
NOVA
IMS
Information
Management
School
Data Mining
Association Rules
9/12/2021
NOVA-IMS
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
Association Rules
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

2

NOVA IMS Information Management School

Association Rules

- **Association Rules:**
  - Aims at the extraction of compact patterns that describe subsets of data
    - Events that occur together (market basket analysis);
    - The mais purpose is to establish relationships between fields;
    - Are rules of the form if X then Y;
    - Association rules provide information about things that tend to happen together.

3

NOVA IMS Information Management School

Association Rules

- **Association Rules:**
  - Table with a set of purchases from a supermarket, with 5 purchases and 5 items.

Cliente	Itens
1	Orange juice, soda
2	Milk, Orange juice, Glass cleaner
3	Orange juice
4	Orange juice, detergent, soda
5	Glass cleaner, soda

  - With these data we can create a table of co-occurrences with the number of times that any pair of products was purchased together.

4

Association Rules

	Orange Juice	Glass Cleaner	Milk	Soda	Detergent
Orange Juice	**4**	1	1	2	1
Glass Cleaner	1	**2**	1	1	0
Milk	1	1	**1**	0	0
Soda	2	1	0	**3**	1
Detergent	1	0	0	1	**2**

5

Association Rules

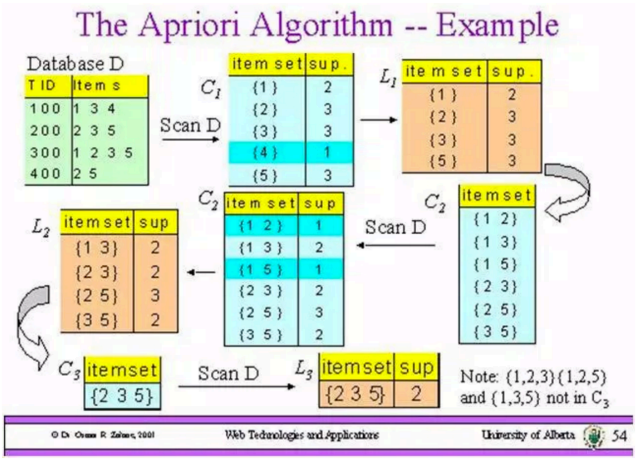
- From the analysis, one can conclude that:
  - Orange juice and soda are more likely purchased together than any other two items;
  - Detergent is never purchased with milk or window cleaner;
  - Milk is never purchased with soda or detergent.

	OJ	GC	Milk	Soda	Det.
OJ	**4**	1	1	2	1
GC	1	**2**	1	1	0
Milk	1	1	**1**	0	0
Soda	2	1	0	**3**	1
Det.	1	0	0	1	**2**

6

Association Rules

- **Association Rules:**



7

Association Rules

- **Association Rules:**
  - The rules are expressed in the form of:

"*if the item A is part of an event, then the item B will also be part of the event X percent of the time*"

  - The rules should not be interpreted as a direct causation, but only as an association;
  - It is not legitimate to infer rules of causality.

8

- **How association rules work:**

*If a customer buys **shoes**, then 10% of the time also buys **socks***

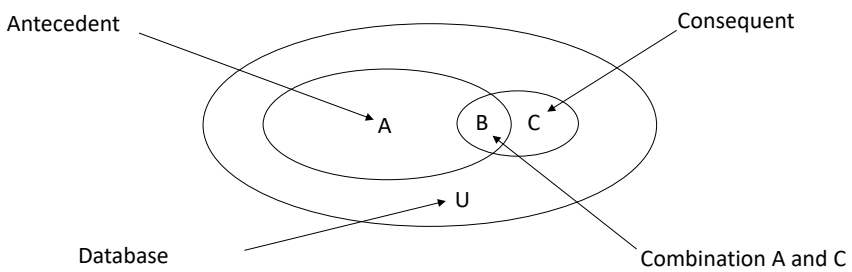
*If a customer buys **plastic paint**, then 85% of the time also buys **brushes***

- All rules have an antecedent and a consequent

*Shoes and plastic paint – **antecedents***

*Socks and brushes - **consequents***

- **Venn Diagram:**



NOVA IMS Information Management School

Association Rules

- **How association rules work:**
  - *Evaluating the quality of Association Rules:*

    Confidence

    Support

    Expected Confidence

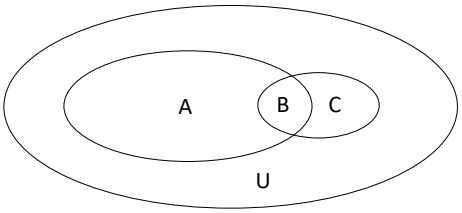
    Lift
  - *The most important measures to assess the rules*

11

NOVA IMS Information Management School

Association Rules

- **How association rules work:**
  - *Evaluating the quality of Association Rules:*
  - ***Confidence** – the **strenght of an association**, the percentage of a consequent appears given that the antecedent has occurred*



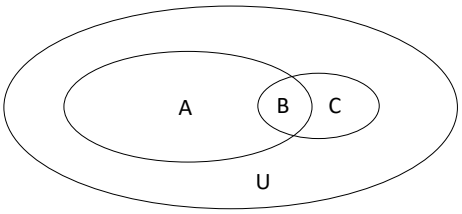
$$Confidence = B/A$$

12

NOVA IMS Information Management School

Association Rules

- **How association rules work:**
  - *Evaluating the quality of Association Rules:*
  - ***Support** – shows **how frequently** the combination occurs in the database.*



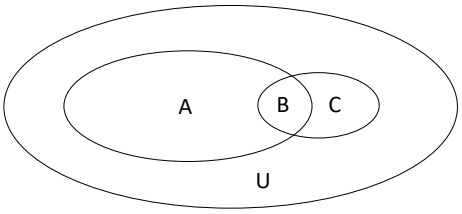
$$Support = B/U$$

13

NOVA IMS Information Management School

Association Rules

- **How association rules work:**
  - *Evaluating the quality of Association Rules:*
  - ***Expected confidence** – equal to the number of consequent transactions, divided by the total number of transactions*



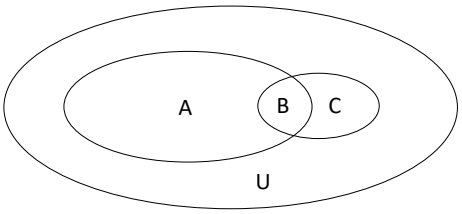
$$ExpectedConfidence = C/U$$

14

NOVA IMS Information Management School

Association Rules

- **How association rules work:**
  - *Evaluating the quality of Association Rules:*
  - ***Lift** – equal to the confidence factor divided by the expected confidence. Lift is a factor by which the **likelihood** of consequent increases given an antecedent*



$$Lift = Confidence / Expected\ Confidence$$

15

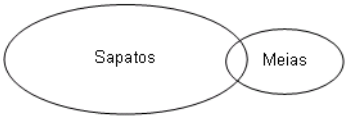
NOVA IMS Information Management School

Association Rules

Tabela de Transações

1,000,000	Total de Transações
200,000	Sapatos
50,000	Meias
20,000	Sapatos e Meias

Diagrama de Venn



$$Confidence = B/A$$

$$Support = B/U$$

$$ExpectedConfidence = C/U$$

$$Lift = Confidence / Expected\ Confidence$$

16

Association Rules

Tabela de Transações

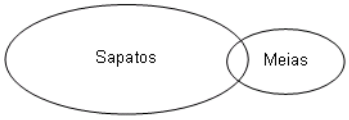
1,000,000	Total de Transações
200,000	Sapatos
50,000	Meias
20,000	Sapatos e Meias

Regra

Se um cliente compra sapatos, então 10% das vezes ele comprará meias.

Critérios de Avaliação

Diagrama de Venn



17

Association Rules

Tabela de Transações

1,000,000	Total de Transações
200,000	Sapatos
50,000	Meias
20,000	Sapatos e Meias

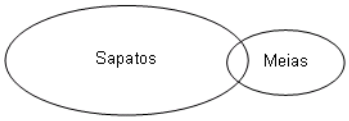
Regra

Se um cliente compra sapatos, então 10% das vezes ele comprará meias.

Critérios de Avaliação

Confiança: 20,000/200,000 = 10%
Suporte: 20,000/1,000,000 = 2%
Confiança Esperada: 50,000/1,000,000 = 5%
Lift: Confiança/Confiança Esperada = 2

Diagrama de Venn



18

Association Rules

- **How association rules work:**
  - *Evaluating the quality of Association Rules:*

Lift *– a value of 2 means that, if one bought shoes, one is twice as likely to buy socks than those who have not bought shoes*

A ***credible rule*** *must have a good confidence factor, a high level of support and lift higher than 1*

Rules with a ***high level of confidence*** *but with a* ***low support*** *should be interpreted with caution, as it can result in* ***idiosyncrasies*** *that rise due to the small number of cases to support the rule*

19

Association Rules

- **Use:**

20

Association Rules

- **Use:**

You could win a **$50 Amazon gift certificate** or the **Grand Prize**. Start by adding five items to your Wish List. Learn more.

Customers who bought this book also bought:

- How to Taste : A Guide to Enjoying Wine by Jancis Robinson (Rate it)
- The World Atlas of Wine by Hugh Johnson, Jancis Robinson (Rate it)
- The Wall Street Journal Guide to Wine: New and Improved: How to Buy, Drink, and Enjoy Wine by Dorothy J. Gaiter, John Brecher (Rate it)
- The University Wine Course: A Comprehensive Text and Tutorial by Marian W. Baldy Ph.D. (Rate it)
- Parker's Wine Buyer's Guide 6th Edition : The Complete, Easy-to-Use Reference on Recent Vintages, Prices, and Ratings for More Than 8,000 Wines from All the Major Wine Regions by Robert M. Parker (Rate it)

Explore Similar Items: 20 in Books, 20 in DVD, and 16 in Magazine Subscriptions

Association Rules

- **Types of rules**
  - Trivial rules;
  - Inexplicable Rules;
  - Actionable Rules;
  - Rules that result of promotions made.

NOVA IMS Information Management School

Association Rules

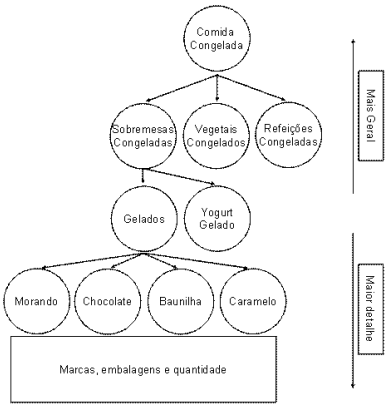
- **Additional aspects**
  - Available data are essential to success;
  - It takes large amounts of data;
  - What is an item
    - Frozen pizza, or frozen pizza 4 seasons;
  - Setting the appropriate level of detail;
  - Classifications.

23

NOVA IMS Information Management School

Association Rules

- **Additional aspects**



24

- **Additional aspects**
  - Use of virtual items;
  - Allow you to get information from data beyond the information expressed in the products;
  - May include information about the purchase (paid in cash, credit card…), about the customer (new, old…) day of the week, hour of the day.

NOVA IMS Information Management School

Data Mining

S3

NOVA-IMS 2021/2022

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

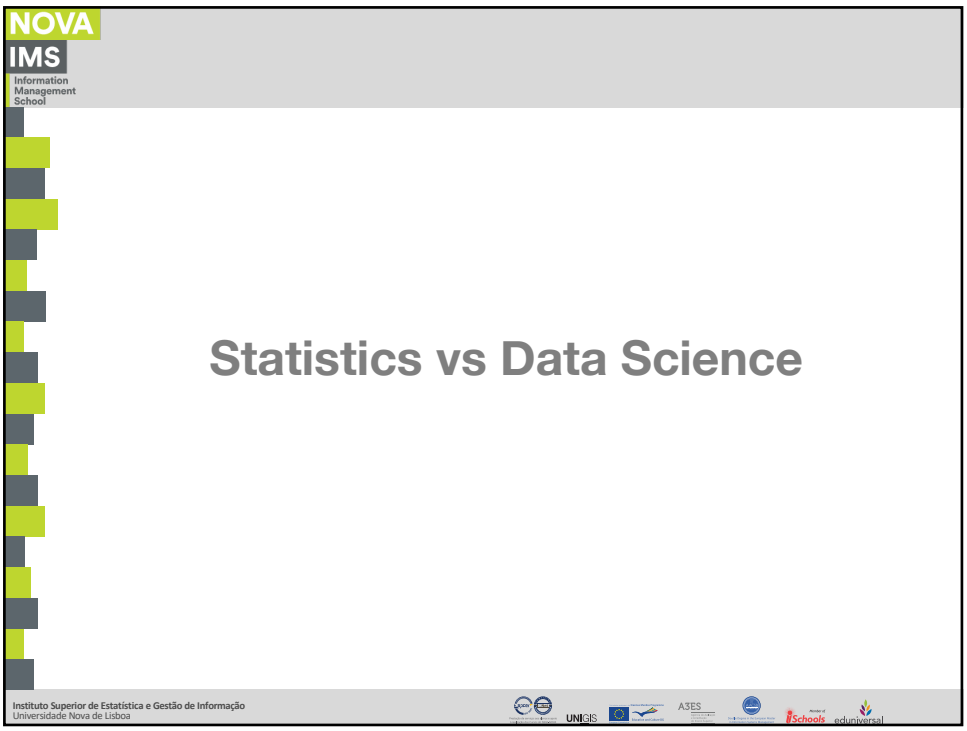
1

NOVA IMS Information Management School

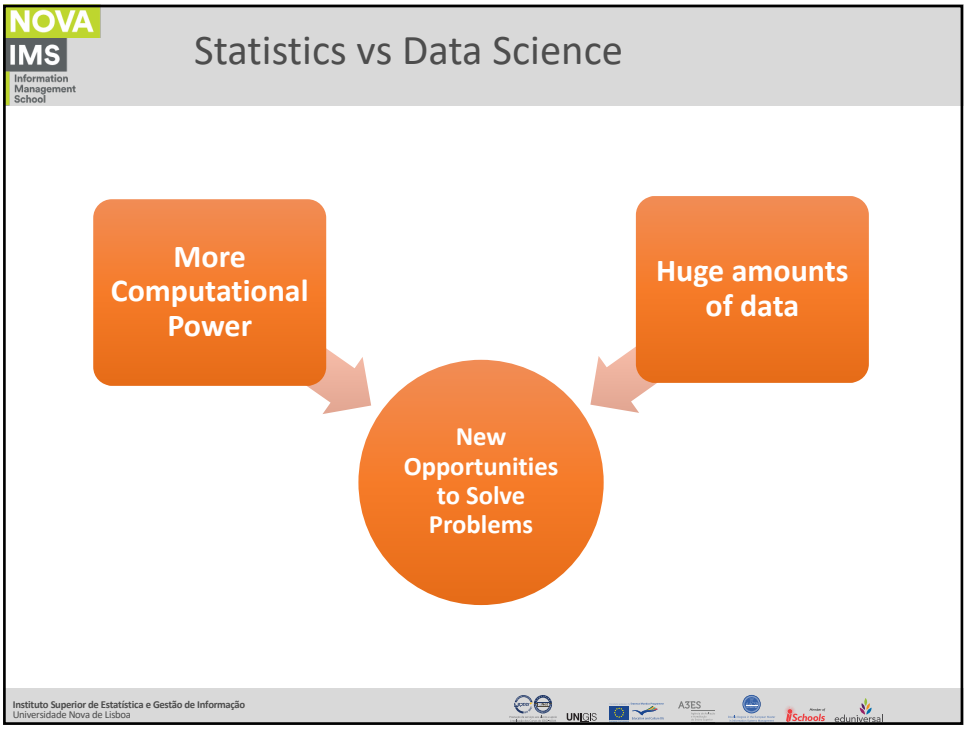
Agenda

- Data Mining
  - Statistics vs data science
  - The canonical tasks in data mining
  - Exercise
  - The data mining process
  - General aspects of problem definition
  - General aspects of data collection

2


NOVA
IMS
Information Management School
Statistics vs Data Science
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

3


NOVA
IMS
Information Management School
Statistics vs Data Science
More Computational Power
Huge amounts of data
New Opportunities to Solve Problems
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

4

What new about Data Mining?
(Data Mining: Statistics and More? David J. Hand)

- Statistics might be described as being characterized by **data sets which are small and clean**, which are **static**, which were **sampled in an iid manner**, which were often **collected to answer the particular problem** being addressed, and which are **solely numeric**.
- None of these apply in the data mining context…..

What new about Data Mining?
(Data Mining: Statistics and More? David J. Hand)

- Size of the data sets
  - Data will **not all fit into the main memory** of the computer this means that, if all of the data is to be processed during an analysis, **adaptive or sequential techniques** have to be developed (nonstatistical communities especially to those working in pattern recognition and machine learning).
  - Data sets may be large because the number of records is large or because the number of variables is large (**deep and large**).
  - Data may not be **stored as the single flat** file so but as multiple interrelated flat files.

What new about Data Mining?
(Data Mining: Statistics and More? David J. Hand)

Incremental (online)

- Examples are presented one at a time and the structure of representation changes.
- In the online learning, the system will handle each instance incrementally, the algorithm itself is updatable, and the knowledge will be updated by every instance in time.

Non incremental (batch)

- Examples are presented all at the same time and are considered together.

Incremental learning algorithms are usually faster than non-incremental algorithms, and for extremely large data sets, non-incremental algorithms may not be applicable at all.

7

What new about Data Mining?
(Data Mining: Statistics and More? David J. Hand)

- Size of the data sets
  - In the past, in many situations where statisticians have classically worked, the problem has been one of **lack of data** rather than abundance.
  - However, when data exists in the **superabundance** the results of tests (significance tests) will lead to **very strong evidence that even tiny effects exist**, effects which are so minute as to be of doubtful practical value.
  - In place of statistical significance, we need to consider more carefully substantive significance: **is the effect important or valuable or not?**

8

What new about Data Mining?
(Data Mining: Statistics and More? David J. Hand)

- Contaminated data
  - In the data mining context, when the analysis is necessarily **secondary**, data is always “dirty”.
  - When the data sets are large, it is practically certain that some of the data will be **invalid in some way**.
  - This is especially true when the **data describe human interactions** of some kind, such as marketing data, financial transaction data, or human resource data.

9

What new about Data Mining?
(Data Mining: Statistics and More? David J. Hand)

- Nonstationarity, Selection Bias, and Dependent Observations
  - Very large data sets are **unlikely to arise in an iid** manner;
  - **Population drift**, can arise because the underlying population is changing (for example, the population of applicants for bank loans may evolve as the economy heats and cools). Supermarket transactions or Telco phone calls occur every day, not just one day, so that the database is a constantly evolving entity
  - **Selection bias**, it arises when developing scoring rules, typically in this situation comprehensive data is available only for those previous applicants who were graded good risk by some previous rule. Those graded bad would have been rejected and hence their true status never discovered.

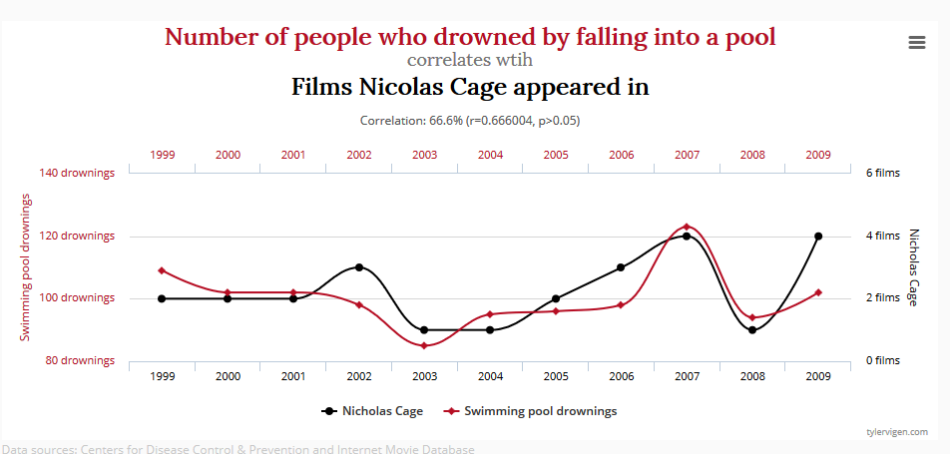
10

What new about Data Mining?
(Data Mining: Statistics and More? David J. Hand)

- Spurious Relationships and Automated Data Analysis
  - Because the pattern searches will throw up a **large numbers of candidate patterns**, there will be a **high probability that spurious** (chance) data configurations will be identified as patterns.
  - The bottom line is that those patterns and structures identified as potentially interesting will be presented to a **domain expert for consideration to be accepted or rejected** in the context of the substantive domain and objectives, and not merely on the basis of internal statistical structure.

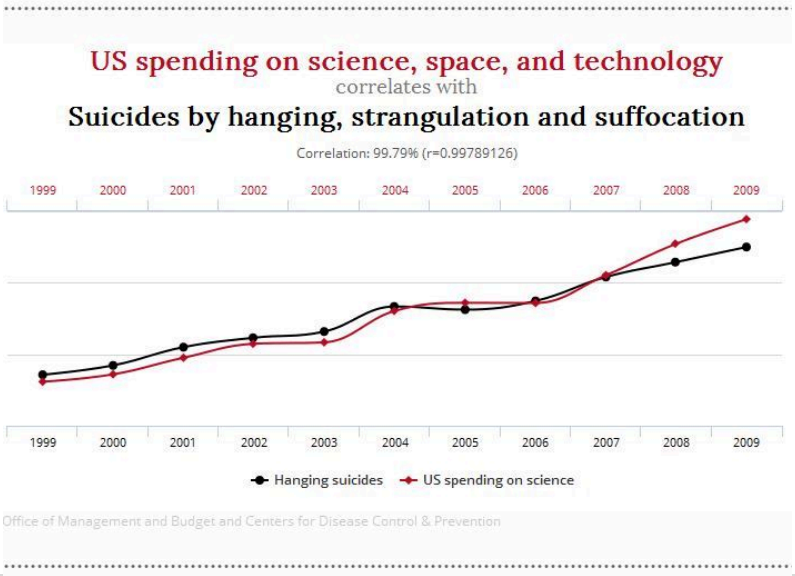
11

What new about Data Mining?
(Data Mining: Statistics and More? David J. Hand)



12

What new about Data Mining?
(Data Mining: Statistics and More? David J. Hand)



13

Statistics vs Data Science

	Experimental Primary	Opportunistic Secundary
Purpose	Research	Operational
Value	Scientific	Commercial
Origin	Controlled	Passively observed
Size	Small	Massive
Hygiene	Clean	Dirty
Status	Static	Dynamic

14

What is Data Mining?

*Data mining is a new **discipline lying at the interface** of statistics, database technology, pattern recognition, machine learning, and other areas.*

*It is concerned with **the secondary analysis of large databases** in order to find previously unsuspected relationships which are of interest or value to the database owners.*

*Data analysis is as much an **art as a science**.*

David J. Hand

15

The Canonical Tasks in Data Mining

16

Different Tasks

Data Mining

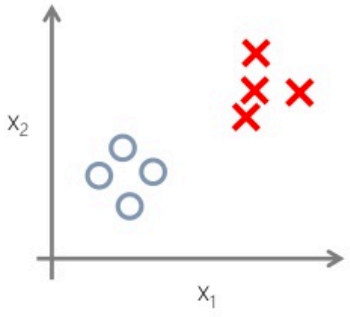
- Predictive modelling* — To learn a decision criterion that allows us to classify new and unknown examples;
- Descriptive modeling** — To describe, to provide a summary of a data set;

* aka supervised learning
** aka unsupervised learning

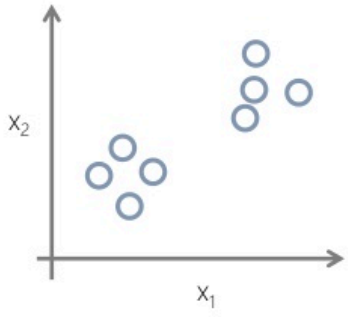
17

Different Tasks

Predictive modelling
Supervised Learning

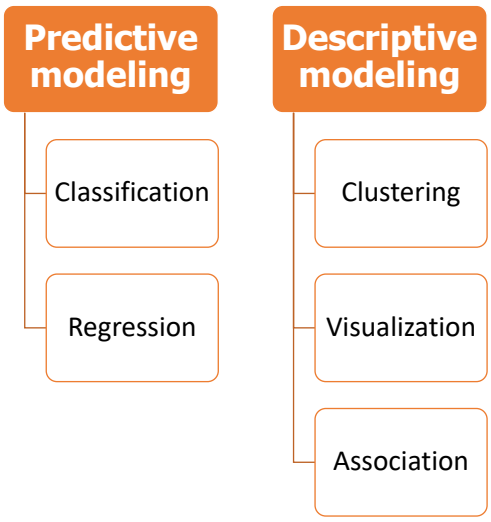


Descriptive modeling
Unsupervised Learning



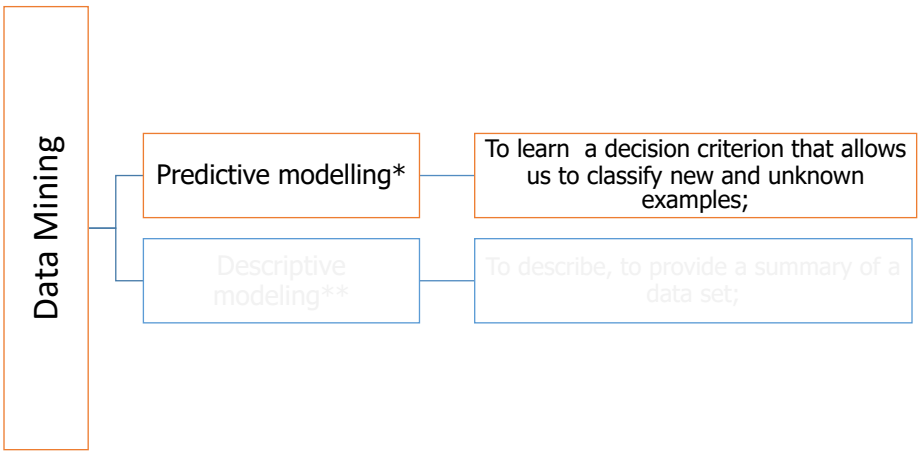
18

Different Tasks



19

Different Tasks



20

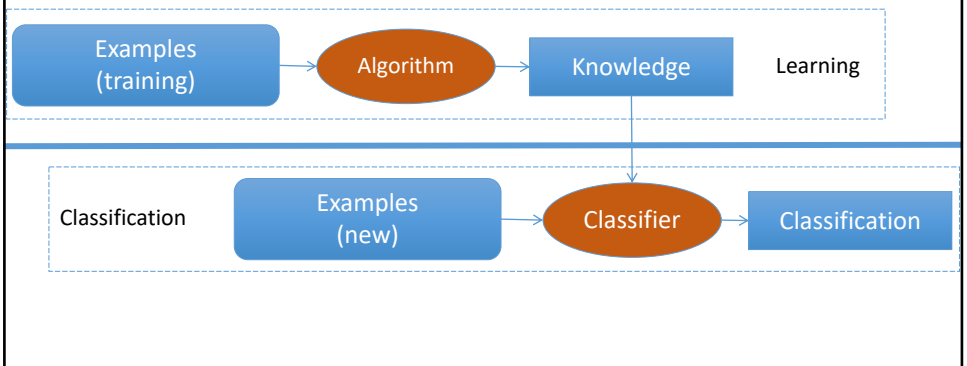
Predictive Modeling

Feature → **Label** →

Height	Weight	Sex	Age	Income	Physical Activity	Insurance Costs
1.60	79	M	41	3000	S	N
1.72	82	M	32	4000	S	N
1.66	65	F	28	2500	N	N
1.82	87	M	35	2000	N	S
1.71	66	F	42	3500	N	S

21

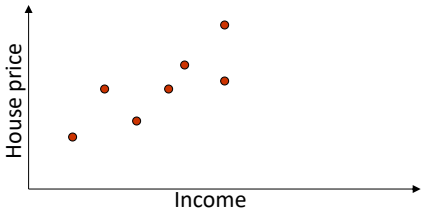
Predictive Modeling

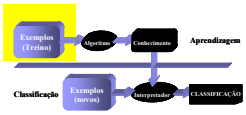


22

Predictive Modeling

- A real estate agency wants to estimate the price range for each customer based on their income;
- Training examples:
  - Historical data;
  - Income vs sold house prices.

