
    - Population/area
    - Euros spent/nº of purchases
    - Euros spent/time as customer
    - Debt/income

13

NOVA IMS Information Management School

Data Preprocessing

- **Size Reduction of the Input Space:**
  - To create input combinations

1. Customer ID (could be anonymous)
2. Total revenue for the customer
3. Number of transactions per customer (frequency)
4. Average time between transactions (transaction interval)
5. Variance of transaction interval
6. Customer stability index (ratio of (5)/(4))
7. Days since last visit (recency)
8. Average number of different products purchased per transaction
9. Relative spend on each product
10. NRS on each product

(and where a product taxonomy exists):

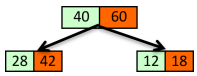
11. Relative spend in each product subgroup
12. NRS in each product subgroup
13. NRS in each product group

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

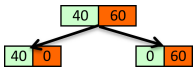
14

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



Perfect Split

15

Reducing Input Space

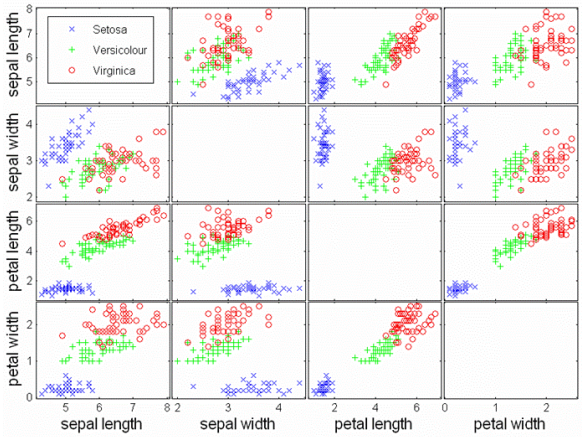
Redundancy

16

NOVA IMS Information Management School

Data Preprocessing

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



NOVA IMS Information Management School

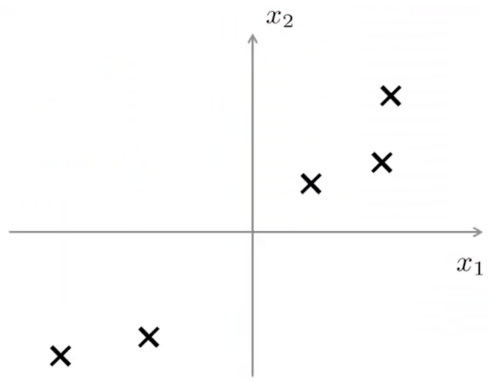
Data Preprocessing

- **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.

NOVA IMS Information Management School

Data Preprocessing

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

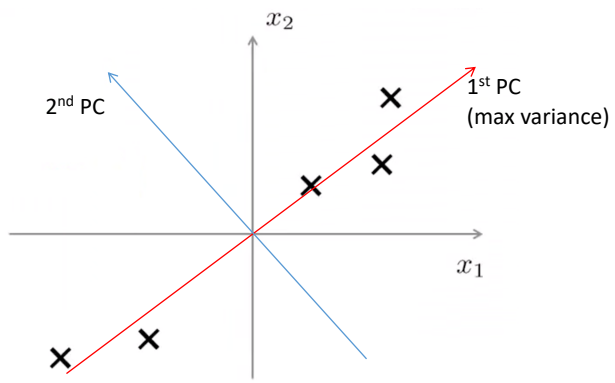


19

NOVA IMS Information Management School

Data Preprocessing

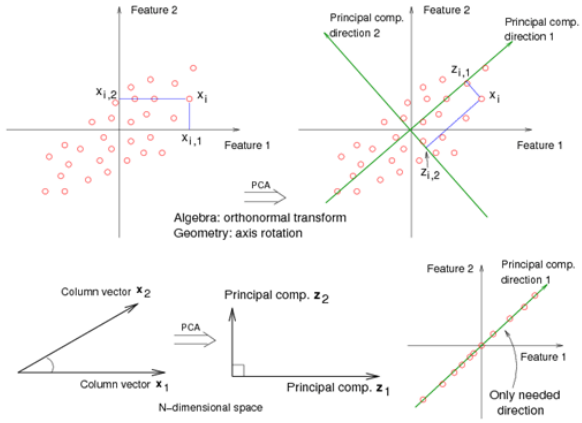
- **Size Reduction of the Input Space:**
  - Principal Component Analysis



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

20

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

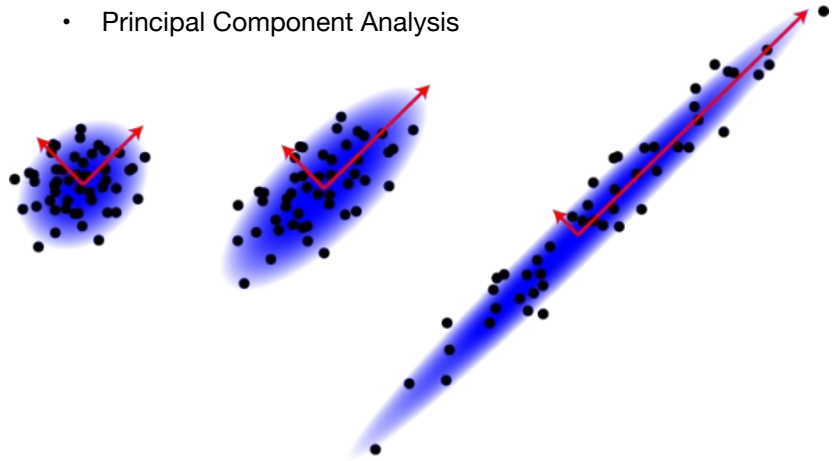


- **Size Reduction of the Input Space:**
  - Each principal component (PC) captures a portion of the total variance in the dataset — that is, how much of the information (or variability) in the original variables it represents.
    - PC1 captures the largest possible variance.
    - PC2 captures the next largest, uncorrelated with PC1.
    - Subsequent PCs each explain less and less variance.
  - By summing the variance explained by each component, we get the cumulative variance explained, which tells us how much of the total information in the data is retained when using only the first k components.
  - Example:
    - PC1 explains 60% of the variance,
    - PC2 explains 25%,
    - PC3 explains 10%,
  - then the first three components together explain 95% of the total variance.