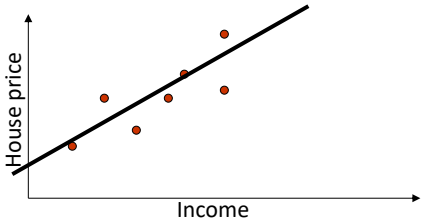


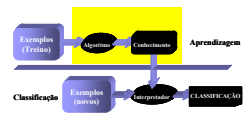


23

Predictive Modeling

- Algorithm
  - Linear regression
- Knowledge representation
  - Regression line (slope and origin)

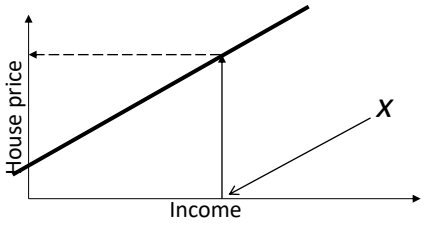


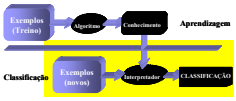


24

Predictive Modeling

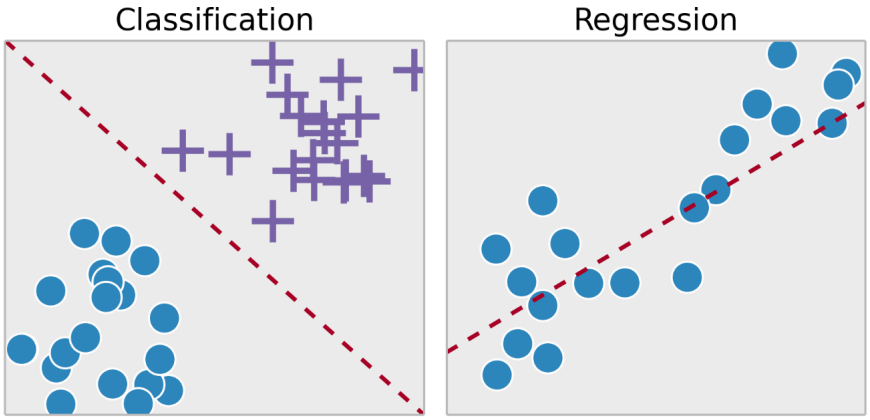
- New examples
  - A customer with an income of *x*
- Interpretation
  - Use the line (prediction method) to obtain an estimate





25

Predictive Modeling

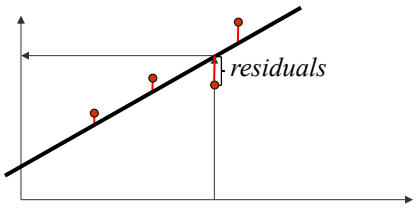


Source: http://ipython-books.github.io/featured-04/

26

Predictive Modeling

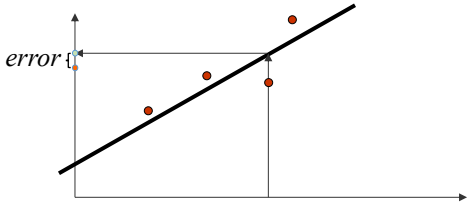
- If we are facing a regression problem, then we have to consider the average deviations produced by the model.



27

Predictive Modeling

- If we are facing a regression problem, then we have to consider the average deviations produced by the model.



28

Predictive Modeling

- If we are facing a classification problem, then we only need to count the number of times that the model was wrong;

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

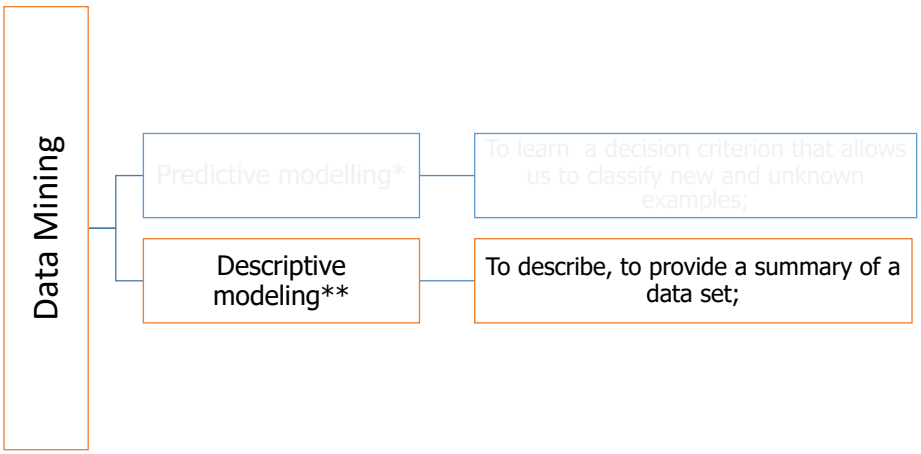
$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

29

Different Tasks



30

Different Tasks

Feature

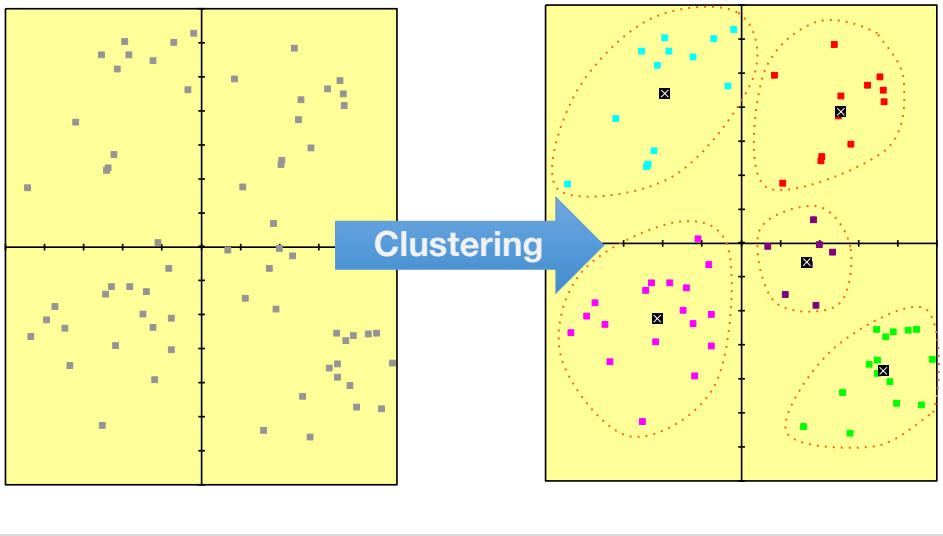
Height	Weight	Sex	Age	Income	Physical Activity
1.60	79	M	41	3000	S
1.72	82	M	32	4000	S
1.66	65	F	28	2500	N
1.82	87	M	35	2000	N
1.71	66	F	42	3500	N

31

Clustering

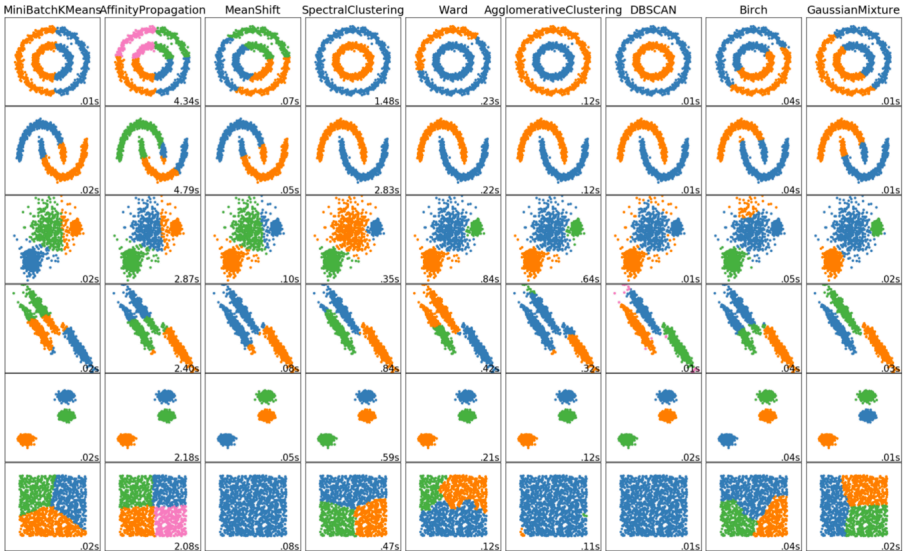
32

Different Tasks



33

Different Tasks



34

NOVA IMS Information Management School

Association Rules

35

NOVA IMS Information Management School

Different Tasks

Transaction Table

1,000,000 Total Transactions

200,000 Shoes

50,000 Socks

20,000 Shoes and Socks

Rule

If a customer purchases shoes, then 10% of the time he or she will purchase socks.

Evaluation Criteria:

Confidence: 20,000/200,000 = 10%

Support 20,000/1,000,000 =2%

Expected Confidence= 50,000/1,000,000 = 5%

Lift= Confidence/Expected Confidence =2

Note: The confidence factor with socks on the left-hand side and shoes on the right-hand side is 40% (20,000/50,000).

The lift value of two implies that you are twice as likely to buy socks if you bought shoes than if you did not buy shoes.

Figure 1. Association Discovery Statistics Example

36

NOVA IMS Information Management School

Visualization

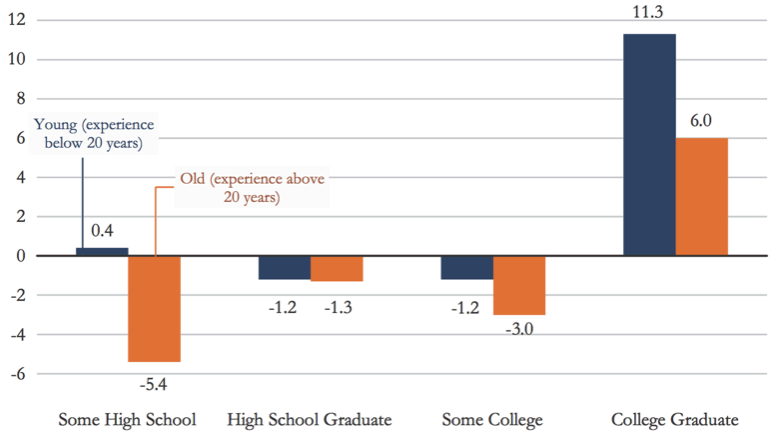
37

NOVA IMS Information Management School

Different Tasks

Flattening a 3D Chart

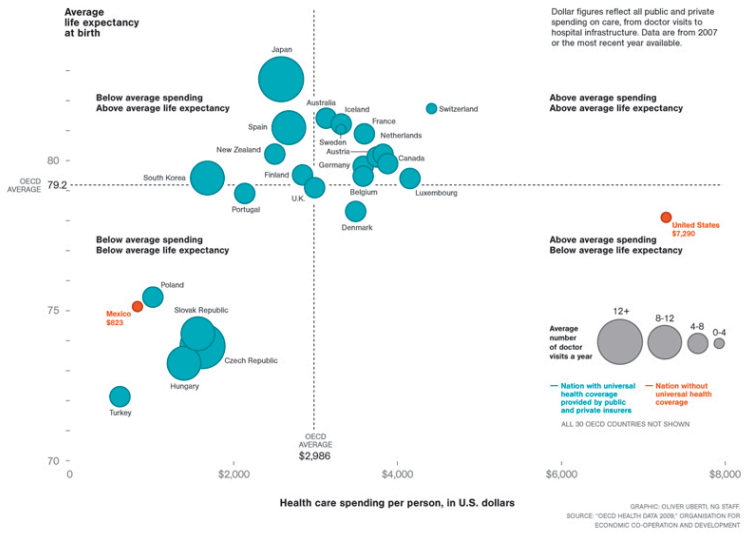
Change in real weekly wages of US-born workers by group, 1990–2006
(Percent)



38

NOVA IMS Information Management School

Different Tasks



39

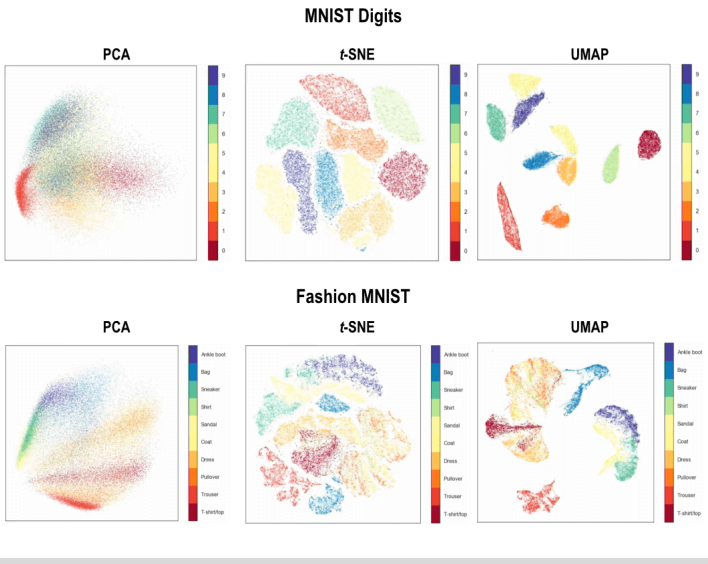
NOVA IMS Information Management School

Different Tasks

40

NOVA IMS Information Management School

Different Tasks



41

NOVA IMS Information Management School

Let's do a predictive model!!!

42

43

NOVA IMS Information Management School

Data Mining

S3

NOVA-IMS 2021/2022

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

Agenda

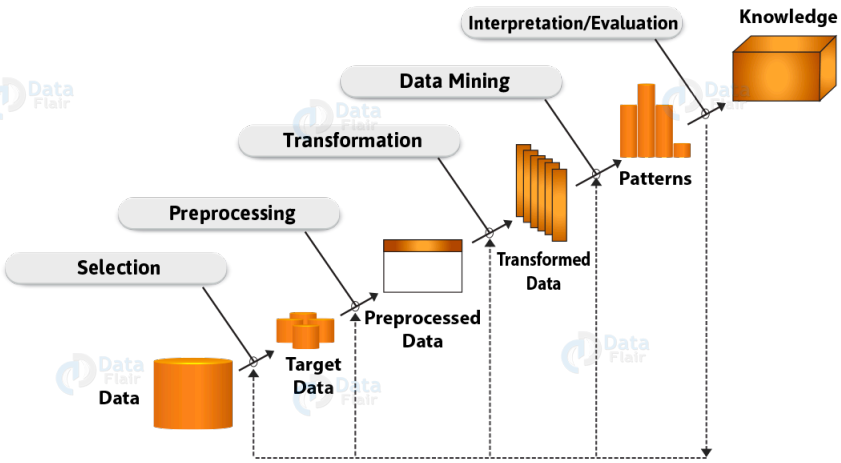
- Data Mining
  - The data mining process
  - General aspects of problem definition
  - Input space
  - The curse of dimensionality
  - Input space coverage
  - Binary and multiclass classification
  - Separability and Bayes error
  - Different types of variables
  - Spurious correlations and confounding variables
  - Performance evaluation

2

The Data Mining Process

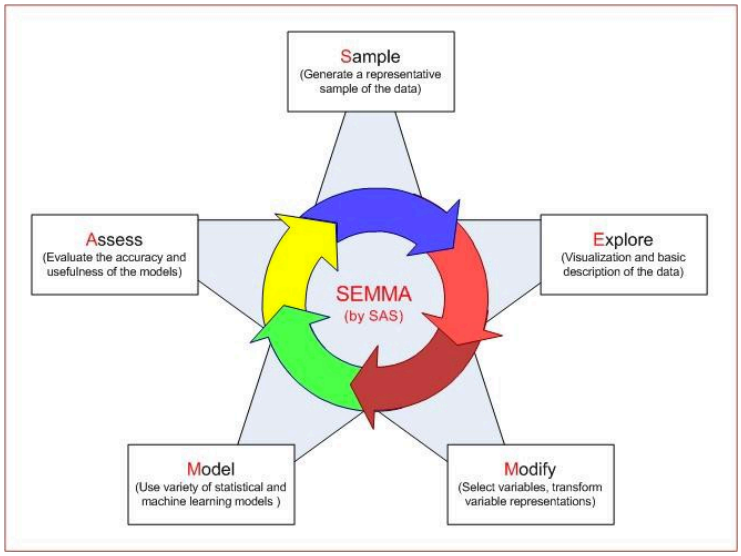
3

Data Mining and Knowledge Discovery Database (KDD) Process



4

Sample, Explore, Modify, Model, and Assess (SEMMA)



5

CRISP-DM, Cross Industry Standard Process for Data Mining

6

General aspects of the problem definition

7

Problem definition

- "We're doomed to complex theories that will never have the **elegance of physics equations**. But if that's so, we should stop acting as if our goal is to author extremely elegant theories, and instead **embrace complexity** and make use of the best ally we have: **the unreasonable effectiveness of data**."
- Invariably, **simple models** and a **lot of data** trump more elaborate models based on less data.

8

Problem definition

- So, follow the data. Choose a representation that can use **unsupervised learning** on **unlabeled data**, which is so **much more plentiful than labeled data**.
- Represent all the data with a nonparametric model rather than trying to summarize it with a parametric model: **"let the data speak for themselves"**

9

Problem definition

- Suppose you've constructed the **best set of features** you can, but the **classifiers you're getting are still not accurate enough**.
- What can you do now? There are **two main choices**:
  - design a **better learning algorithm**,
  - or **gather more data** (more examples, and possibly more raw features, subject to the curse of dimensionality).
- Machine learning researchers are mainly concerned with the former, but **pragmatically the quickest path to success is often to just get more data**.
- As a rule of thumb, **a dumb algorithm with lots and lots of data beats a clever one with modest amounts of it**. (After all, machine learning is all about **letting data do the heavy lifting**.)

10

NOVA IMS Information Management School

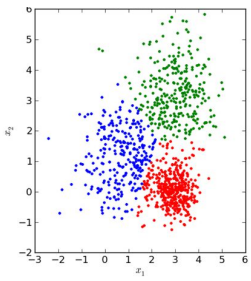
Input Space

11

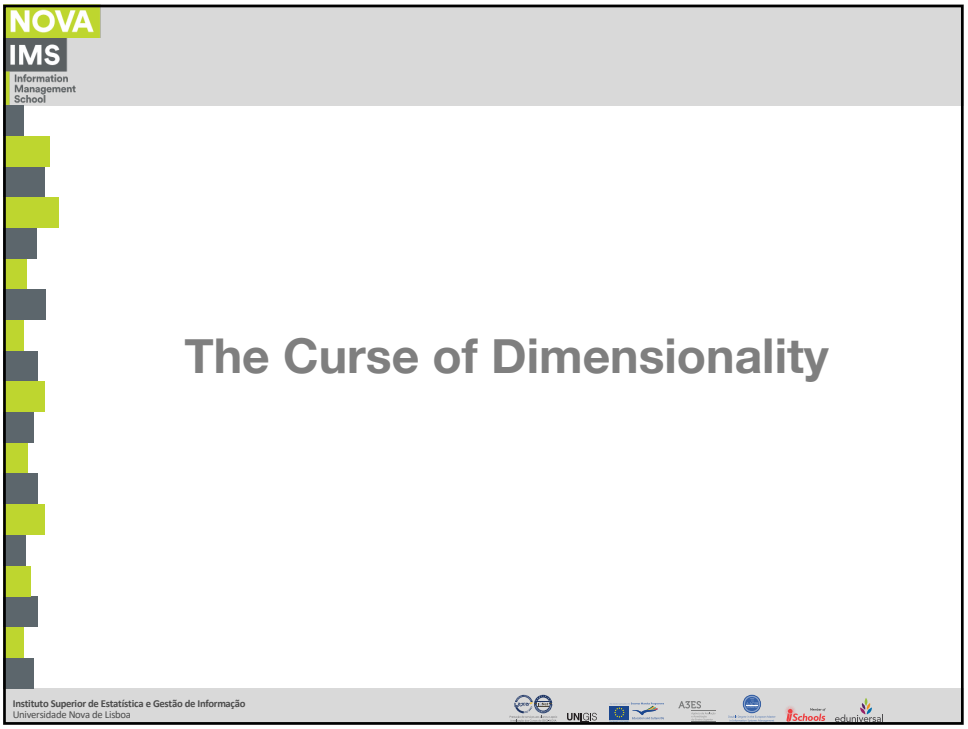
NOVA IMS Information Management School

Problem definition

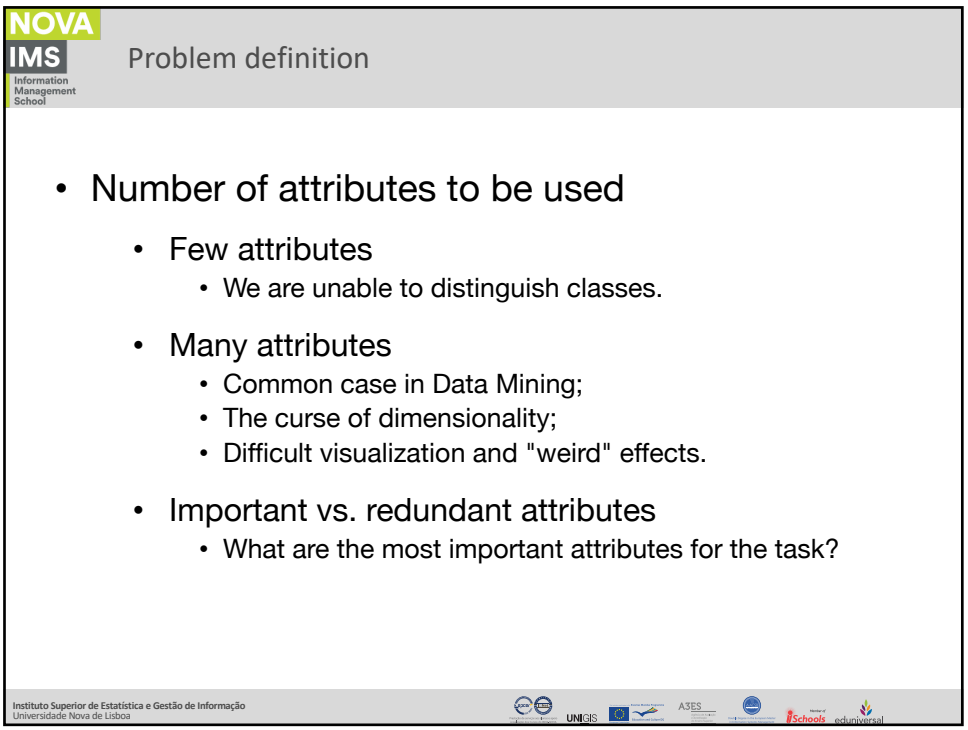
- **Input Space**
  - The input space is defined by the input feature vectors.
  - Where the algorithms will try to find a solution to the problem



12


NOVA
IMS
Information Management School
The Curse of Dimensionality
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

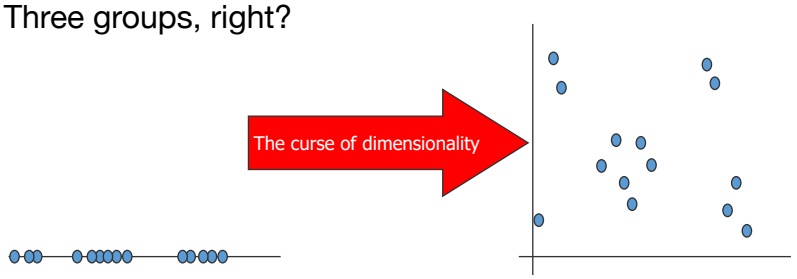
13


NOVA
IMS
Information Management School
Problem definition
• Number of attributes to be used
• Few attributes
• We are unable to distinguish classes.
• Many attributes
• Common case in Data Mining;
• The curse of dimensionality;
• Difficult visualization and "weird" effects.
• Important vs. redundant attributes
• What are the most important attributes for the task?
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

14

Problem definition

Three groups, right?

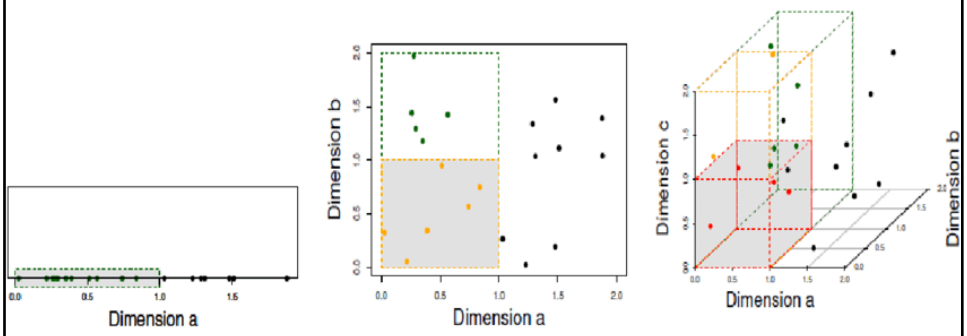


Well… not exactly.

When the dimensionality increases, the space becomes more sparse and it becomes more difficult to find groups (you need even more data)

15

Problem definition



16

Problem definition

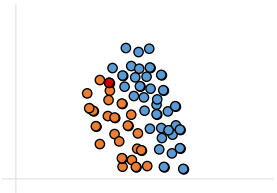
- The curse of dimensionality
  - **Generalizing correctly becomes exponentially harder** as the dimensionality (number of features) of the examples grows, because a fixed-size training set covers a dwindling fraction of the input space.
  - With a **dimension of 100 and a huge training set of a trillion examples**, the latter covers only a fraction of about 10^{-18} of the input space. This is what makes machine learning both necessary and hard.

17

Input Space Coverage

18

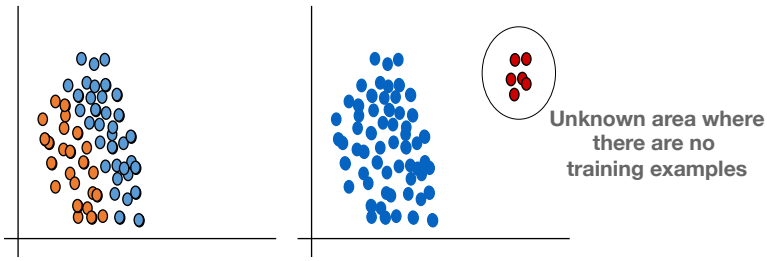
Problem definition



- Good coverage of the problem space increases confidence in the results and in its quality.

19

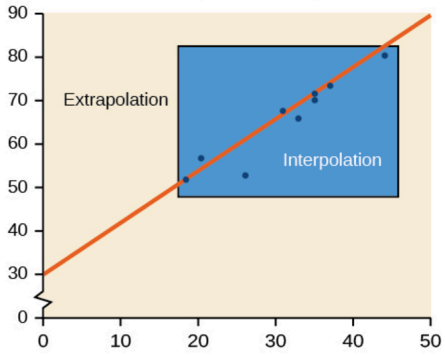
Space coverage



- If a model is developed based on a set of examples, but in fact the examples would be very different, it is natural that the results will be bad (@men women photo).

20

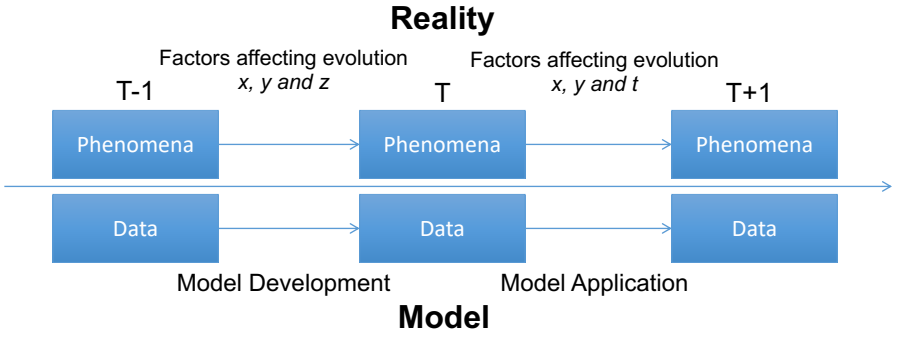
Extrapolation vs Interpolation



- **Interpolation** involves predicting a value inside the domain and/or range of the data.
- **Extrapolation** involves predicting a value outside the domain and/or range of the data.

21

General aspects of data collection



22

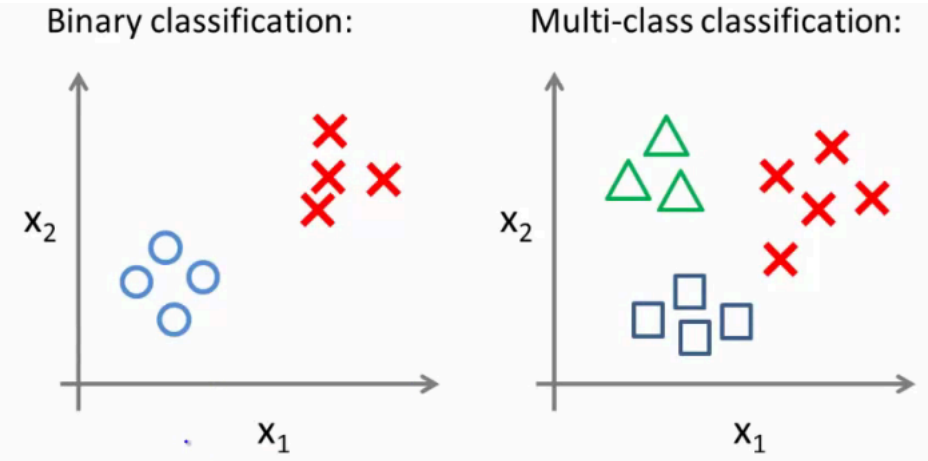
NOVA IMS Information Management School

Binary and Multiclass Classification

23

NOVA IMS Information Management School

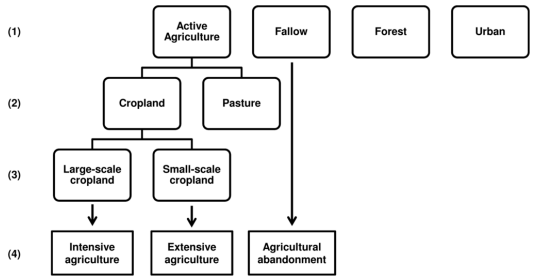
Binary and Multi-class



24

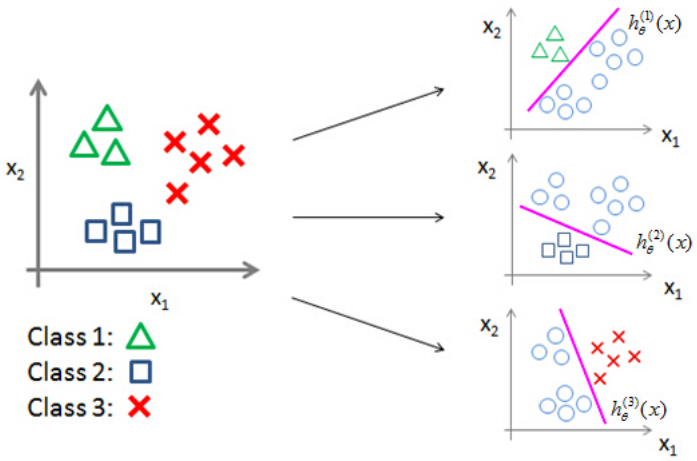
Granularity

- **Granularity of classification problems:**
  - Initially a small number of output classes;
  - The more output classes, the greater the number of data needed for training.



25

Multi-class strategies

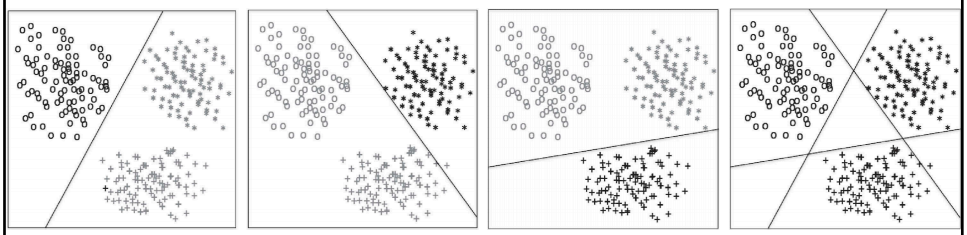


26

NOVA IMS Information Management School

Multi-class strategies

- **One-vs-rest (OVR) strategy:**
  - Breaks the multi-class classification down into a series of binary classification
  - N-class classification problem is decomposed into N binary classification problems.

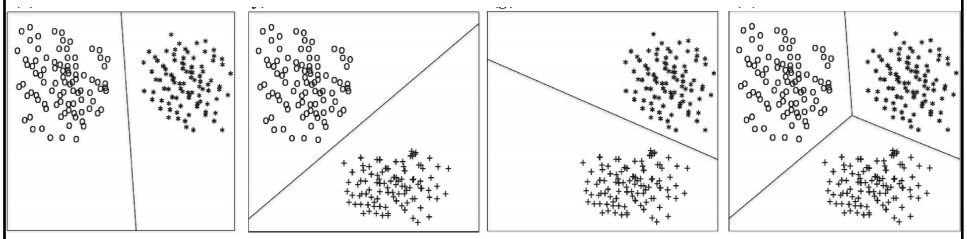


27

NOVA IMS Information Management School

Multi-class strategies

- **One-vs-one (OVO) strategy:**
  - Enumerating all possible pairs of classes and then to develop a binary classifier for each pair of classes
  - Classification is then done by inputting the data point into each particular binary classifier and labelling by majority voting 1/2N(N-1)



28

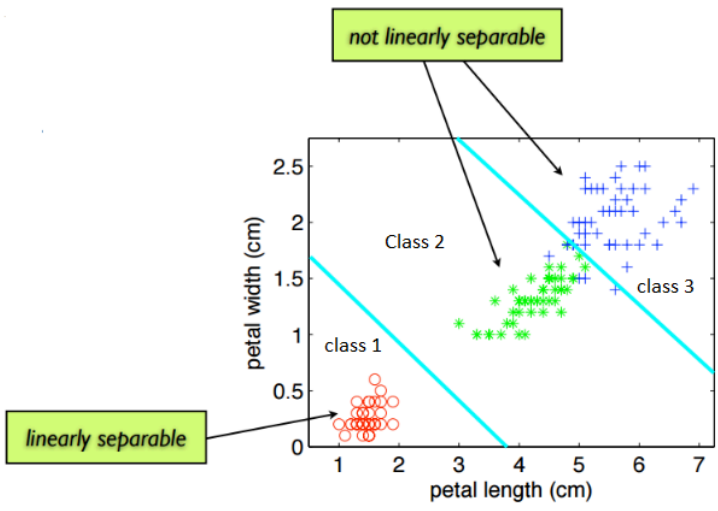
NOVA IMS Information Management School

Separability and Bayes Error

29

NOVA IMS Information Management School

Separation and error



30

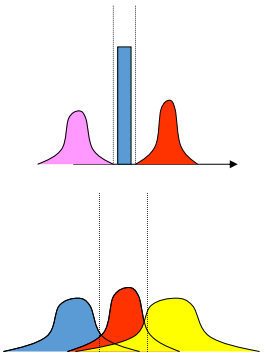
Problem definition

Separable

- Ø error possible

Not separable

- Always error > Ø
- Bayes error
  - Lowest possible error for a classifier



31

Different Types of Variables

32

Variables

- **Different types of variables:**
  - **Nominal** are just labels, e.g. ‘red’, ‘green’, ‘blue’, no particular order. Think in classes.
  - **Ordinal** have an order, e.g. ‘satisfied’, ‘very satisfied’, ‘extremely satisfied’. Think in ranks.
  - **Discrete** are just counting data, e.g. 0, 1, 2, ...
  - **Continuous** are just measurement data, e.g. 1.23, 0.001, etc

33

Variables

- **Different types of variables:**
  - **Interval** data are measured and have constant, equal distances between values, but the zero point is arbitrary. The zero isn’t meaningful, it doesn’t mean a true absence of something.
  - When a **ratio** between two values of a quantitative variable is meaningful, it's a ratio scaled variable. Ratio measurement assumes a zero point where there is no measurement.

34

Problem definition

Provides:	Nominal	Ordinal	Interval	Ratio
The "order" of values is known		✔	✔	✔
"Counts," aka "Frequency of Distribution"	✔	✔	✔	✔
Mode	✔	✔	✔	✔
Median		✔	✔	✔
Mean			✔	✔
Can quantify the difference between each value			✔	✔
Can add or subtract values			✔	✔
Can multiple and divide values				✔
Has "true zero"				✔

Summary of data types and scale measures

35

Metadata

- **Metadata:**
  - Metadata is "data [information] that provides information about other data".
  - Three distinct types of metadata exist:
    - Descriptive metadata describes a resource for purposes such as discovery and identification.
    - Structural metadata is metadata about containers of data and indicates how compound objects are put together, for example, how pages are ordered to form chapters.
    - Administrative metadata provides information to help manage a resource, such as when and how it was created, file type and other technical information, and who can access it.

36

NOVA IMS Information Management School

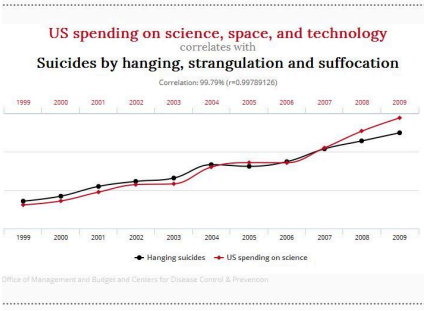
Spurious Correlations and Confounding Variables

37

NOVA IMS Information Management School

Problem definition

- **Input variables should be causally related to the outputs**
  - Spurious correlations
    - Low number of training examples;
    - Large number of input variables.
  - It is important that there is a plausible reason to chose the input variables.



38

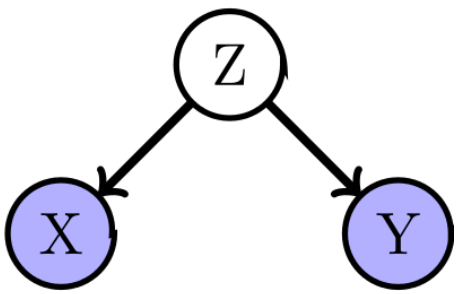
Causality and correlation

- **Input variables should be causally related to the outputs**
  - Confounding variables
    - In statistics, a confounding variable (also confounding factor) is an **extraneous variable** in a statistical model that **correlates** (directly or inversely) **with both the dependent variable and the independent variable**.
    - A **spurious relationship** is a perceived relationship between an independent variable and a dependent variable that has been **estimated incorrectly** because the estimate **fails to account for a confounding factor**.

39

Problem definition

- **Input variables must be causally related to the outputs**
  - Confounding variables



40

NOVA IMS Information Management School

Causality and correlation

- **Spurious correlations**
  - Example n.º 1: Ice cream sales and the number of drowning's;
  - Example n.º 2: Correlation between the measuring of a patient's temperature on admission to hospital and the probability of his survival;

41

NOVA IMS Information Management School

Performance Evaluation

42

NOVA IMS Information Management School

Problem definition

- **Result Evaluation:**

		Actual Value: Positives	Actual Value: Negatives
Predicted Value	Positives	TP True Positives	FP False Positives
Predicted Value	Negatives	FN False Negatives	TN True Negatives

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Population}$$

$$Specificity\ (true\ negative\ rate) = \frac{True\ Negatives}{Total\ Negatives}$$

43

NOVA IMS Information Management School

Problem definition

- **Result Evaluation:**

		Actual Value: Positives	Actual Value: Negatives
Predicted Value	Positives	TP True Positives	FP False Positives
Predicted Value	Negatives	FN False Negatives	TN True Negatives

$$Recall = \frac{True\ Positives}{Total\ Positives}$$

$$Precision = \frac{True\ Positives}{Predicted\ Positives}$$

44

NOVA IMS Information Management School

Problem definition

- **Result Evaluation:**

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

$$R^2 = 1 - \frac{MSE(model)}{MSE(baseline)}$$

$$MSE(baseline) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2$$

45

NOVA IMS Information Management School

Problem definition

- **Result Evaluation:**
  - What is the level of accuracy required to consider the application a success?
  - How to compare the quality of an obtained solution?
  - What are the existing alternatives that can serve as a standard of comparison?
  - What type of data to use to evaluate the various models?

46

47

Data Mining

S4

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

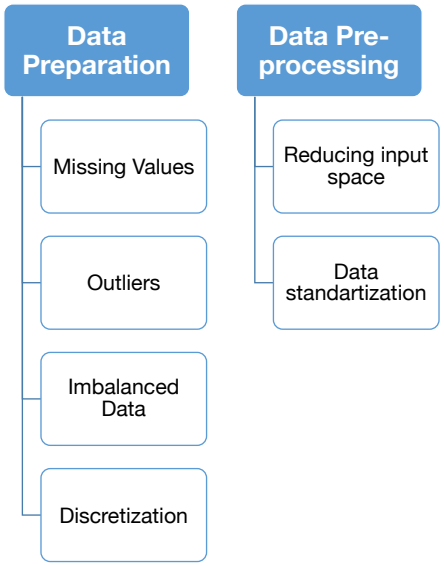
1

Data Preparation and Pre-processing

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

2

Data Preparation



3

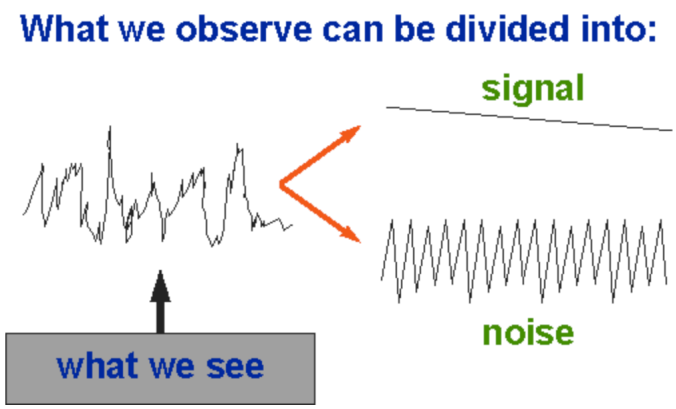
Data Preparation

- **Objective**
  - Transform data sets to **best expose its information content**
  - The **quality** of the models **should be better** (or at least the same) after preparation
  - **Good data** is a prerequisite for good models
  - Some techniques are theoretically based, other are just based on experience

4

- **Signal vs. Noise**
  - In science (physics and telecommunications) noise is defined as **fluctuations and external disturbances** in the flow of information (signal) received;
  - An **undesired disturbance** in relevant information;
  - A disturbance that affects a signal and that may distort the information carried by the signal.

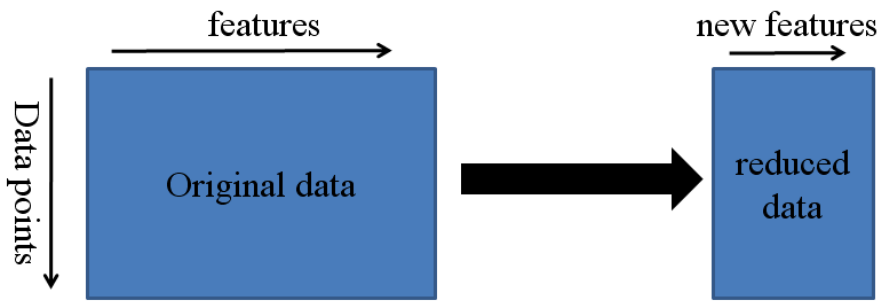
5



6

NOVA IMS Information Management School

Data Preparation



7

NOVA IMS Information Management School

Data Preparation

- **Relevance**
  - Data preparation takes **60 to 80 percent of the whole analytical pipeline** in a typical machine learning / deep learning project.
  - Data Preparation and pre-processing **puts data into the right shape and quality for analysis**. It includes many different functions, for example the following:

8

NOVA IMS Information Management School

Data Preparation

- **Real data suffers from several problems**
- Incomplete
  - Missing values, lacking attributes of interest, levels of aggregation
- Noisy
  - Errors and outliers
- Inconsistent
  - E.g. Age=42 Birthday=31/07/1997
  - Changes in scales
  - Duplicate records with different values

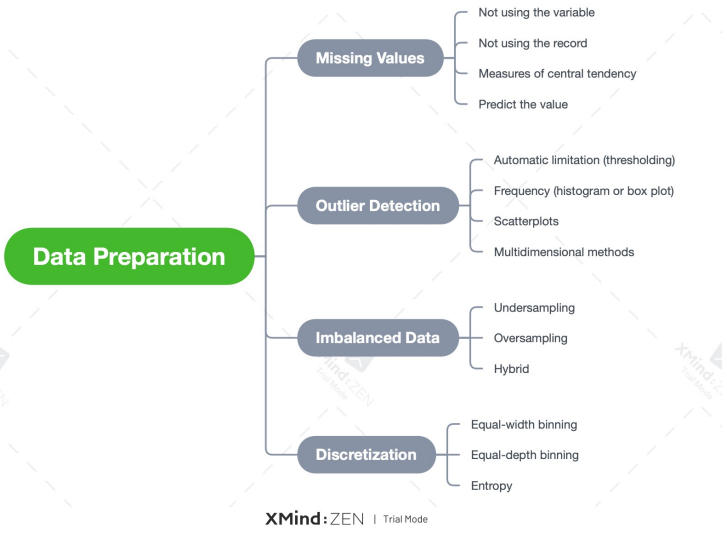
9

NOVA IMS Information Management School

Data Preparation

10

Data Preparation



11

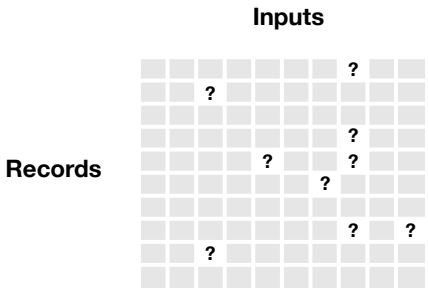
Treatment of Missing Data

12

NOVA IMS Information Management School

Data Preparation

- **Missing Data**

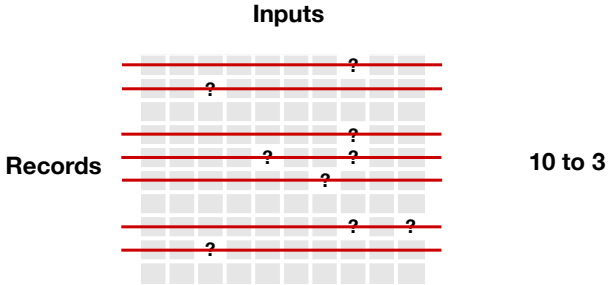


13

NOVA IMS Information Management School

Data Preparation

- **Missing Data**

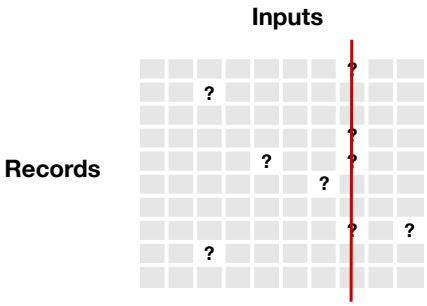


14

NOVA IMS Information Management School

Data Preparation

- **Missing Data**



15

NOVA IMS Information Management School

Data Preparation

OS	Revenue
Android	1,804
iOS	3,027
iOS	8,788
Android	NA
Android	3,735
Android	1,056
iOS	9,319
Android	6,199
Android	2,235
iOS	NA
Android	1,146

16

Data Preparation

OS	Revenue
Android	1,804
iOS	3,027
iOS	8,788
Android	NA
Android	3,735
Android	1,056
iOS	9,319
Android	6,199
Android	2,235
iOS	NA
Android	1,146

OS	Global Mean
Android	1,804
iOS	3,027
iOS	8,788
Android	4,145
Android	3,735
Android	1,056
iOS	9,319
Android	6,199
Android	2,235
iOS	4,145
Android	1,146

17

Data Preparation

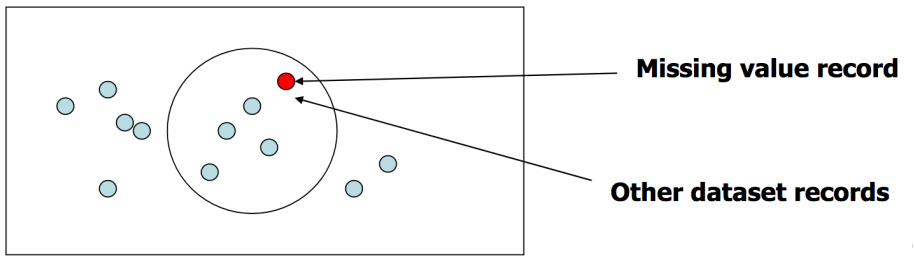
OS	Revenue
Android	1,804
iOS	3,027
iOS	8,788
Android	NA
Android	3,735
Android	1,056
iOS	9,319
Android	6,199
Android	2,235
iOS	NA
Android	1,146

OS	Global Mean	Group Mean
Android	1,804	1,804
iOS	3,027	3,027
iOS	8,788	8,788
Android	4,145	2,696
Android	3,735	3,735
Android	1,056	1,056
iOS	9,319	9,319
Android	6,199	6,199
Android	2,235	2,235
iOS	4,145	7,045
Android	1,146	1,146

18

NOVA IMS Information Management School

Data Preparation

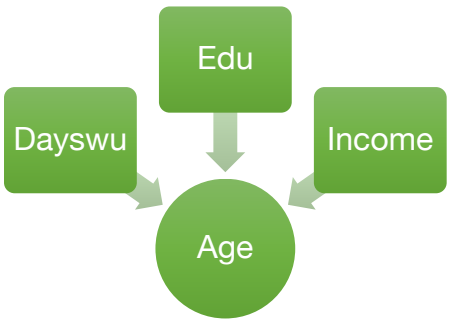


19

NOVA IMS Information Management School

Data Preparation

CUSTID	DAYSWUS	AGE	EDU	INCOME
1138	951	29	17	55637
6313	1231	22	16	32795
9185	1186	31	18	40717
7735	1097	27	18	29184
7918	1196	19	13	16479
1585	849	34	19	55403
7897	1120	46	15	63603
7389	792	39	16	57402
6764	1060	missing	15	69803
8416	1217	31	18	56028
6541	1074	54	15	91934
8263	1185	46	18	78262
8052	1007	52	16	91292
2448	840	51	14	62699
10978	553	44	19	62675
1409	1191	missing	14	64449
6063	1222	32	18	55048
9767	961	34	20	53217
4489	1224	69	16	103191
10738	1190	missing	18	64412
10381	939	54	16	73709
8999	1054	45	15	73074
4482	1089	50	19	63983
4257	1074	missing	20	75006
7722	876	50	19	55332
7491	1165	49	14	63150
1037	1127	47	16	69713
10419	1112	48	15	74961
5124	1217	47	17	72890
8140	1229	54	16	74670
3218	1159	54	15	83956
4124	1247	53	15	81937
8990	1025	54	17	79364
7155	1235	54	18	81948
6346	1170	42	17	71182
9547	1125	missing	16	81442
8250	989	12	16	75566
2527	1234	51	15	77635
6663	1185	40	14	61974



20

NOVA IMS Information Management School

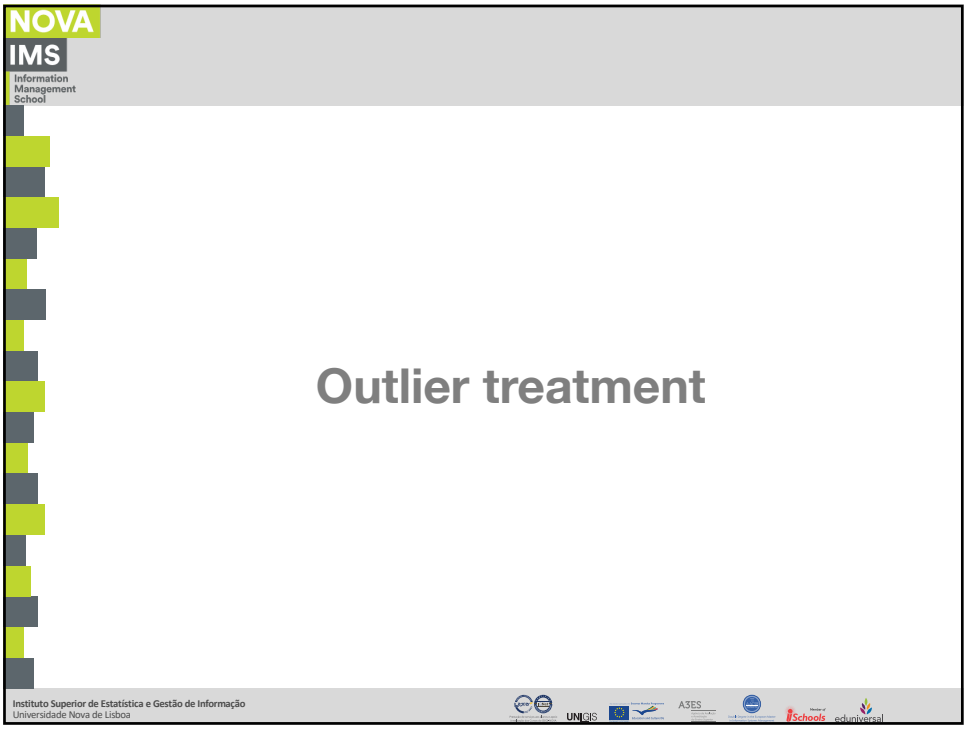
Data Preparation

- **Missing data**
  - Delete variables (you loose information)
  - Delete records (potential bias);
  - To examine and manually enter a probable value (tedious + infeasible);
  - Automatically fill in with a measure of central tendency (i.e. mean, median, mode);
  - Automatically fill in with a measure of central tendency of a subset (e.g. men and women);
  - To fill in with values from similar individuals (nearest neighbours);
  - Predictive model (linear regression, multiple linear regression);
  - Code the missing data explicitly.