NOVA IMS Information Management School

Data Preprocessing

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

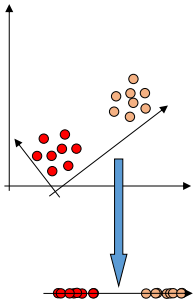


23

NOVA IMS Information Management School

Data Preprocessing

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

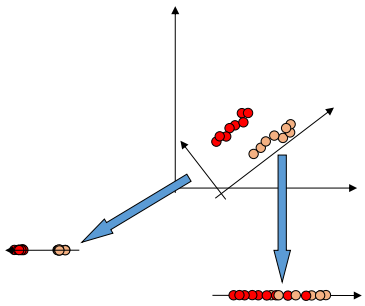


24

NOVA IMS Information Management School

Data Preprocessing

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

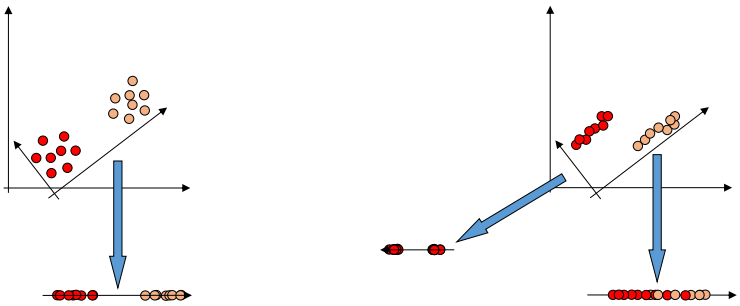


25

NOVA IMS Information Management School

Data Preprocessing

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



PCA preserves variance, not class separability
Dimensionality reduction can hurt supervised learning

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

26

NOVA IMS Information Management School

Data Standardization

27

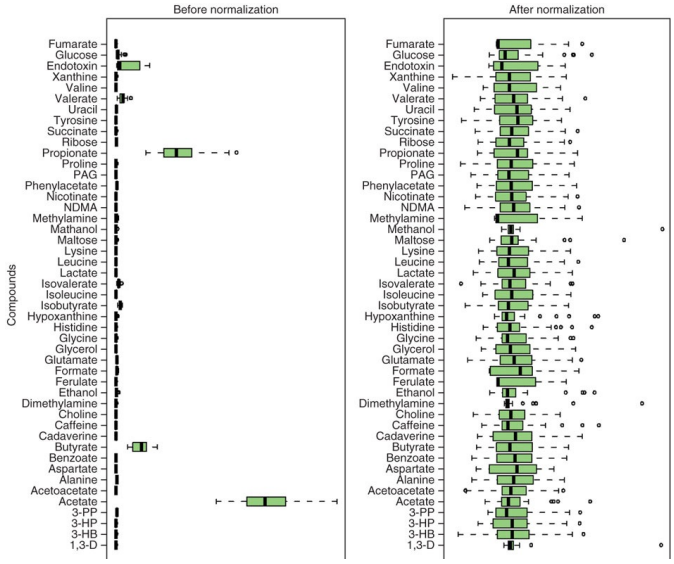
NOVA IMS Information Management School

Data Preprocessing

- **Standardization or Normalization:**
  - Many algorithms (e.g., k-means, SVMs) assume features are on comparable scales.
    - Variables come in many different scales (percentages, euros, kilos, meters, days…)
    - Normalization: is about adjusting values measured on different scales to a common scale

28

Data Preprocessing



29

Data Preprocessing

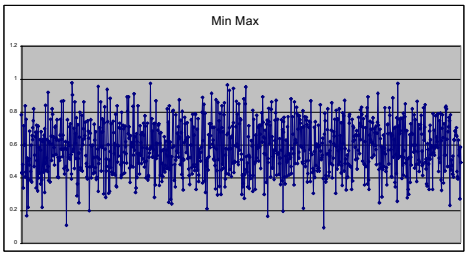
- **Standardization or Normalization:**
  - Min-Max: $x' = \frac{x - \min(x)}{\max(x) - \min(x)} \underbrace{(\max 2 - \min 2) + \min 2}_{\text{optional}}$
  - Zscore: $z = \frac{x - \mu}{\sigma}$

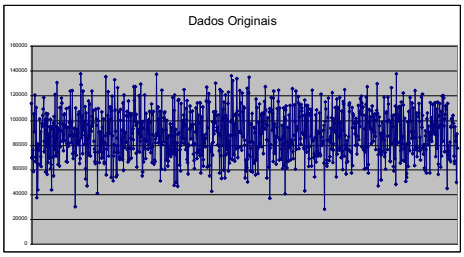
30

NOVA IMS Information Management School

Data Preprocessing

- **Normalization:**





31

Questions?

32