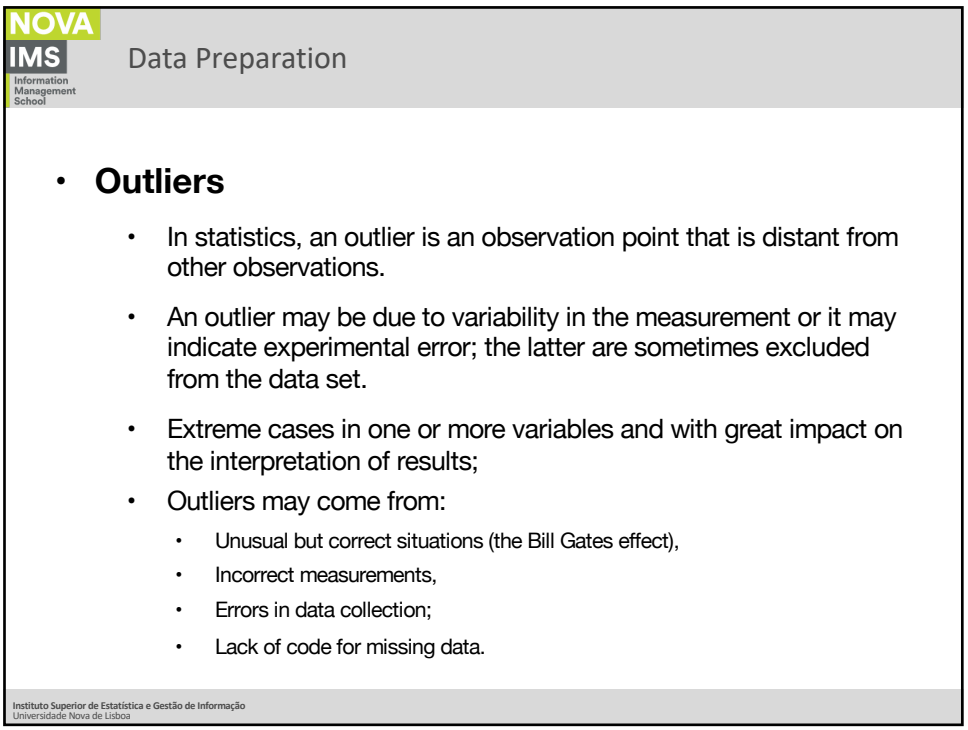
NOVA IMS Information Management School

Data Preparation

- **Missing data**
  - The most practical approach to the problem is to initially use the **quickest and simplest option**;
  - After achieving some **preliminary results** we can comparatively analyze the performance of the model in the full sample patterns an in those where there was a need to estimate missing values;
  - In the event that the **error is significantly higher** than in other data, then we will try to use another method in order to improve results.


NOVA
IMS
Information Management School
Outlier treatment
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

23

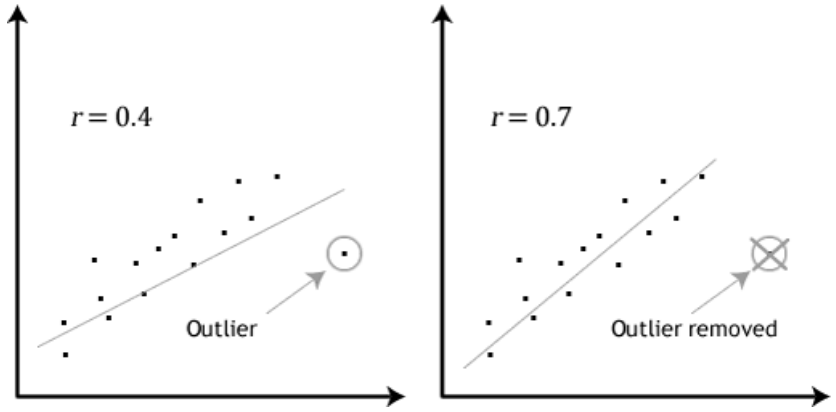

NOVA
IMS
Information Management School
Data Preparation
Outliers
In statistics, an outlier is an observation point that is distant from other observations.
An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set.
Extreme cases in one or more variables and with great impact on the interpretation of results;
Outliers may come from:
Unusual but correct situations (the Bill Gates effect),
Incorrect measurements,
Errors in data collection;
Lack of code for missing data.
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

24

NOVA IMS Information Management School

Data Preparation

- **Outliers (leverage effect)**

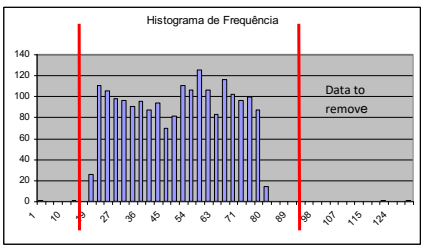


25

NOVA IMS Information Management School

Data Preparation

- **Remove Data Outliers**
  - Automatic limitation (thresholding)
    - The imposition of maximum and minimum values for the variables (age – 0 e 100)

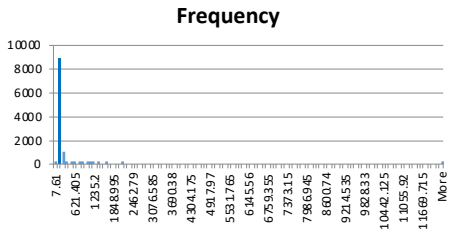


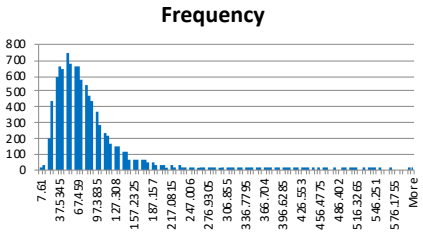
26

NOVA IMS Information Management School

Data Preparation

- **Remove Data Outliers**

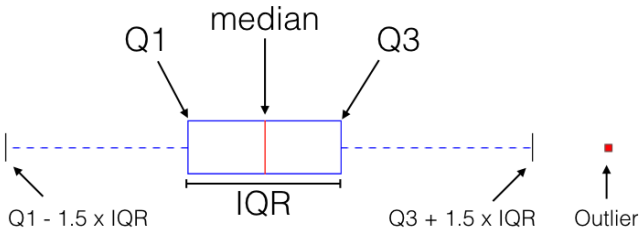




27

NOVA IMS Information Management School

Data Preparation



Q1: *Quartile 1*, or median of the *left* data subset
after dividing the original data set into 2 subsets via the median
(25% of the data points fall below this threshold)

Q3: *Quartile 3*, median of the *right* data subset
(75% of the data points fall below this threshold)

IQR: *Interquartile-range*, Q3 - Q1

Outliers: Data points are considered to be outliers if
value < *Q1 - 1.5 x IQR* or
value > Q3 + 1.5 x IQR

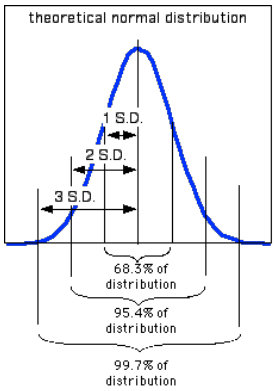
fmi:
http://sebastianraschka.com/Articles/2014_dixon_test.html
http://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm

28

NOVA IMS Information Management School

Data Preparation

- **Remove Data Outliers**
  - Normal data distribution
    - 3σ +/- Average

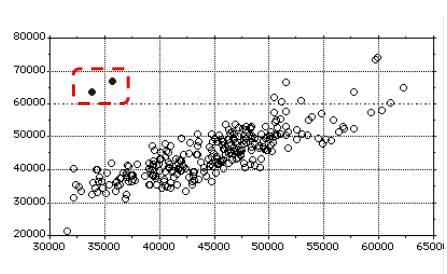


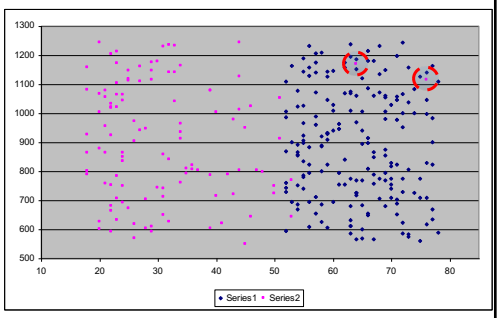
29

NOVA IMS Information Management School

Data Preparation

- **Remove Data Outliers**
  - In two dimensions…





30

NOVA IMS Information Management School

Data Preparation

- **Remove Data Outliers**
  - Cluster Analysis (K-means)
  - Self-Organizing Maps

31

NOVA IMS Information Management School

Imbalanced Learning

32

NOVA IMS Information Management School

Quick quizz:

	Yes	No
Is it possible to have a useless classifier with 99% accuracy?		
Is it possible to achieve a 99.9% accuracy with a trivial classifier?		

33

NOVA IMS Information Management School

Quick quizz:

	Yes	No
Is it possible to have a useless classifier with 99% accuracy?	X	
Is it possible to achieve a 99.9% accuracy with a trivial classifier?	X	

34

NOVA IMS Information Management School

Quick quizz:

	Yes	No
Is it possible to have a useless classifier with 99% accuracy? If the minority class is 1%	X	
Is it possible to achieve a 99.9% accuracy with a trivial classifier? If the minority class is 0,1%	X	

35

NOVA IMS Information Management School

- Class Imbalance
  - Is this frequent in real-world applications?
    - **Credit Card** frauds - **~2%** per year.
    - **HIV prevalence** in the USA - **~0.4%**.
    - **Disk drive** failures - **~1% per year**.
    - **Factory production** defects - **~ 0.1%**.
    - **Business churn** - **~3%**

36

Imbalanced Learning

- Imbalanced Learning
  - An **imbalanced learning** problem is defined as a classification task for binary or multi-class datasets where a **significant asymmetry** exists between the **number of instances for the various classes**.
  - The dominant class is called the **majority class (negative cases)** while the rest of the classes are called the **minority classes (positive cases)**
  - The **Imbalance Ratio** (IR), is the ratio between the majority class and the minority class, (depends on the type of application and for binary problems values between 100 and 100.000 have been observed)

37

Imbalanced Learning

- Imbalanced Learning
  - **Standard learning methods induce a bias** in favor of the majority class during training.
  - This happens because the **minority classes contribute less to the maximization** of the objective function, which is usually accuracy.

$$Accuracy = \frac{TruePositive + TrueNegative}{TotalPopulation}$$

38

Imbalanced Learning

- Imbalanced Learning
  - **Standard learning methods induce a bias** in favor of the majority class during training.
  - This happens because the **minority classes contribute less to the maximization** of the objective function, which is usually accuracy.

		Actual Value: Positives	Actual Value: Negatives
Predicted Value	Positives	TP True Positives	FP False Positives
Predicted Value	Negatives	FN False Negatives	TN True Negatives

39

Imbalanced Learning

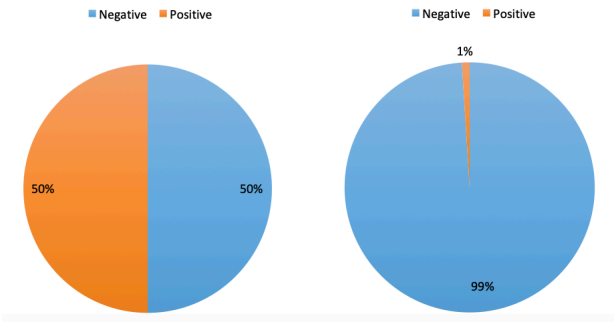
- Imbalanced Learning
  - **Standard learning methods induce a bias** in favor of the majority class during training.
  - This happens because the **minority classes contribute less to the maximization** of the objective function, which is usually accuracy.

		Actual Value: Positives	Actual Value: Negatives
Predicted Value	Positives	TP True Positives	FP False Positives
Predicted Value	Negatives	FN False Negatives	**TN True Negatives**

40

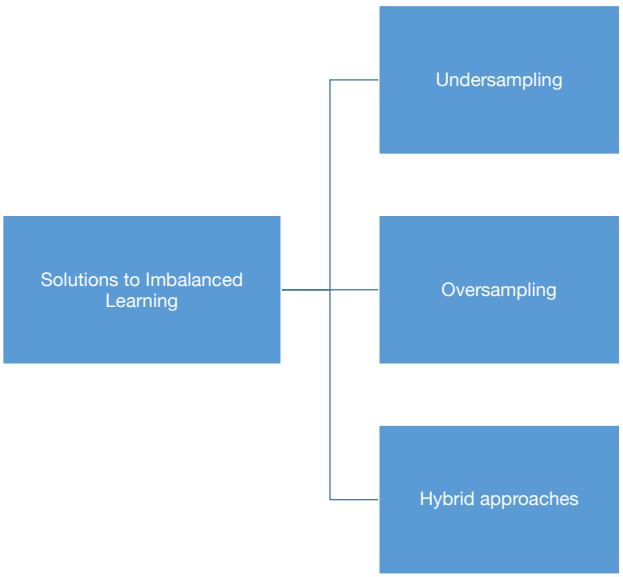
Imbalanced Learning

- Imbalanced Learning
  - By optimizing classication accuracy, **most algorithms assume a balanced class distribution**



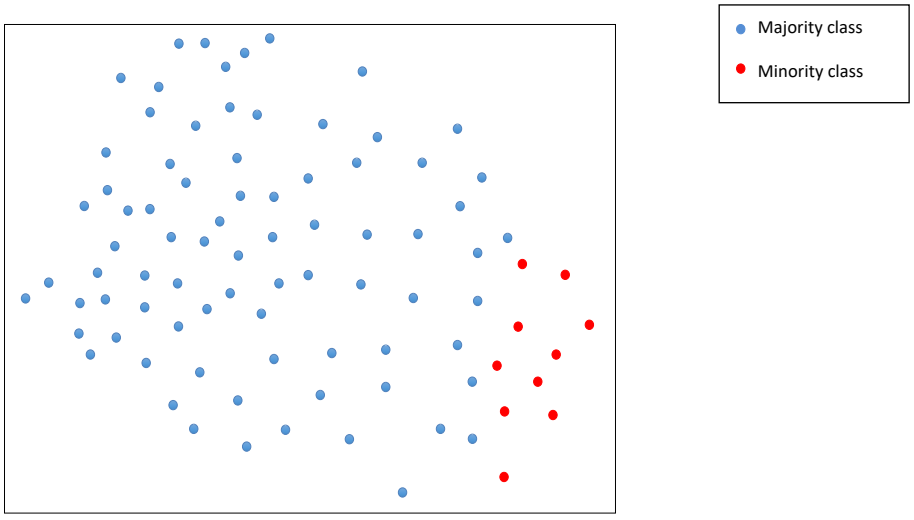
41

Approaches to the Imbalanced Learning Problem



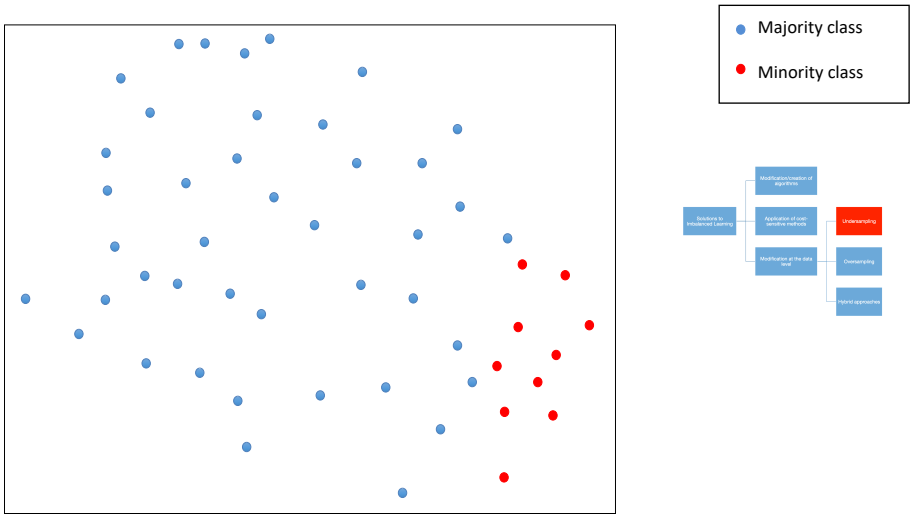
42

Approaches to the Imbalanced Learning Problem



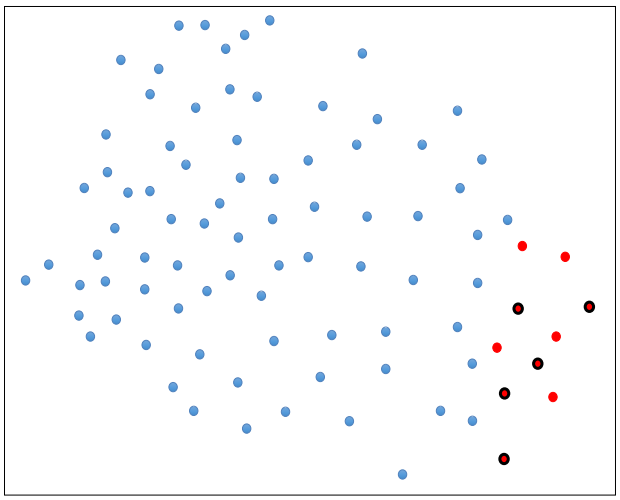
43

Approaches to the Imbalanced Learning Problem
Random Undersampling

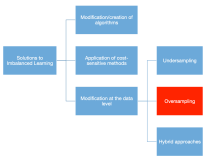


44

Approaches to the Imbalanced Learning Problem
Random Oversampling

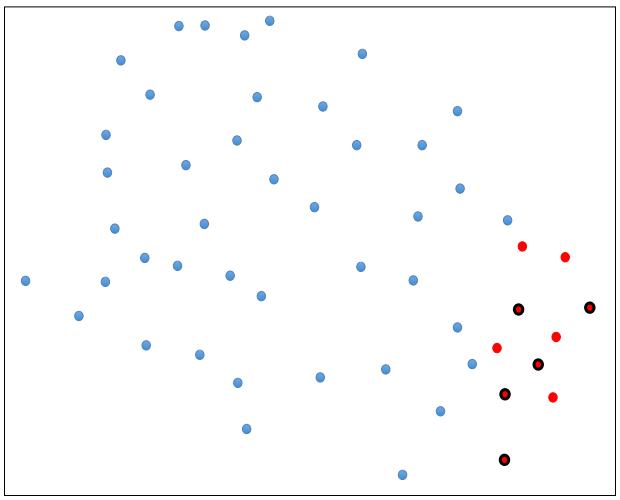


Majority class
Minority class
Generated sample

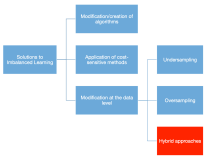


45

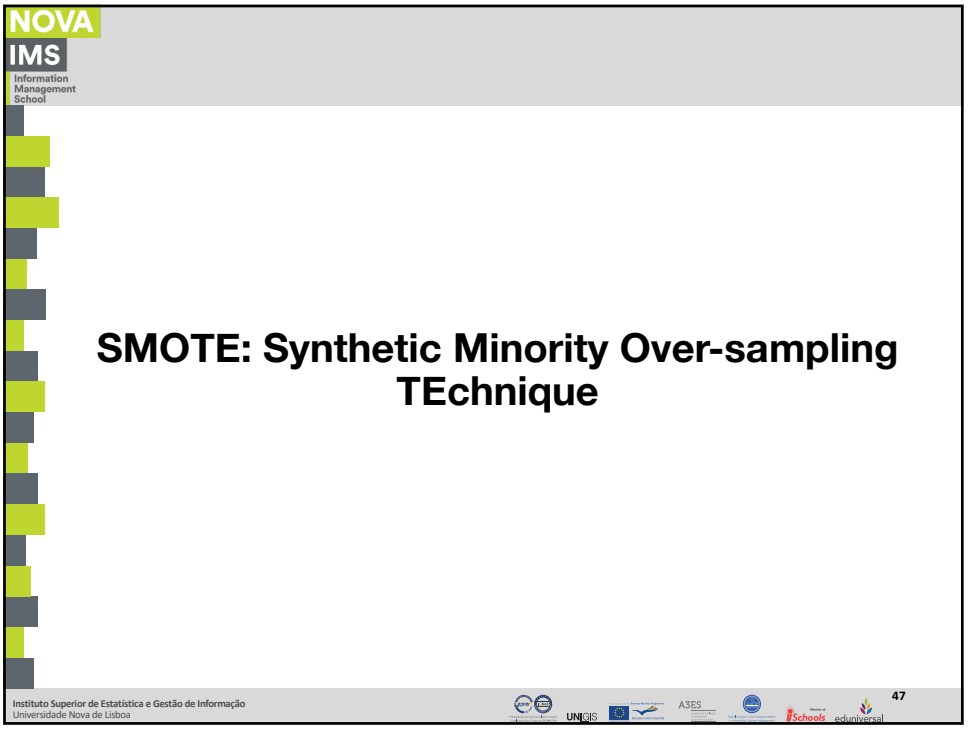
Approaches to the Imbalanced Learning Problem
Hybrid Approach



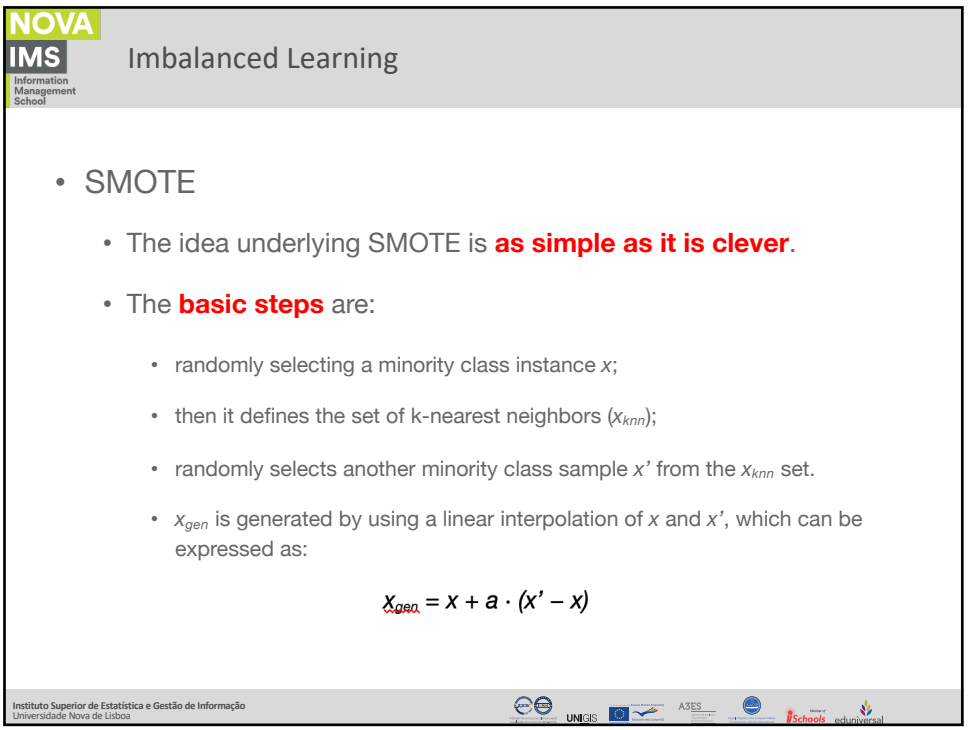
Majority class
Minority class
Generated sample



46

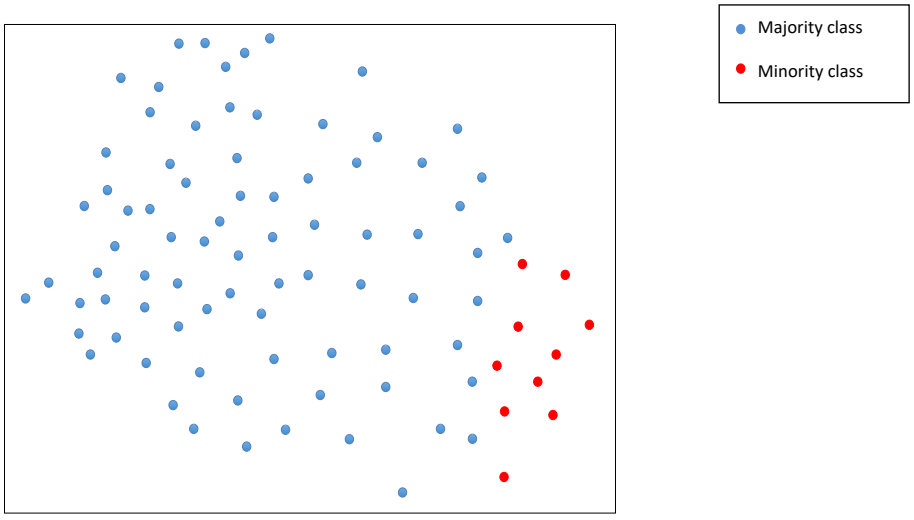

NOVA
IMS
Information Management School
SMOTE: Synthetic Minority Over-sampling TEchnique
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa
47

47


NOVA
IMS
Information Management School
Imbalanced Learning
• SMOTE
• The idea underlying SMOTE is as simple as it is clever.
• The basic steps are:
• randomly selecting a minority class instance x;
• then it defines the set of k-nearest neighbors (x_{knn});
• randomly selects another minority class sample x' from the x_{knn} set.
• x_{gen} is generated by using a linear interpolation of x and x', which can be expressed as:
$x_{gen} = x + a \cdot (x' - x)$
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

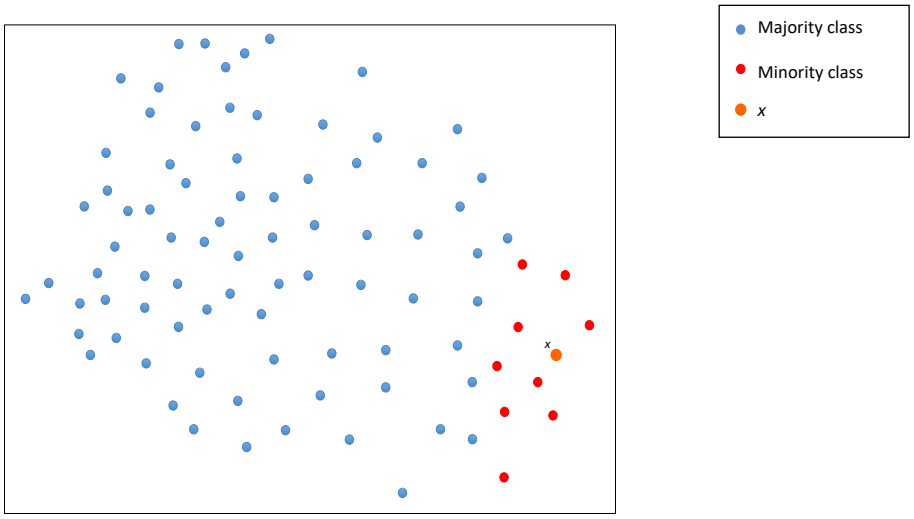
48

General aspects of data collection – acquiring knowledge



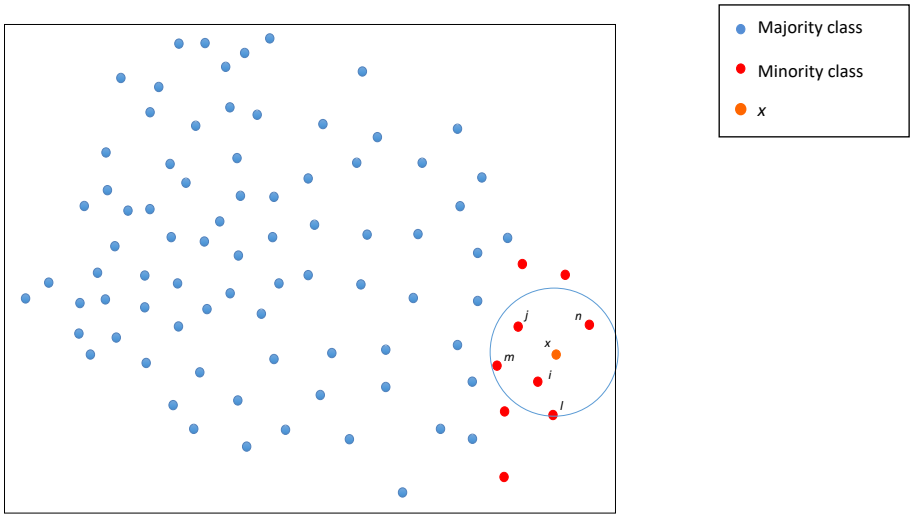
49

Informed Oversampling



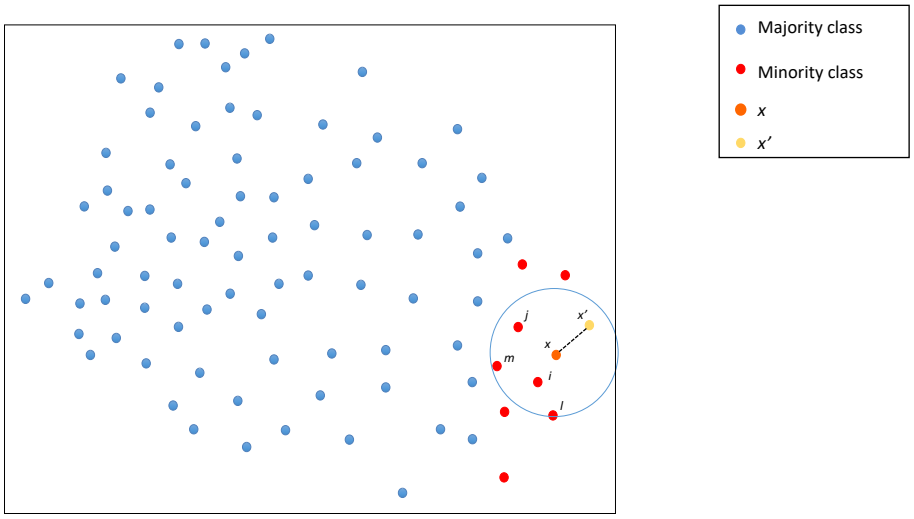
50

Informed Oversampling



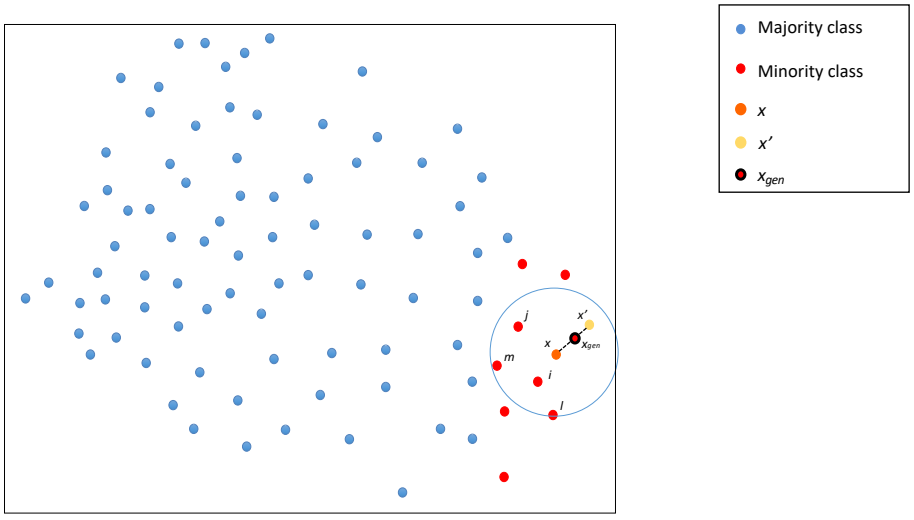
51

Informed Oversampling



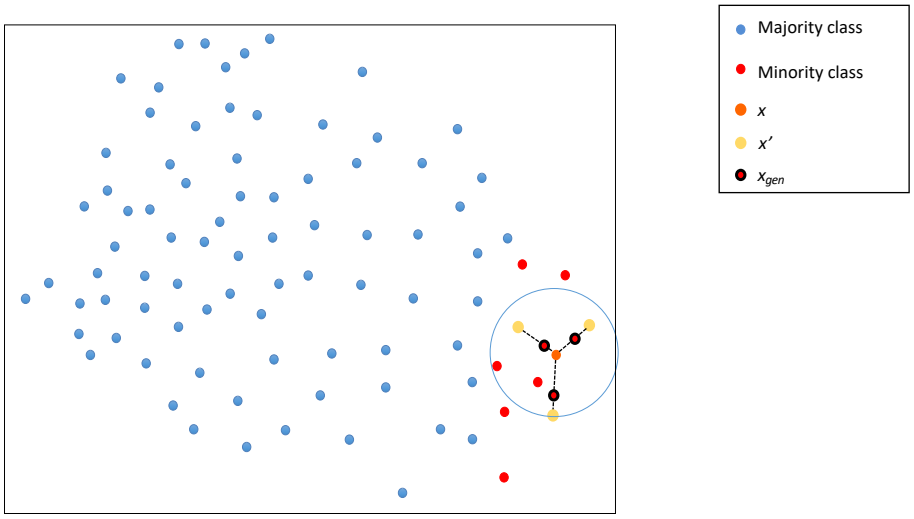
52

Informed Oversampling



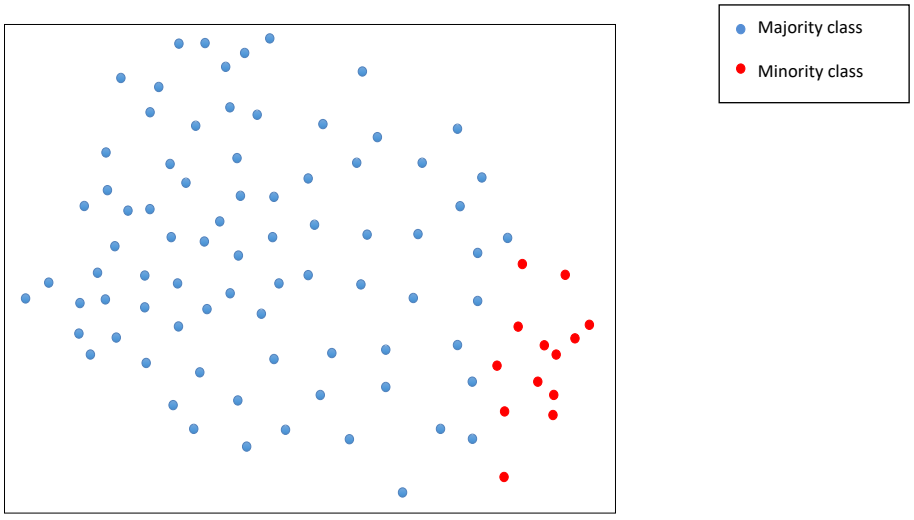
53

Informed Oversampling



54

Informed Oversampling

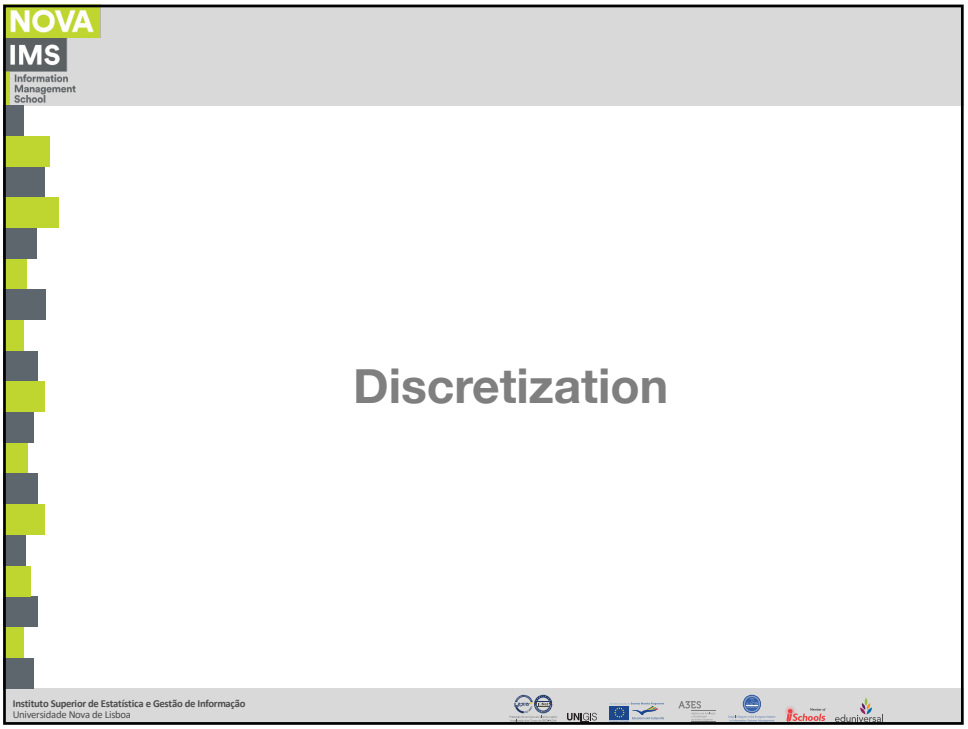


55

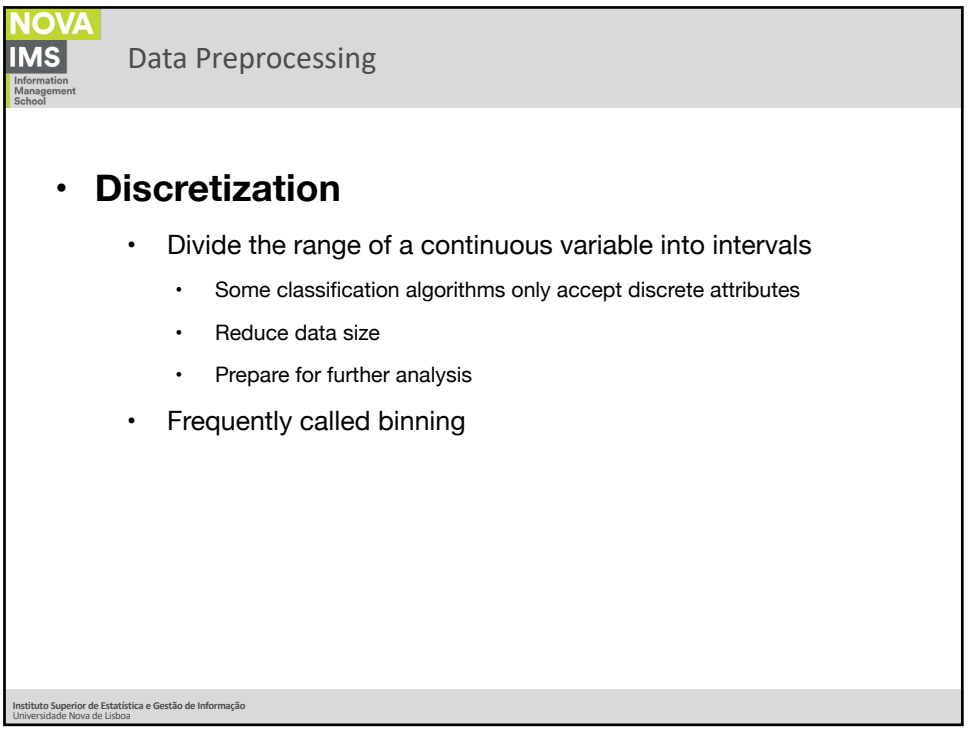
General aspects of data collection

- Use of artificial data:
  - It is always preferable to use real data;
    - Create data as realistic as possible;
    - Make artificial data as representative as possible.
  - The quality of the model is constrained by the quality of the data;
  - Creating artificial data translates into the introduction of some noise.

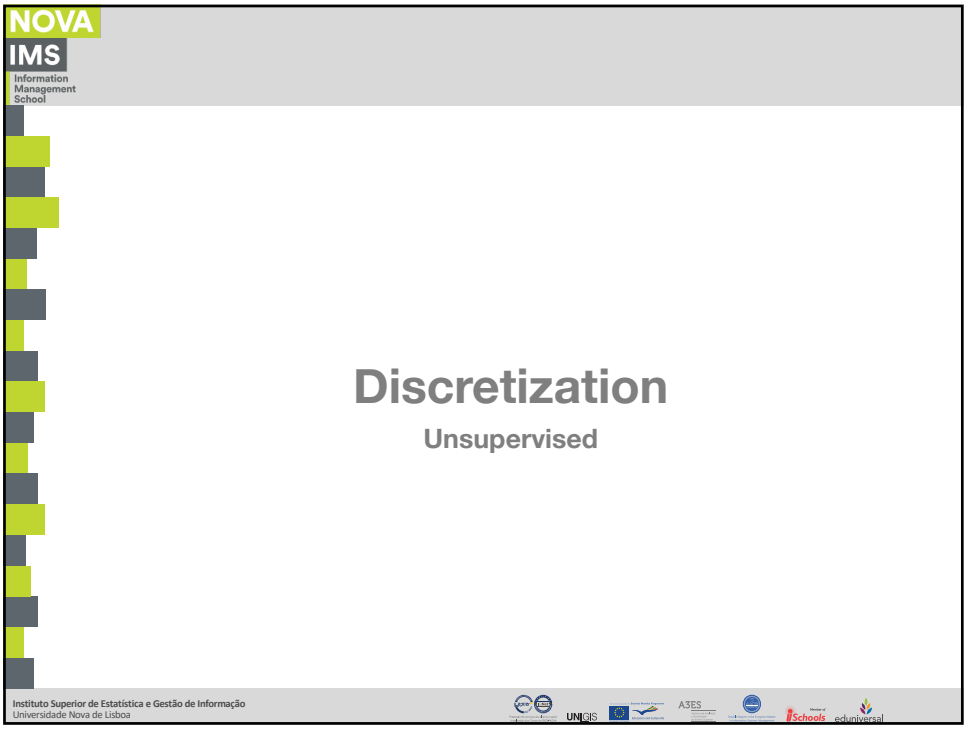
56


NOVA
IMS
Information Management School
Discretization
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

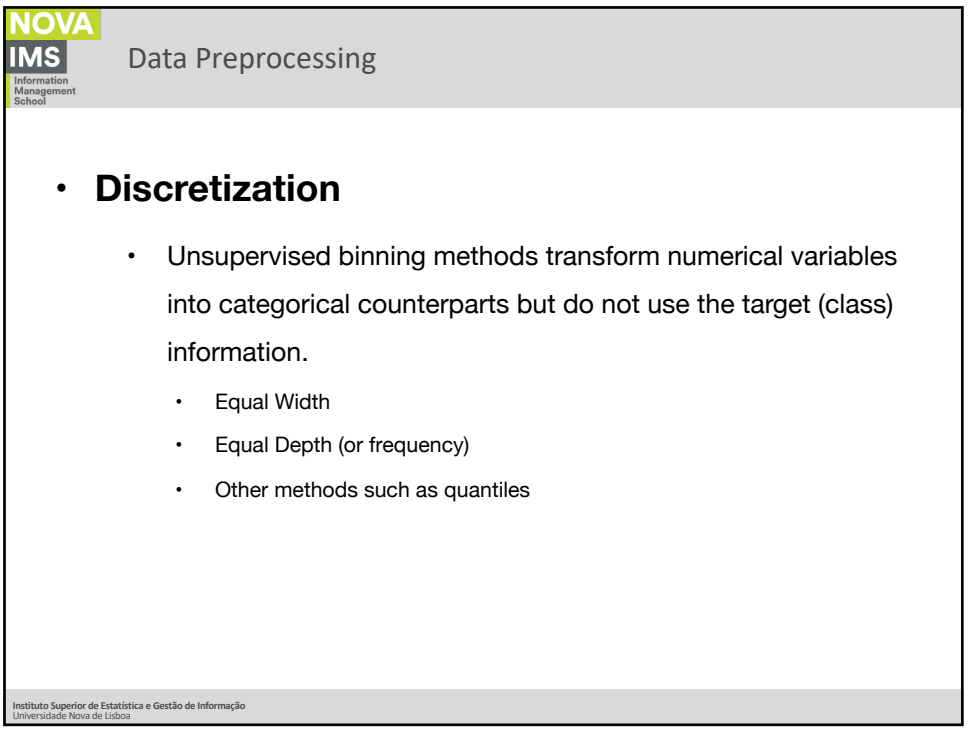
57


NOVA
IMS
Information Management School
Data Preprocessing
Discretization
Divide the range of a continuous variable into intervals
Some classification algorithms only accept discrete attributes
Reduce data size
Prepare for further analysis
Frequently called binning
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

58

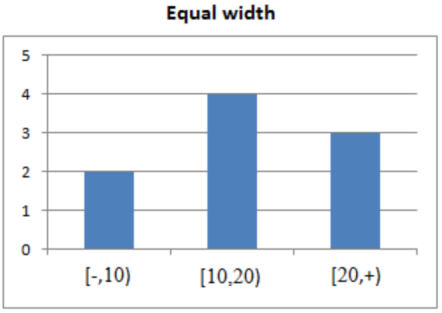

NOVA
IMS
Information Management School
Discretization
Unsupervised
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

59


NOVA
IMS
Information Management School
Data Preprocessing
Discretization
Unsupervised binning methods transform numerical variables into categorical counterparts but do not use the target (class) information.
Equal Width
Equal Depth (or frequency)
Other methods such as quantiles
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

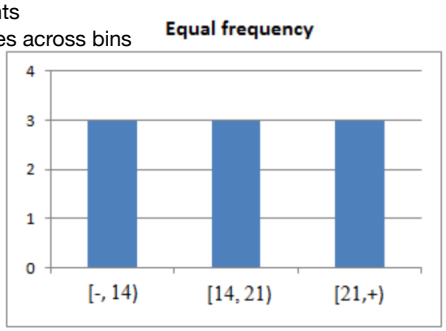
60

- **Discretization**
  - Equal-width binning
    - Divides the range into *n* intervals of equal size
    - If A and B are the minimum and the maximum values of the attribute, the width of the intervals will be: w=(B-A)/N
    - Most simple method
    - Outliers may dominate



61

- **Discretization**
  - Equal-depth binning
    - Divides the range into n intervals, each containing approximately the same number of samples
    - Generally preferred avoids clumps
    - Gives more intuitive breakpoints
    - Shouldn't break frequent values across bins



62

- **Discretization**
  - Class-independent methods (unsupervised)
    - Equal Width is simpler, good for many classes
      - can fail miserably for unequal distributions
    - Equal Height gives better results
  - Class-dependent methods can be better for classification
    - Decision tree methods build discretization on the fly
    - Naïve Bayes requires initial discretization
  - Many other methods exist …

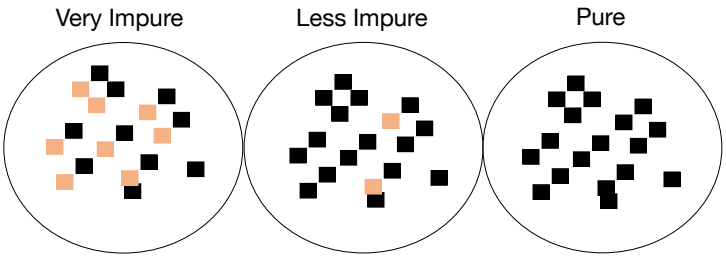
63

Discretization

Supervised

64

- **Discretization**
  - Entropy (also called Expected Information) based discretization

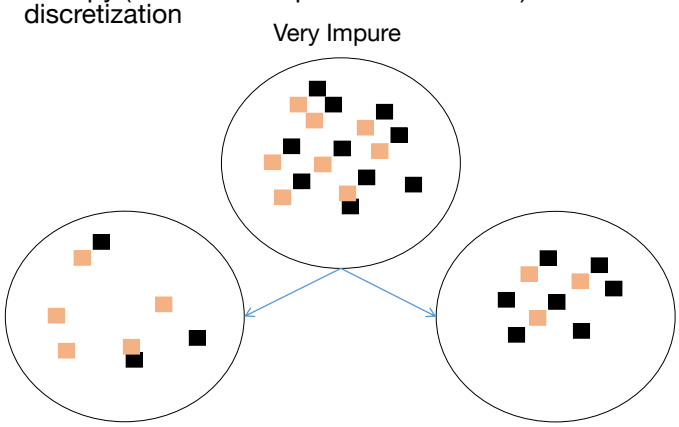


65

- **Discretization**
  - Entropy (also called Expected Information) based discretization
    - Sort examples in increased order
    - Each value forms an interval (m intervals)
    - Calculate the entropy measure of each discretization
    - Find the binary split boundary that minimizes the entropy function over all possible partitions. The split is selected as a binary discretization
    - Apply the process recursively until some stopping criteria is met

66

- **Discretization**
  - Entropy (also called Expected Information) based discretization

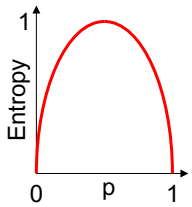


67

- **Discretization**
  - Entropy based discretization
    - Entropy
    - Idea: maximize info
    - It measures the purity of a partition:

$$E = - p \log2(p)$$

    - Where p is the probability of the examples belong to a specific class



68

- **Discretization**
  - Entropy based discretization

	Income <= 50K	Income > 50K
Age < 25	4	6

	Income <= 50K	Income > 50K
Age < 25	9	1

- **Discretization**
  - Entropy based discretization
    - **Partition entropy:**

$$Ent(S) = -\sum_{i=1}^{\#C} p_i \log_2(p_i)$$

    - **Gain in choosing A attribute:**

$$Gain(Ent_{new}) = Ent_{initial} - Ent_{new}$$

$$Gain(S,A) = Ent(S) - \sum_{v \in Valores(A)} \frac{\#S_v}{\#S} Ent(S_v)$$

- **Discretization**
  - Entropy based discretization

	Income <= 50K	Income > 50K
	13	7

$$Ent(S) = -(\frac{13}{20}\log_2(\frac{13}{20}) + \frac{7}{20}\log_2(\frac{7}{20})) = 0.403 + 0.530 = 0.934$$

- **Discretization**
  - Entropy based discretization

	Income <= 50K	Income > 50K
Age < 25	4	6

$$Ent(Age < 25) = -(\frac{4}{10}\log_2(\frac{4}{10}) + \frac{6}{10}\log_2(\frac{6}{10})) = 0.529 + 0.442 = 0.971$$

	Income <= 50K	Income > 50K
Age ≥ 25	9	1

$$Ent(Age \geq 25) = -(\frac{9}{10}\log_2(\frac{9}{10}) + \frac{1}{10}\log_2(\frac{1}{10})) = 0.137 + 0.332 = 0.469$$

- **Discretization**
  - Entropy based discretization

	Income <= 50K	Income > 50K
Age < 25	9	1
Age ≥ 25	4	6

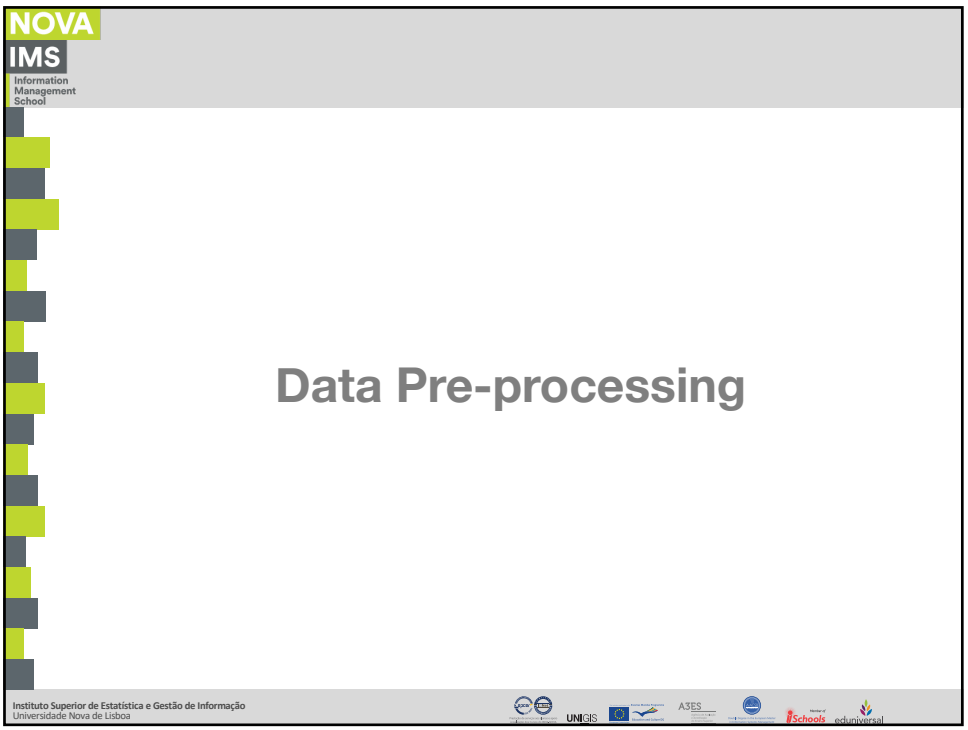
$$\sum_{v \in Valores(A)} \frac{\#S_v}{\#S} Ent(S_v) = \frac{1}{2}(0.469) + \frac{1}{2}(0,971) = 0.72$$

$$Gain(S,A) = Ent(S) - \sum_{v \in Valores(A)} \frac{\#S_v}{\#S} Ent(S_v)$$

$$Ent(S) = -(\frac{13}{20}\log_2(\frac{13}{20}) + \frac{7}{20}\log_2(\frac{7}{20})) = 0.403 + 0.530 = 0.934$$


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

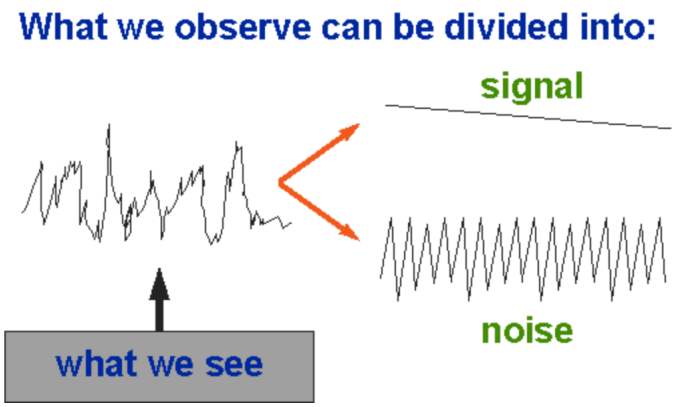
2

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;

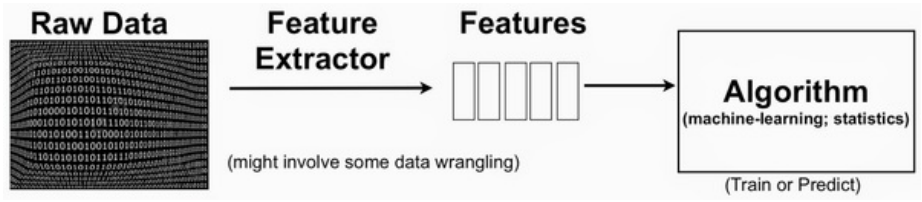
3

Data Preparation



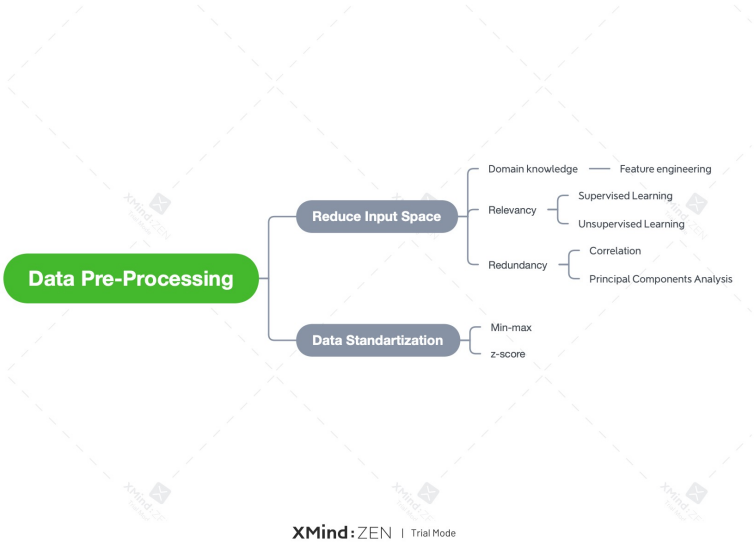
4

Data Preparation



5

Data Preparation



6

Reducing Input Space

7

Data Preprocessing

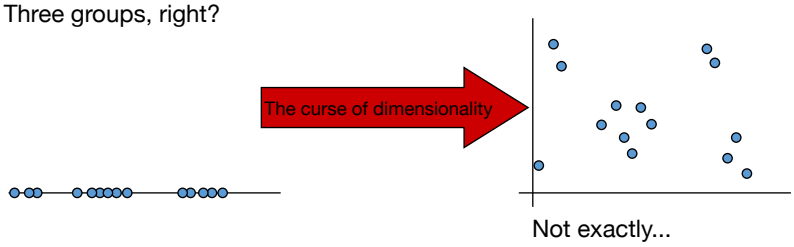
- **Additional considerations about data:**
  - Curse of dimensionality – the input space grows exponentially with the number of input variables;
  - The larger the input space, the more data and computing power we need.

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

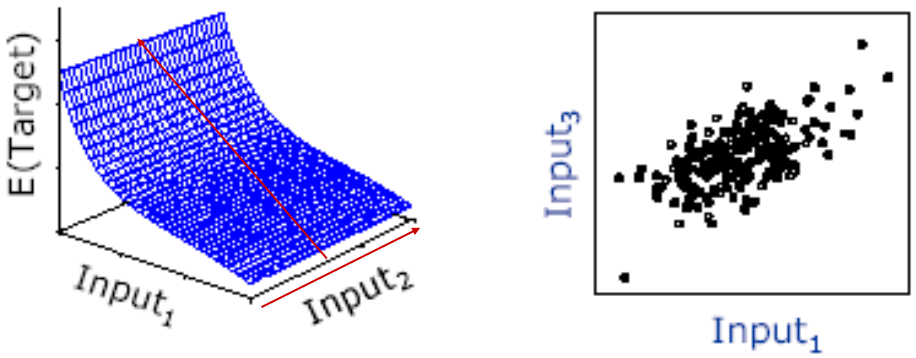
NOVA IMS Information Management School

Data Preprocessing

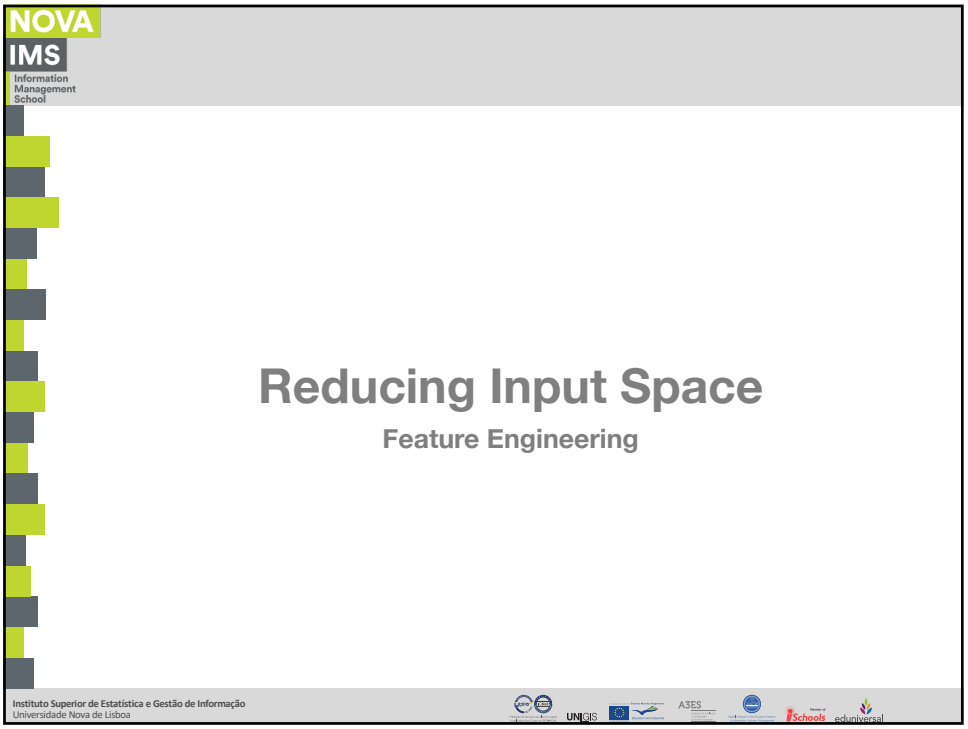
- **Size Reduction of the Input Space** (or feature selection):

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))

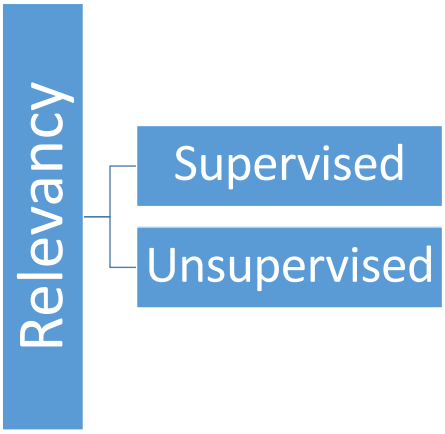
13

Reducing Input Space

Relevancy

14

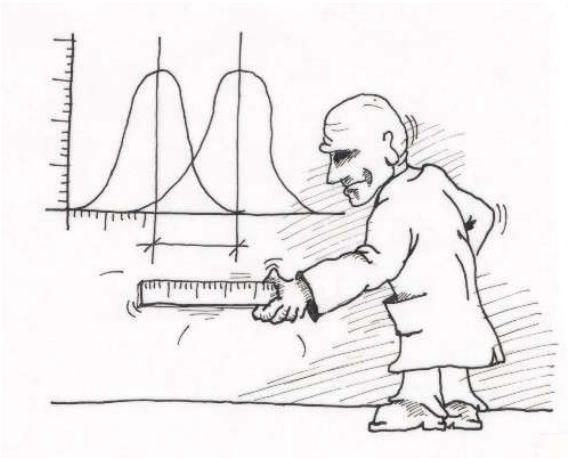
Data Preparation



15

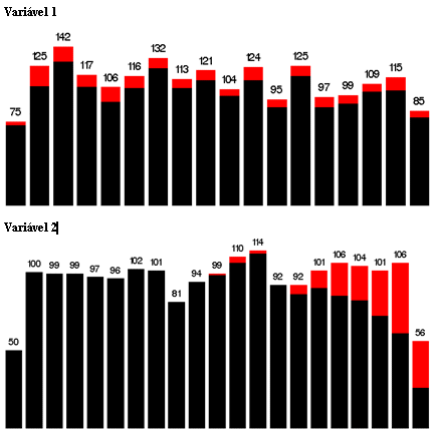
Data Preprocessing

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



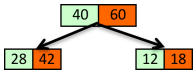
16

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

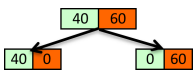


17

- **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

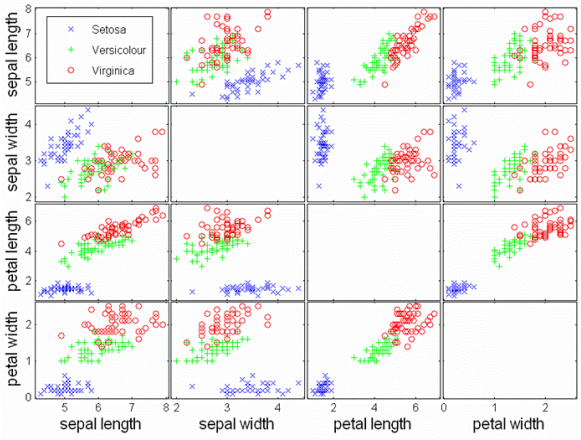
Reducing Input Space

Redundancy

19

Data Preprocessing

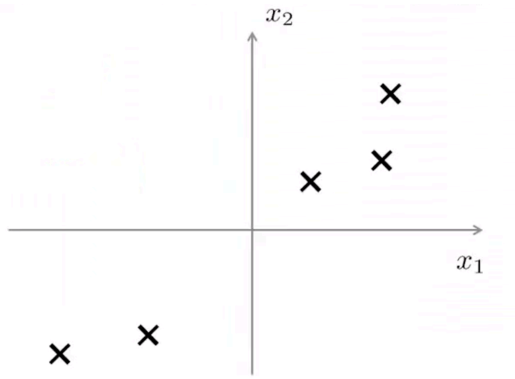
- **Size Reduction of the Input Space:**



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.

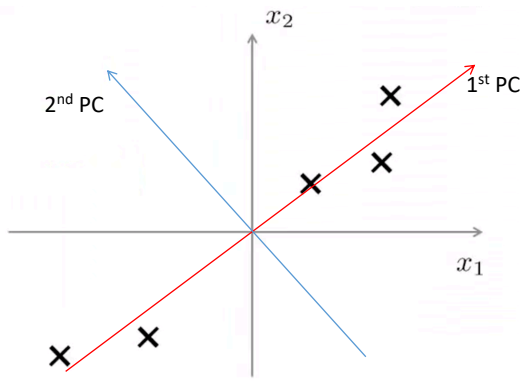
- **Size Reduction of the Input Space:**
  - Principal Component Analysis



NOVA IMS Information Management School

Data Preprocessing

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

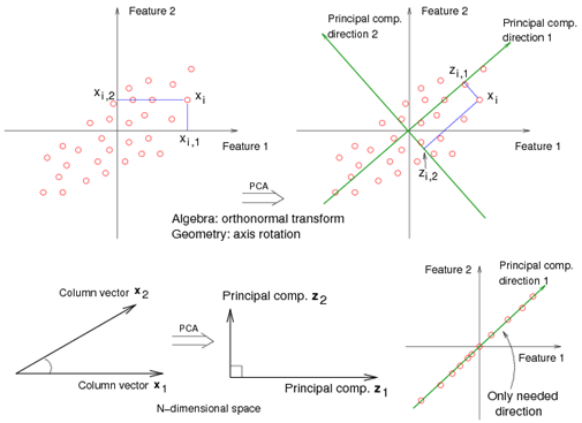


23

NOVA IMS Information Management School

Data Preprocessing

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

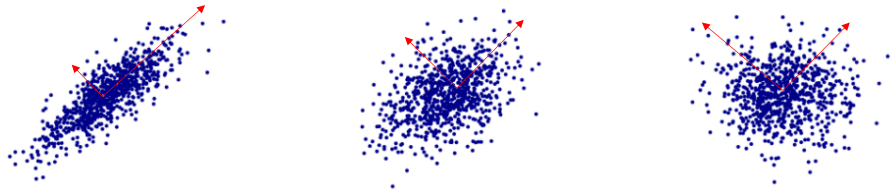


24

NOVA IMS Information Management School

Data Preprocessing

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

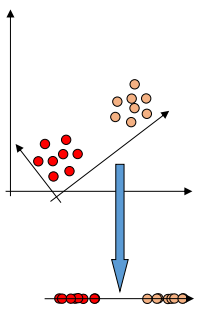


25

NOVA IMS Information Management School

Data Preprocessing

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

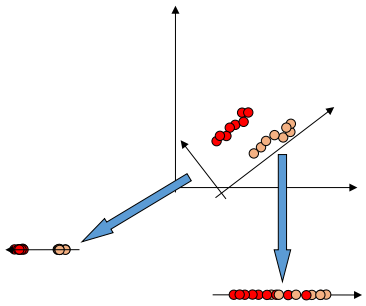


26

NOVA IMS Information Management School

Data Preprocessing

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

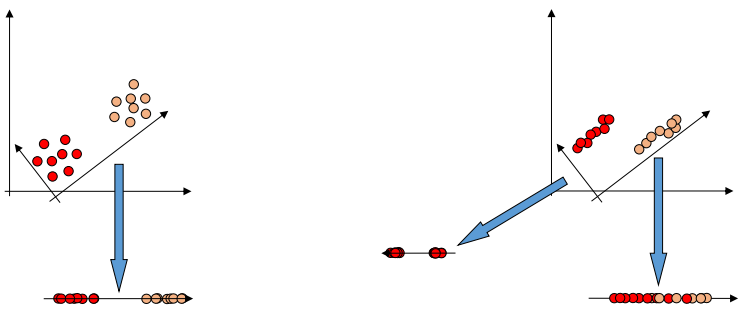


27

NOVA IMS Information Management School

Data Preprocessing

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



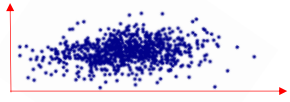
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

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

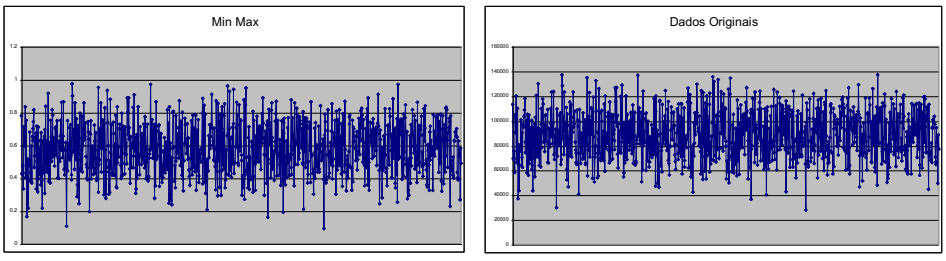
- **Normalization:**
  - Min-Max $$y' = \left(\frac{y - \min 1}{\max 1 - \min 1} \right) \underbrace{(\max 2 - \min 2) + \min 2}_{\text{optional}}$$
  - Zscore $$y' = \frac{y - \mu}{std}$$

31

NOVA IMS Information Management School

Data Preprocessing

- **Normalization:**



32

33

Data Mining

NOVA-IMS

02/11/2021

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

Agenda

- Cluster analysis
- Variables to use
- Similarity criterion
- Clustering algorithms
  - A Priori Grouping
  - RFM Analysis

2

Cluster Analysis

3

Clustering

- **Cluster Analysis**
  - Cluster Analysis is a **basic conceptual activity of human beings**;
  - A **fundamental process**, common to many sciences, essential to the development of scientific theories;
  - The possibility of **reducing the infinite complexity of real** to sets of objects or similar phenomena, is one of the most powerful tools in the service of mankind.

4

NOVA IMS Information Management School

Clustering

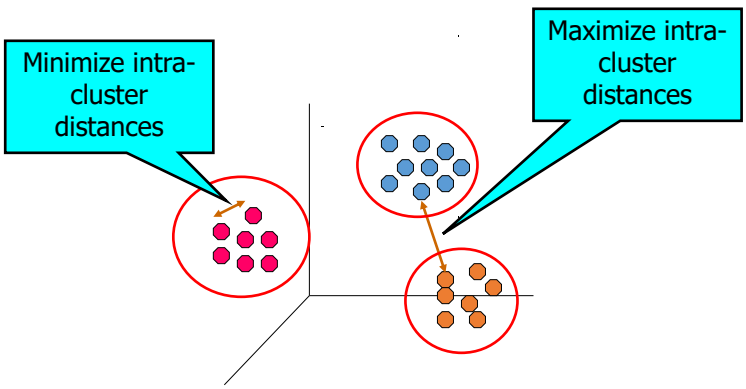
- **Cluster Analysis**
  - Cluster analysis is a generic name for a variety of methods that are used to **group entities**;
  - Objective: **To form groups of objects that are similar to each other**;
  - From a data collection about a group of entities, seeks to organize them **in homogeneous groups**, assessing a "frame" of similarities/differences between units.

5

NOVA IMS Information Management School

Clustering

- **Cluster Analysis**



Source: Tan, Steinbach, Kumar, Introduction to Data Mining 2004

6

NOVA IMS Information Management School

Clustering

- **Cluster Analysis**
  - Classification:
    - Starts out with a **pre-classified training set**, that is, the method has a set of data which contains not only the variables to use in classification but also the class to which each of the records belongs;
    - Attempts to develop a model capable of predicting how a new record will be classified.

7

NOVA IMS Information Management School

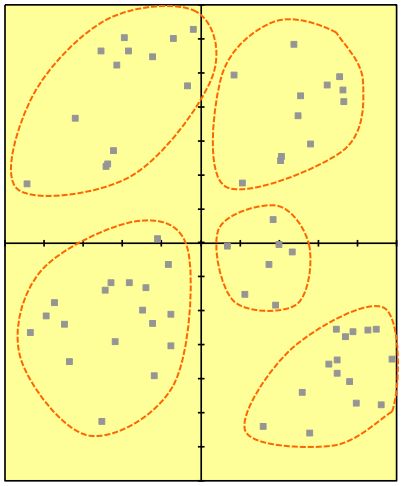
Clustering

- **Cluster Analysis**
  - Clustering:
    - There is **no pre-classified data**;
    - We search for groups of records (clusters) that are similar to one another;
    - Underlying is the expectation that similar customers in terms of the variables used will behave in similar ways.

8

NOVA IMS Information Management School

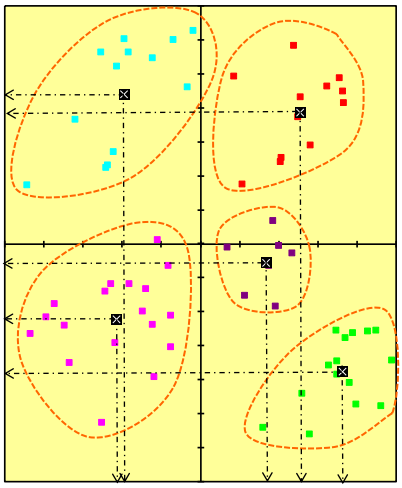
Clustering



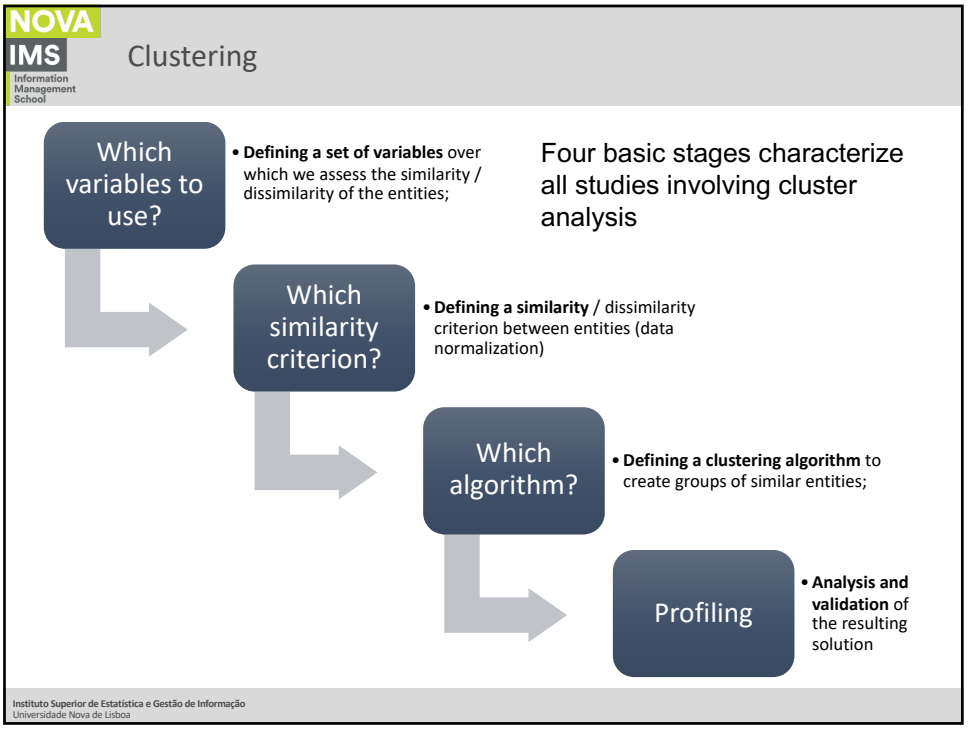
9

NOVA IMS Information Management School

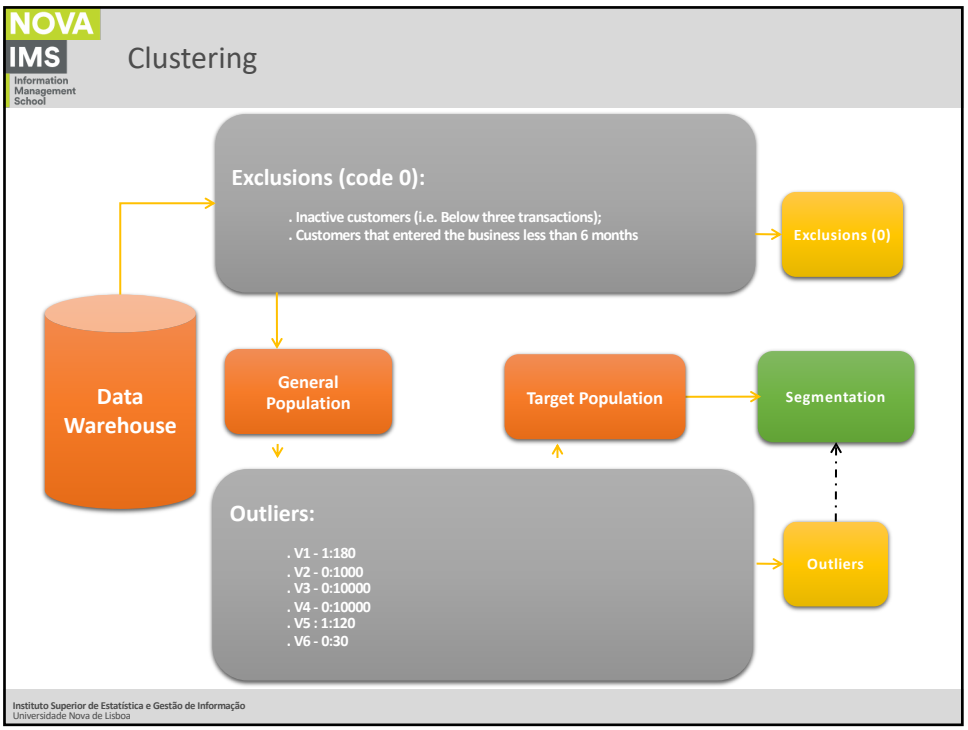
Clustering



10


NOVA IMS Information Management School
Clustering
Which variables to use?
• Defining a set of variables over which we assess the similarity / dissimilarity of the entities;
Four basic stages characterize all studies involving cluster analysis
Which similarity criterion?
• Defining a similarity / dissimilarity criterion between entities (data normalization)
Which algorithm?
• Defining a clustering algorithm to create groups of similar entities;
Profiling
• Analysis and validation of the resulting solution
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

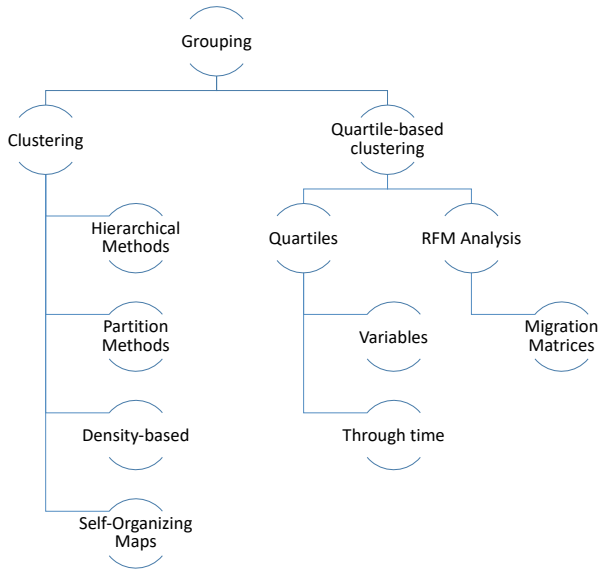
11


NOVA IMS Information Management School
Clustering
Exclusions (code 0):
. Inactive customers (i.e. Below three transactions);
. Customers that entered the business less than 6 months
Exclusions (0)
Data Warehouse
General Population
Target Population
Segmentation
Outliers:
. V1 - 1:180
. V2 - 0:1000
. V3 - 0:10000
. V4 - 0:10000
. V5 : 1:120
. V6 - 0:30
Outliers
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

12

NOVA IMS Information Management School

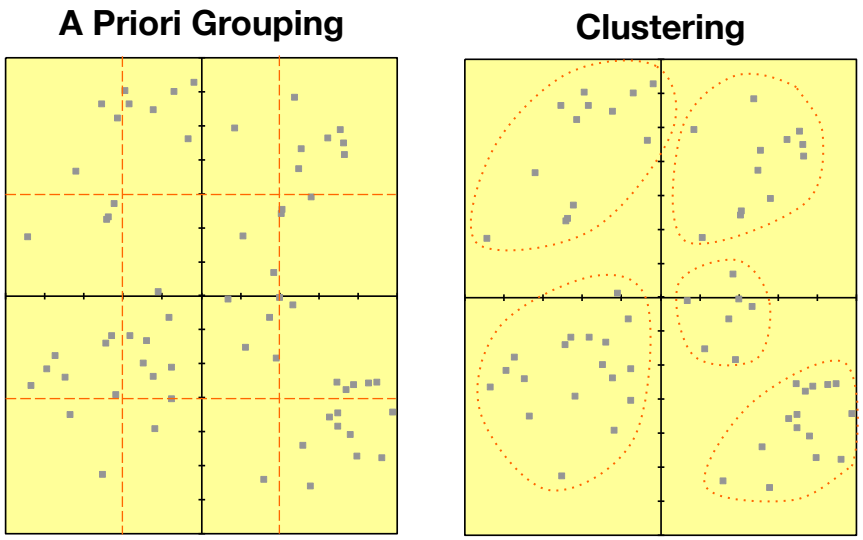
Clustering



13

NOVA IMS Information Management School

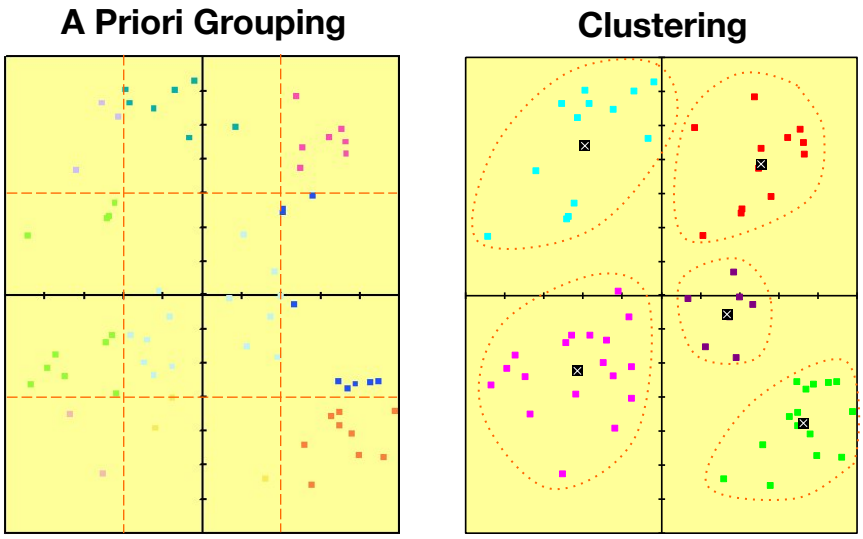
Clustering



14

NOVA IMS Information Management School

Clustering



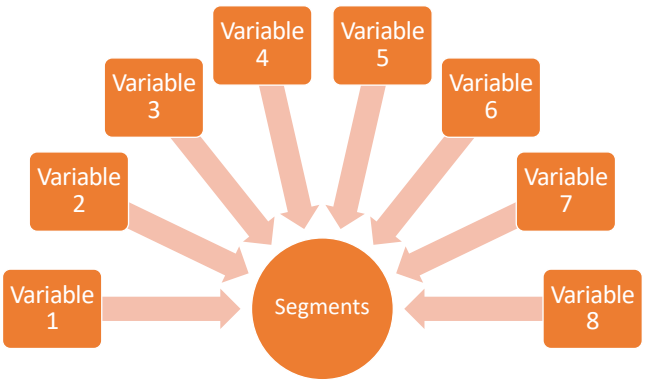
15

NOVA IMS Information Management School

Decide on the Variables to Use

16

Clustering



17

Clustering

- **Deciding which variables to use:**
  - **Objective of the segmentation**
    - Value/Engagement;
    - Needs;
    - Behaviors/Consumption;
    - GeoDemographics/Socio-economic characteristics.

18

NOVA IMS Information Management School

Clustering

- **Deciding which variables to use:**
  - The type of problem determines the variables to choose;
  - If the purpose is to group objects, the choice of variables with discrimination ability is crucial;
  - The quality of any cluster analysis is, first of all, conditioned by the variables used.

NOVA IMS Information Management School

Clustering

- **Deciding which variables to use:**
  - The choice of variables should replicate a theoretical context, a reasoning;
  - This process is carried out based on a set of variables that we know to be good discriminators for the problem at hand;
  - First of all, the quality of the cluster analysis reflects the discrimination ability of the variables we decided to use in our study.

Similarity criterion

21

Clustering

- **Similarity criterion:**
  - The analysis of similarity relations has been dominated by metrics based on Euclidean Spaces;
  - Objects as points in a multidimensional space, in a way that the observed dissimilarities between the objects correspond to distances between the respective points;
  - Thus, the use of clustering methods most times means the use of similarity ratios that respect these metrics:

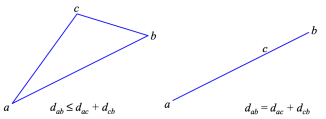
22

- **Similarity criterion:**

- In mathematics, a true measure of distance, called a *metric*, obeys three properties. These metric axioms are as follows, where d_{ab} denotes the distance between objects *a* and *b*:

  1. $d_{ab} = d_{ba}$ *(measure is symmetric)*
  2. $d_{ab} \geq 0$ and = 0 if and only if *a* = *b (distances are always positive except when the objects are identical)*
  3. $d_{ab} \leq d_{ac} + d_{ca}$ *(triangle inequality)*

Exhibit 5.1 Illustration of the triangle inequality for distances in Euclidean space.



- **Similarity criterion:**

- Euclidian distance: the distance between two elements (*i,j*) is the square root of the sum of the squares of the differences between *i* and *j* values for all variables (*v=1, 2,...., p*):

$$d_{ij} = \sqrt{\sum_{v=1}^{p} \left(X_{iv} - X_{jv}\right)^2} \; euclidean \, also \, known \, as \; L_2$$

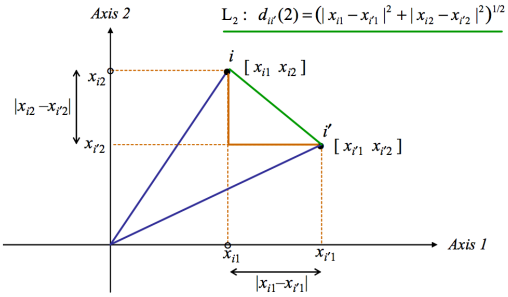
$$d_{ij} = \sum_{v=1}^{p} \left|X_{iv} - X_{jv}\right| City \, Block \, or \, L_1$$

- **Similarity criterion:**

- Euclidian distance: the distance between two elements (*i,j*) is the square root of the sum of the squares of the differences between *i* and *j* values for all variables (*v=1, 2,...., p*):

$$L_1: \; d_{ii'}(1) = |x_{i1} - x_{i'1}| + |x_{i2} - x_{i'2}|$$

$$L_2: \; d_{ii'}(2) = \left(|x_{i1} - x_{i'1}|^2 + |x_{i2} - x_{i'2}|^2\right)^{1/2}$$



- **Similarity criterion:**

- Minkowski distance: is defined from the absolute distance, and can be considered as a generalization of both the Euclidean distance and the Manhattan distance. It coincides with Euclidean distance when *r=2* and with Manhattan distance when *r=1*:

$$d_{ij} = \left(\sum_{v=1}^{p} \left| X_{iv} - X_{jv} \right|^r \right)^{1/r}$$

NOVA IMS Information Management School

Clustering

- **Similarity criterion:**
- If a weight is assigned to each variable, according to their importance for the analysis, the weighted Euclidean distance takes the following form:

$$d_{ij} = \sqrt{\sum_{v=1}^{p} w_v \left(X_{iv} - X_{jv}\right)^2}$$

NOVA IMS Information Management School

Clustering

- **Similarity criterion:**
- Pearson correlation coefficient: its function is to measure the degree of linear correlation between two elements, for a number of variables:

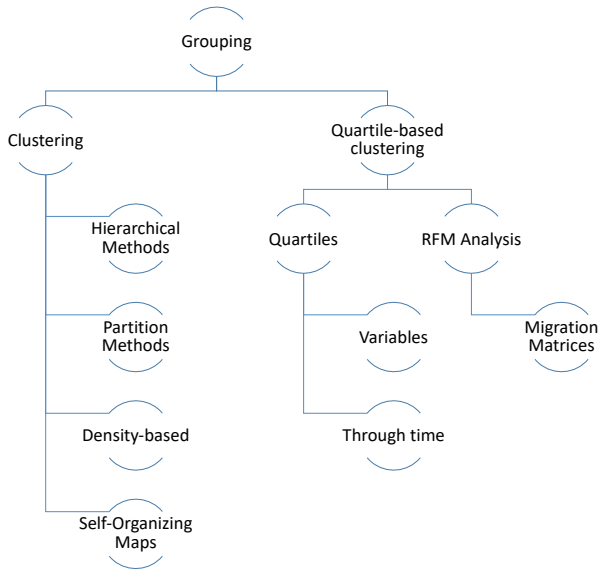
Correlation

$$\rho(x_i, x_{i'}) = \frac{\sum_j (x_{ij} - \bar{x}_i)(x_{i'j} - \bar{x}_{i'})}{\sqrt{\sum_j (x_{ij} - \bar{x}_i)^2 \sum_j (x_{i'j} - \bar{x}_{i'})^2}}$$

Choose the Algorithm

29

Clustering

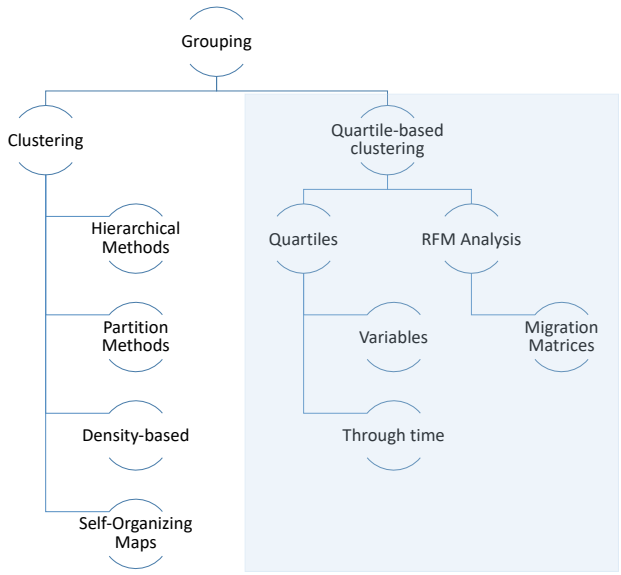


30

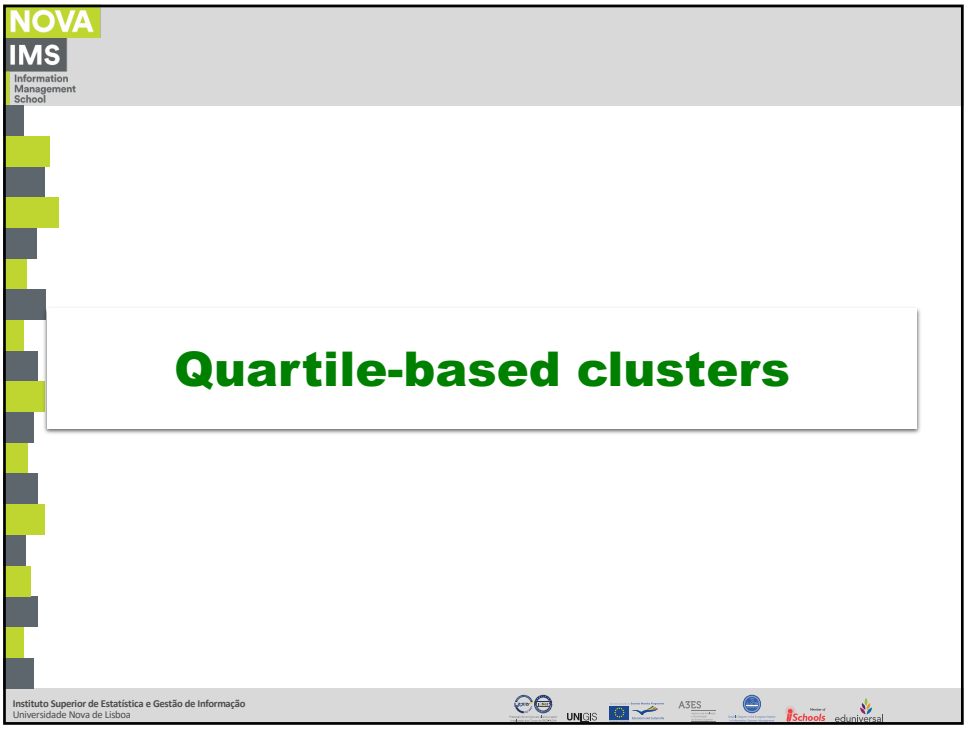
A Priori Grouping

31

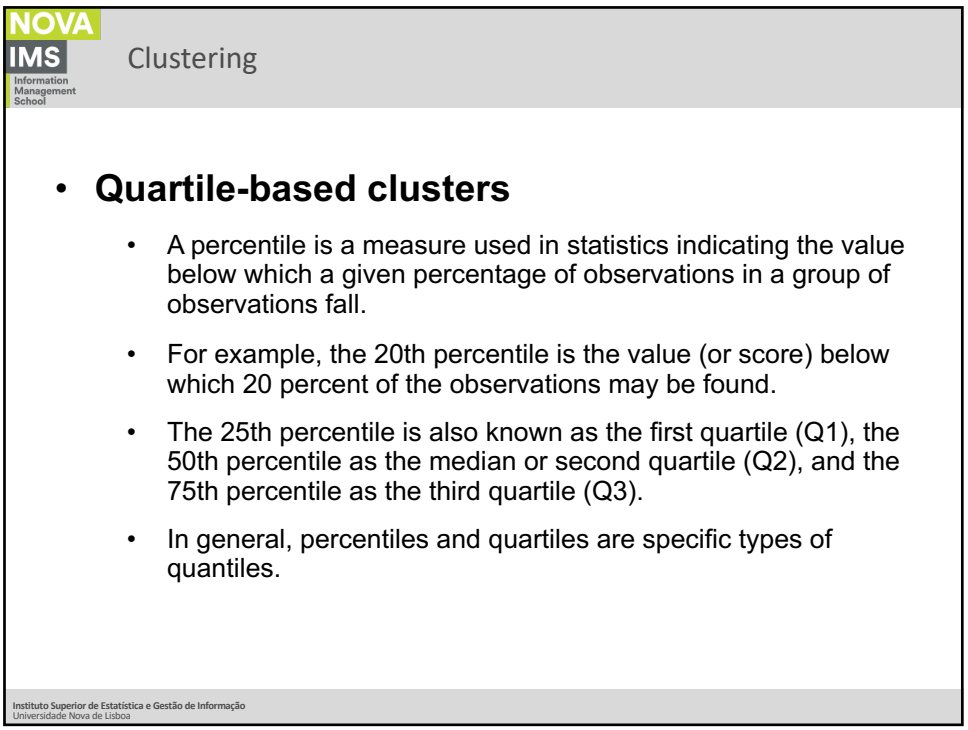
Clustering



32


NOVA
IMS
Information Management School
Quartile-based clusters
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

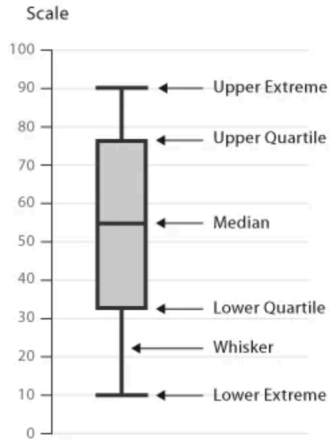
33


NOVA
IMS
Information Management School
Clustering
Quartile-based clusters
A percentile is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations fall.
For example, the 20th percentile is the value (or score) below which 20 percent of the observations may be found.
The 25th percentile is also known as the first quartile (Q1), the 50th percentile as the median or second quartile (Q2), and the 75th percentile as the third quartile (Q3).
In general, percentiles and quartiles are specific types of quantiles.
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

34

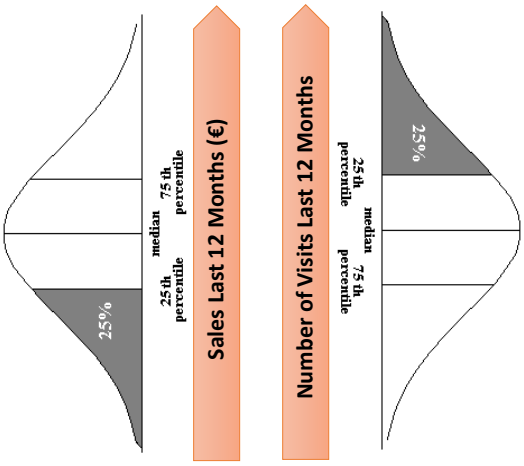
Clustering

- **Quartile-based clusters**



35

Clustering



36

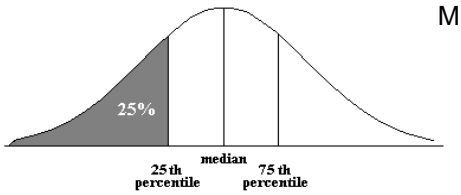
Clustering

Sales last 12M VL X Visits Last 12M NR

Tabela de Contingência		TICKETS_ULT_12M_NR				Totais
		<Q1	[Q1,Q2[	[Q2,Q3[	>=Q3	
VENDAS_ULT_12M_VL	<Q1			Buy very small amounts very frequently		
	[Q1,Q2[					
	[Q2,Q3[					
	>=Q3	A lot of money few times				
Totais						

37

Clustering



Migration Matrix

12/2012 – 12/2013

06/2013 – 06/2014

Evolution Sales Last 12 Months (12.2012 – 06.2014)

38

Clustering

Variable – Sales Last 12M VL

		Period 201412					Totais
		.	<Q1	[Q1,Q2[	[Q2,Q3[	>=Q3	
Period 201406	.				Positive Change		
	<Q1						
	[Q1,Q2[						
	[Q2,Q3[	Negative Change					
	>=Q3						
Totais							

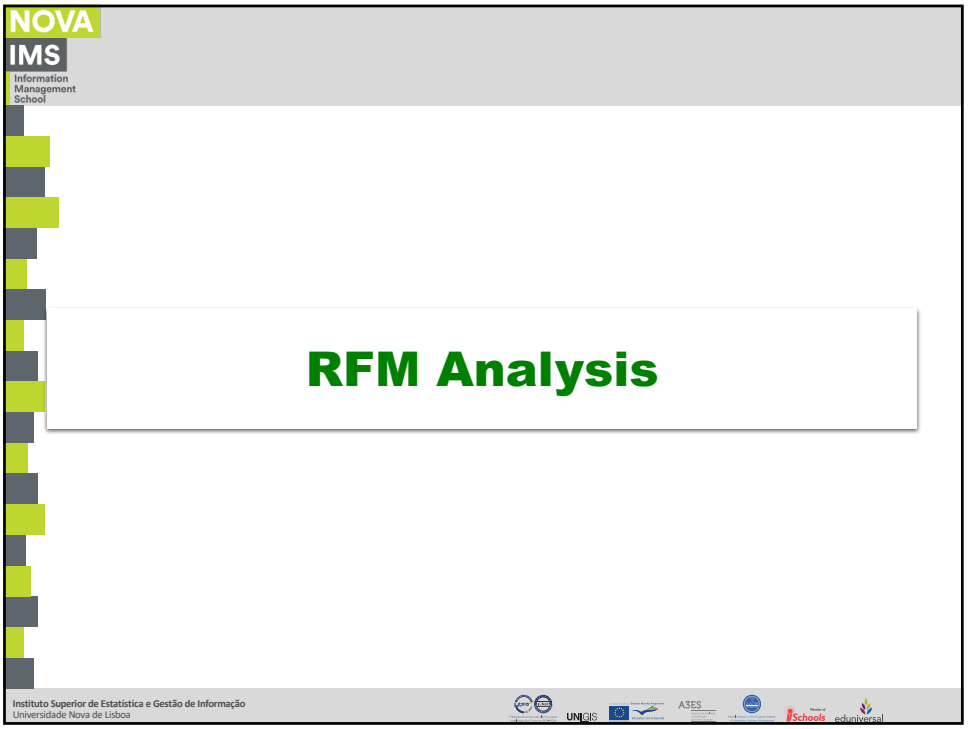
Migration Matrix

39

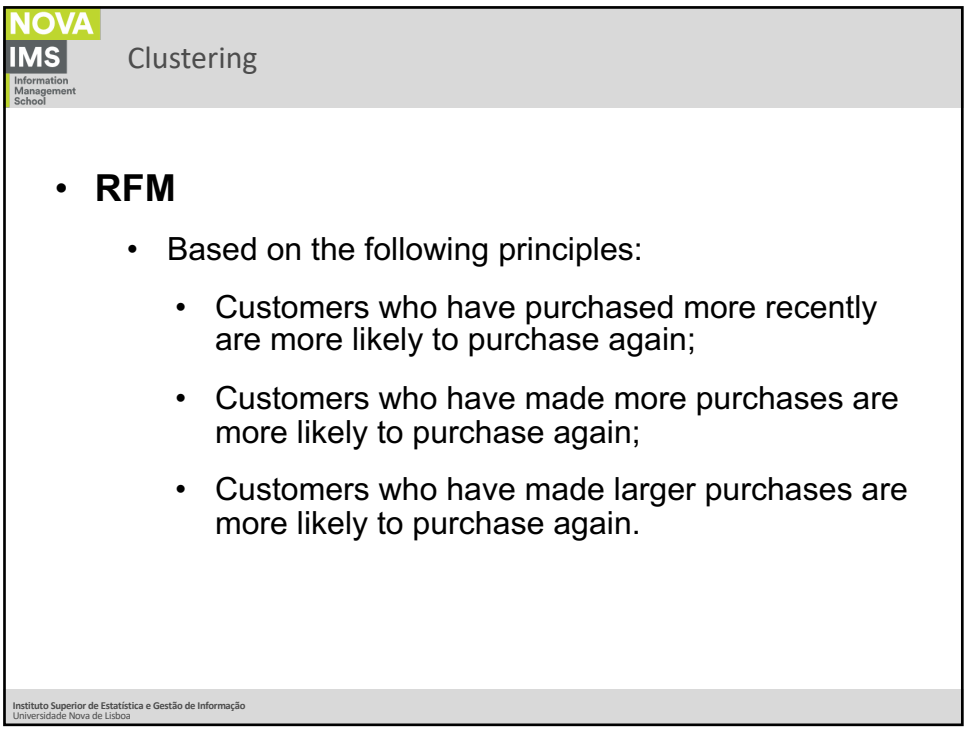
Clustering

Tabela de Contingência		Period 06.2013/06.2014					Totais
		.	<Q1	[Q1,Q2[	[Q2,Q3[	>=Q3	
Periodo 12.2012/ 12.2013	.	0 a	30603 b	17050 c	8815 d	6427 e	62895 13.5%
	<Q1	28734 f	55411 g	13600 h	2772 i	178 j	100695 21.6%
	[Q1,Q2[	14834 k	15506 l	52421 m	16838 n	1097 o	100696 21.6%
	[Q2,Q3[	6540 p	3450 q	19346 r	59756 s	11608 t	100700 21.6%
	>=Q3	3765 u	760 v	1835 w	14733 x	79604 y	100697 21.6%
Totais		53873 11.6%	105730 22.7%	104252 22.4%	102914 22.1%	98914 21.2%	465683 100%

40


NOVA
IMS
Information Management School
RFM Analysis
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

41


NOVA
IMS
Information Management School
Clustering
RFM
Based on the following principles:
Customers who have purchased more recently are more likely to purchase again;
Customers who have made more purchases are more likely to purchase again;
Customers who have made larger purchases are more likely to purchase again.
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

42

NOVA IMS Information Management School

Clustering

- **RFM**
  - Has been in active use in Direct Marketing for more than 40 years;
  - It can be used only for customer files that contain purchase history;
  - There are two methods:
    - Exact Quintiles;
    - Hard coding;

43

NOVA IMS Information Management School

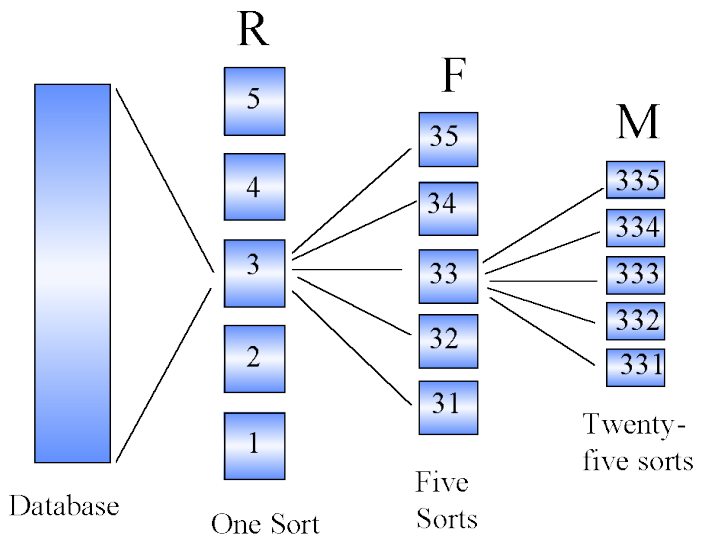
Clustering

- **RFM**
  - How to do it (Exact Quintiles)?
    - We sort the database according to recency and divide into 5 quintiles (5 equal segments);
    - Do the same for the variables frequency and monetary;
    - Result: 125 cells of equal size (5*5*5).

44

NOVA IMS Information Management School

Clustering

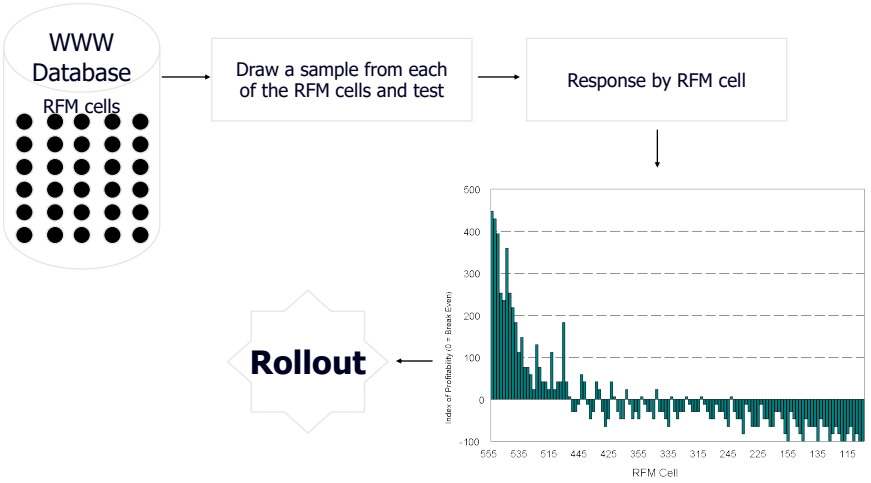


45

NOVA IMS Information Management School

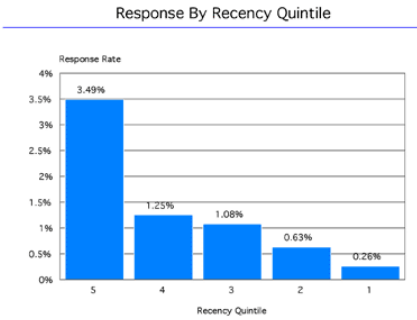
Clustering

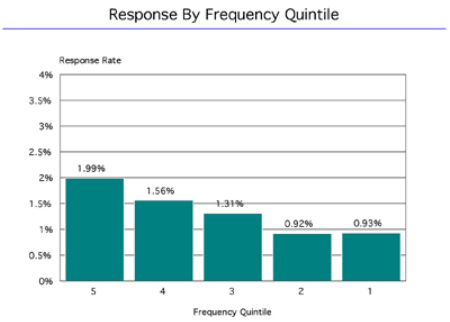
Typical use



46

Clustering





47

Clustering

Migration Matrix

Segment in YY/YY/YYYY	Segment 1 XX/XX/XXXX							
	44	**45**	**51**	**52**	**53**	**54**	**55**	**Total YYYY**
44	.	41914	209	4362	1862	253	22	48622
45	34200	**58714**	7875	14961	8968	1652	128	126498
51	505	9089	**7823**	4895	5420	30	.	27762
52	9109	7044	7151	**83963**	11103	8820	208	127398
53	3572	5758	4211	5578	**29736**	3691	9	52555
54	382	124	93	6507	2190	**36300**	4128	49724
55	69	10	22	156	62	4089	**14446**	18854
Total XXXX	47837	122653	27384	120422	59341	54835	18941	

48

NOVA IMS Information Management School

Clustering

- **RFM**
  - Hard coding
    - Categories are divided by exact values (0-3 months; 4-6 months; 7-9 months; etc.);
    - More expensive in terms of programming, categories tend to change over time;
    - Very different quantities from cell to cell.

NOVA IMS Information Management School

Clustering

- **RFM**
  - Its popularity comes from its simplicity, low cost and capacity to classify customers based on their behavior;
  - Opportunity to carry out tests in small, representative groups of each cell;
  - A more sophisticated modeling is almost always better, but is it worth it? Not always.

51

Data Mining

NOVA-IMS

10/11/2021

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

AGENDA

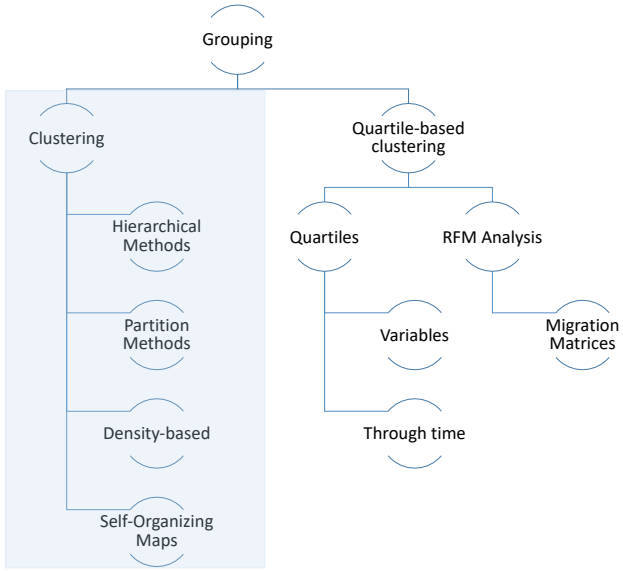
- Cluster analysis
  - Clustering techniques
    - Hierarchical Methods (agglomerative)
    - Partitioning Methods (kmeans and k-meadoids)

2

Clustering

3

Clustering



4

Hierarchical Methods

5

Clustering

- **Hierarchical Clustering**

Data Matrix

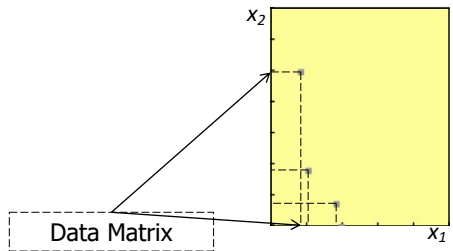
	X_1	X_2	…	X_p
I_1				
I_2				
…				
I_n				

6

NOVA IMS Information Management School

Clustering

- **Hierarchical Clustering**



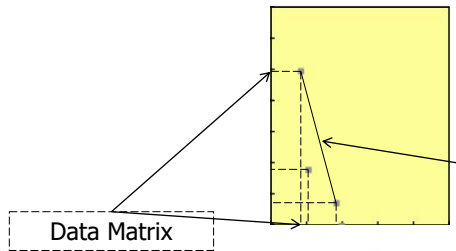
	X_1	X_2
I_1		
I_2		
…		
I_n		

7

NOVA IMS Information Management School

Clustering

- **Hierarchical Clustering**



$$d_{ij} = \sqrt{\sum_{v=1}^{p} (x_{iv} - x_{jv})^2}$$

Dissimilarity Matrix

	X_1	X_2
I_1		
I_2		
…		
I_n		

	I_1	I_2	…	I_n
I_1	0			
I_2	$d(I_2, I_1)$	0		
…	$d(I_{..}, I_1)$	$d(I_{..}, I_2)$	0	
I_n	$d(I_n, I_1)$	$d(I_n, I_2)$	$d(I_n, I_{...})$	0

8

NOVA IMS Information Management School

Clustering

	BA	FI	MI	NA	RM	TO
BA	0	662	877	255	412	996
FI	662	0	295	468	268	400
MI	877	295	0	754	564	138
NA	255	468	754	0	219	869
RM	412	268	564	219	0	669
TO	996	400	138	869	669	0

9

NOVA IMS Information Management School

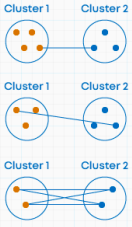
Clustering

- **Hierarchical Clustering**
  - Linkage or Aggregation Rules

- **Single Linkage**
  $D(c_1,c_2) = \min D(x_1,x_2)$
  Minimum distance or distance between closest elements in clusters

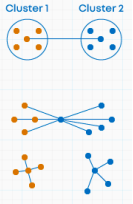
- **Complete Linkage**
  $D(c_1,c_2) = \max D(x_1,x_2)$
  Maximum distance between elements in clusters

- **Average Linkage**
  $D(c_1,c_2) = \frac{1}{|c_1|}\frac{1}{|c_2|}\sum\sum D(x_1,x_2)$
  Average of the distances of all pairs



- **Centroid Method**
  Combining clusters with minimum distance between the centroids of the two clusters

- **Ward's Method**
  - Combining clusters where increase in within cluster variance is to the smallest degree.
  - Objective is to minimize the total within cluster vairance



10

NOVA IMS Information Management School

Clustering

	BA	FI	MI/TO	NA	RM
BA	0	662	877	255	412
FI	662	0	295	468	268
MI/TO	877	295	0	754	564
NA	255	468	754	0	219
RM	412	268	564	219	0

11

NOVA IMS Information Management School

Clustering

	BA	FI	MI/TO	NA/RM
BA	0	662	877	255
FI	662	0	295	268
MI/TO	877	295	0	564
NA/RM	255	268	564	0

12

NOVA IMS Information Management School

Clustering

	BA/NA/RM	FI	MI/TO
BA/NA/RM	0	268	564
FI	268	0	295
MI/TO	564	295	0

13

NOVA IMS Information Management School

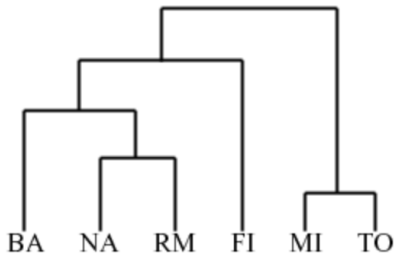
Clustering

	BA/FI/NA/RM	MI/TO
BA/FI/NA/RM	0	295
MI/TO	295	0

14

NOVA IMS Information Management School

Clustering



15

NOVA IMS Information Management School

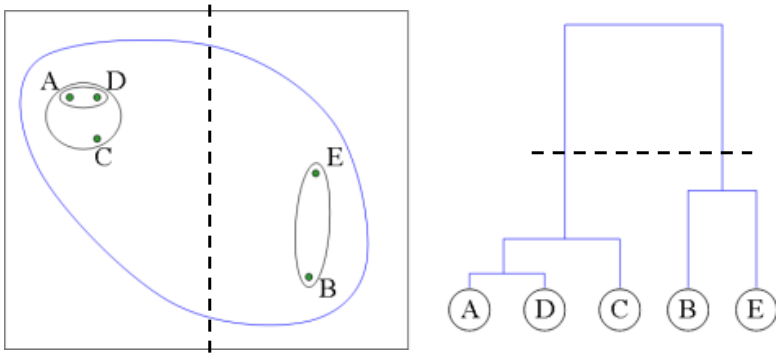
Clustering

- **Hierarchical Clustering**
  - Hierarchical Clustering - Interactive demo
    - You can find a nice worked example of hierarchical clustering at:

https://matteucci.faculty.polimi.it/Clustering/tutorial_html/hierarchical.html

16

- **Hierarchical Clustering**
  - Dendrogram



17

- **Hierarchical Clustering (disadvantages)**
  - Have an essencial problem, once an interaction is performed (merge or separation) we **cannot go back**;
    - This strictness is useful in terms of **computacional costs**, because it avoids the cost that originates from the different combinatorial choices;
    - However, this low cost is related to the **impossibility to correct wrong decisions**;
  - Time complexity: not suitable for large datasets
    - Due to its calculation needs, several operations with large matrices **do not always work very well with large datasets**.

18

- **Hierarchical Clustering (other variants)**
  - There are two ways of improving the performance of hierarchical methods:
    - To perform a careful analysis of the links produced in each hierarchical partition (CURE and Chameleon methods);
    - To integrate hierarchical clustering and optimization, first using an agglomerative algorithm and then refining the results by using iterative optimization (BIRCH method).

19

K-Means Algorithm

20

NOVA IMS Information Management School

Clustering

- **K-means algorithm**
  - K-means is a partitional clustering algorithm
  - Let the set of data points (or instances) D be

$$\{x_1, x_2, \ldots, x_n\},$$

where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ir})$ is a vector in a real-valued space $X \in R^r$, and r is the number of attributes (dimensions) in the data.

  - The k-means algorithm partitions the given data into k clusters.
    - Each cluster has a cluster center, called centroid.
    - k is specified by the user
    - $k \ll n$.

NOVA IMS Information Management School

Clustering

- **K-means algorithm**
  - Classifies the data into K groups, by satisfying the following requirements:
    - each group contains at least one point;
    - each point belongs to exactly one cluster.

- **K-means algorithm**
  - Given k, the partition method creates an initial partition (typically randomly);
  - Next, uses an iterative relocation technique that tries to improve the partition, moving objects from one group to another;
  - Generically, the criterion for a good partitioning is that of objects belonging to the same cluster should be close or related to each other.

- **K-means algorithm**
  - Algorithm:
    1. Choose the seeds;
    2. Each individual is associated with the nearest seed;
    3. Calculate the centroids of the formed clusters;
    4. Go back to step 2;
    5. End when the centroids cease to be recentered.

NOVA IMS Information Management School

Clustering

- **K-means algorithm**
  - The goal is to minimize intra-group variance (sum of squared error):

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

  - x is a data point in cluster C_i and m_i is the representative point (centroid) for cluster C_i
  - One easy way to reduce SSE is to increase K (number of clusters)
  - A good clustering with smaller K can have a lower SSE than a poor clustering with higher K

25

NOVA IMS Information Management School

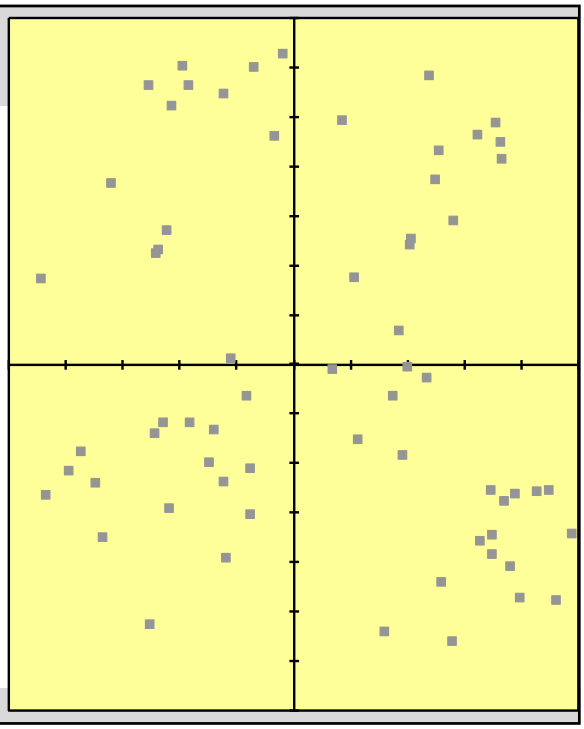
K-Means Algorithm in figures

26

NOVA IMS Information Management School

Clustering

- **K-means**
  - Individuals measured based on two variables;
  - The goal is to group them in homogeneous sets.



Source: Fiona Cameron, Techniques for Neighbourhood Classification

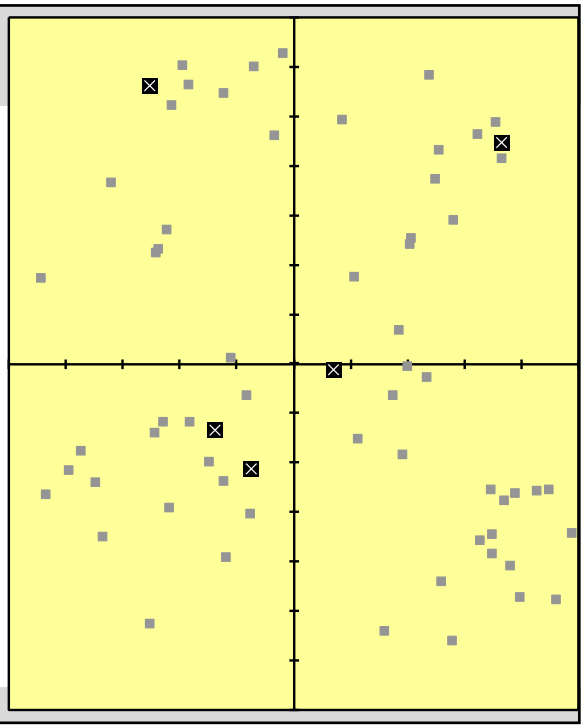
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

27

NOVA IMS Information Management School

Clustering

- **K-means**
  - Seeds
  - Randomly chosen



Source: Fiona Cameron, Techniques for Neighbourhood Classification

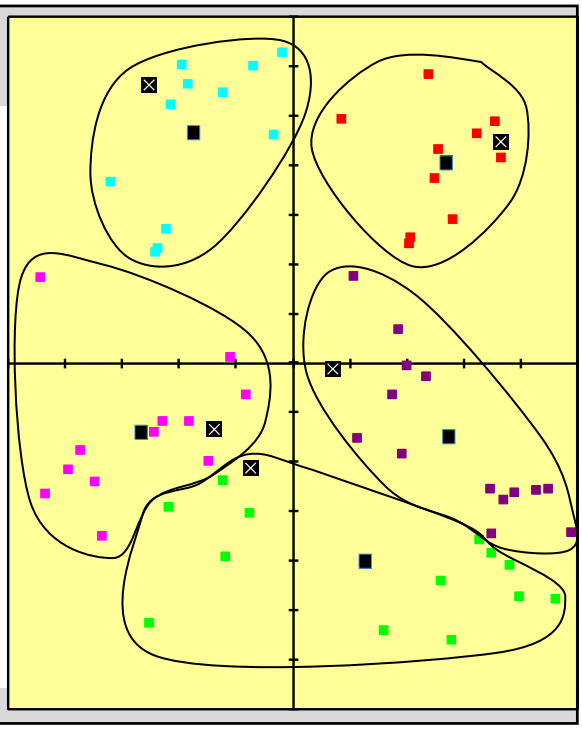
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

28

NOVA IMS Information Management School

Clustering

- **K-means**
- Allocate individuals to the nearest seed



Source: Fiona Cameron, Techniques for Neighbourhood Classification

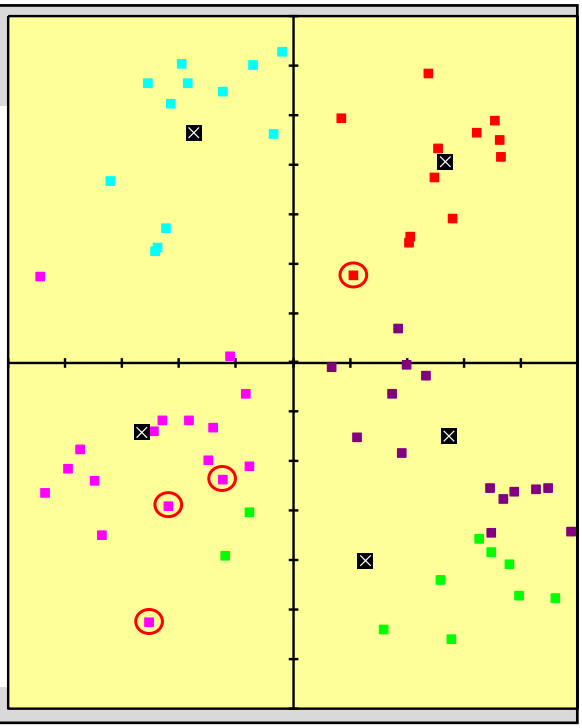
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

29

NOVA IMS Information Management School

Clustering

- **K-means**
- Recenter the seed so that it stays in the center of the cloud of points (called centroid)
- Some individuals change cluster



Source: Fiona Cameron, Techniques for Neighbourhood Classification

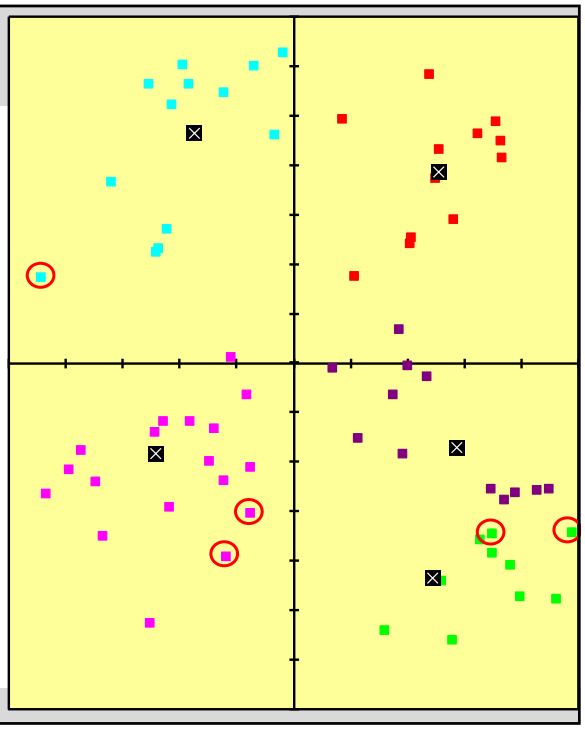
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

30

NOVA IMS Information Management School

Clustering

- **K-means**



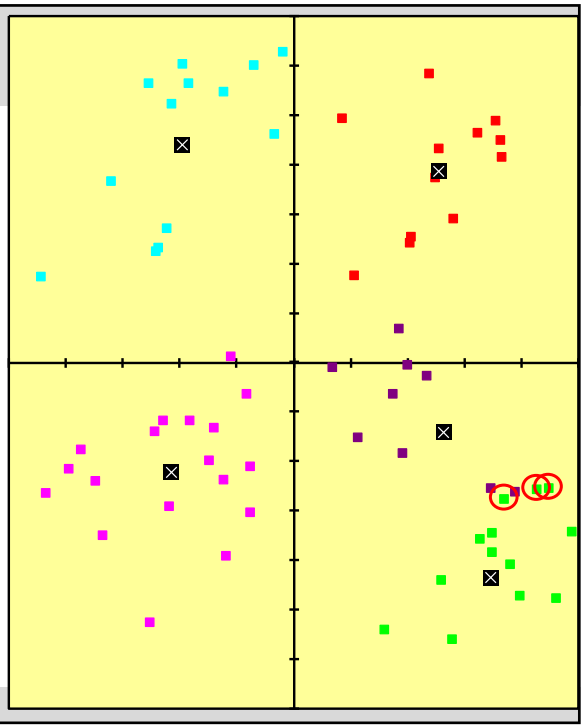
Source: Fiona Cameron, Techniques for Neighbourhood Classification

31

NOVA IMS Information Management School

Clustering

- **K-means**



Source: Fiona Cameron, Techniques for Neighbourhood Classification

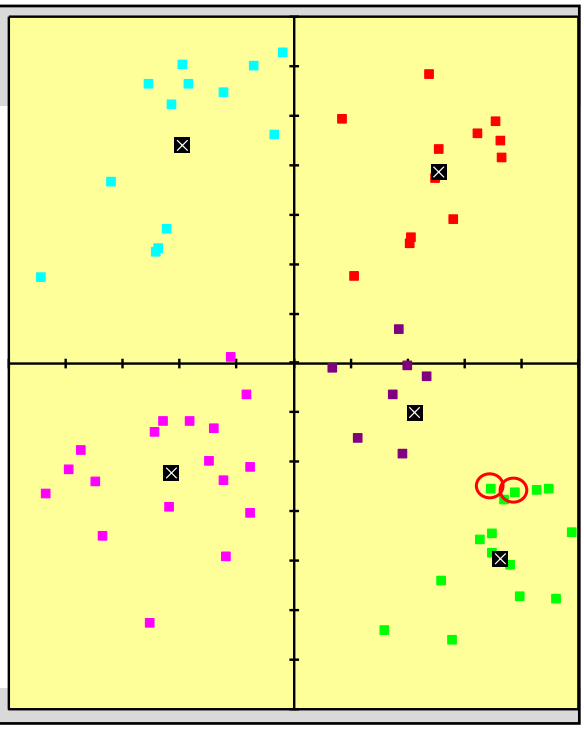
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

32

NOVA IMS Information Management School

Clustering

- **K-means**



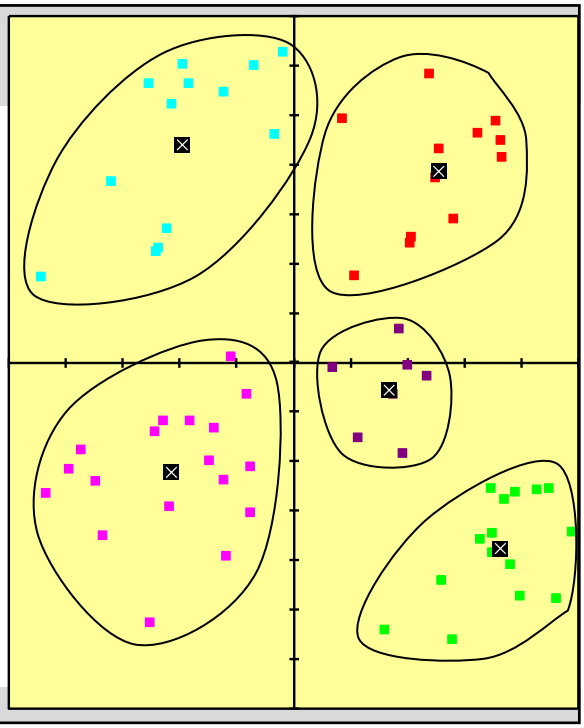
Source: Fiona Cameron, Techniques for Neighbourhood Classification

33

NOVA IMS Information Management School

Clustering

- **K-means**
- Final solution



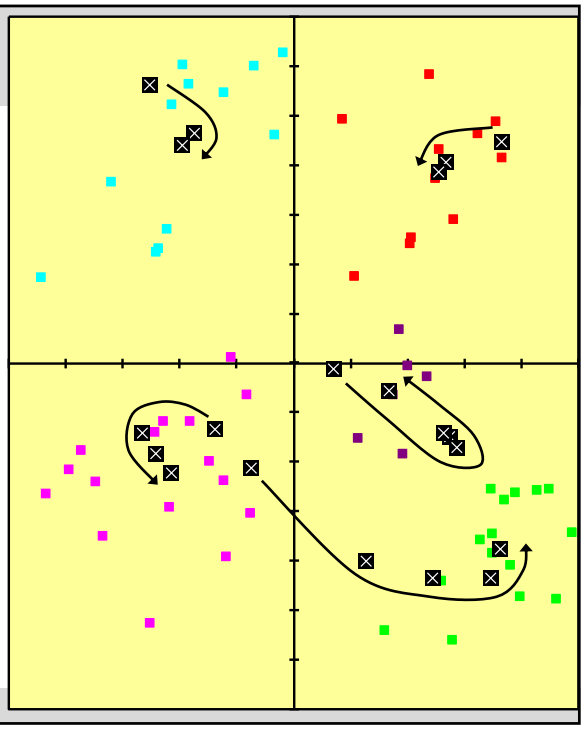
Source: Fiona Cameron, Techniques for Neighbourhood Classification

34

NOVA IMS Information Management School

Clustering

- **K-means**
- Movement of centroids during optimization process



Source: Fiona Cameron, Techniques for Neighbourhood Classification

35

NOVA IMS Information Management School

Clustering

- **K-means algorithm (strengths)**
  - Simple: easy to understand and to implement
  - Efficient: Time complexity O(tkn),
    - where n is the number of data points,
    - k is the number of clusters, and
    - t is the number of iterations.
  - Since both k and t are small, k-means is considered a linear algorithm.
  - K-means is the most popular clustering algorithm.
  - Note that: it terminates at a local optimum if SSE is used. The global optimum is hard to find due to complexity.

36

- **K-means algorithm (weaknesses)**
  - Very sensitive to the existence of outliers;
  - Very sensitive to the initial positions of the seeds;
  - Partitioning methods work well with spherical-shaped clusters;
    - Partitioning methods are not the most suitable to find clusters with complex shapes and different densities;
  - The need to set from the start the number of clusters to create.

37

The Algorithm variant

38

NOVA IMS Information Management School

Clustering

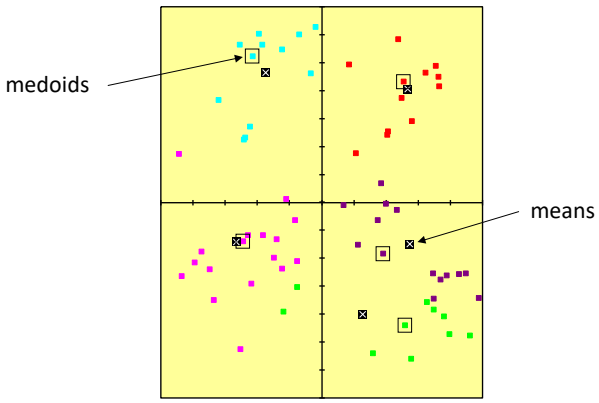
- **K-means and k-medoids algorithms**
  - Most algorithms adopt one of two very popular heuristics:
    - k-means algorithm, where each cluster is represented by the average of the values of the points in a cluster;
    - k-medoids algorithm, where each cluster is represented by one of the points located near the center of the cluster.

39

NOVA IMS Information Management School

Clustering

- **K-means and k-medoids algorithms**



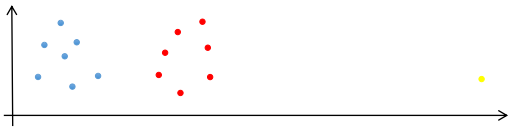
Source: Fiona Cameron, Techniques for Neighbourhood Classification

40

NOVA IMS Information Management School

Clustering

- **K-means and k-medoids algorithms**



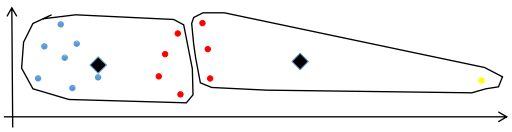
Source: CS583, Bing Liu, UIC, Chapter 4: Unsupervised Learning

41

NOVA IMS Information Management School

Clustering

- **K-means and k-medoids algorithms**



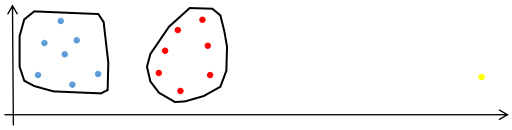
Source: CS583, Bing Liu, UIC, Chapter 4: Unsupervised Learning

42

NOVA IMS Information Management School

Clustering

- **K-means and k-medoids algorithms**



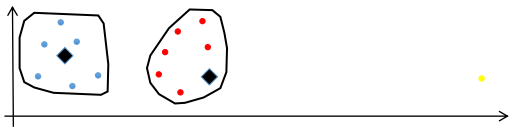
Source: CS583, Bing Liu, UIC, Chapter 4: Unsupervised Learning

43

NOVA IMS Information Management School

Clustering

- **K-means and k-medoids algorithms**



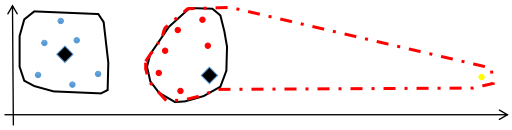
Source: CS583, Bing Liu, UIC, Chapter 4: Unsupervised Learning

44

NOVA IMS Information Management School

Clustering

- **K-means and k-medoids algorithms**



Source: CS583, Bing Liu, UIC, Chapter 4: Unsupervised Learning

45

NOVA IMS Information Management School

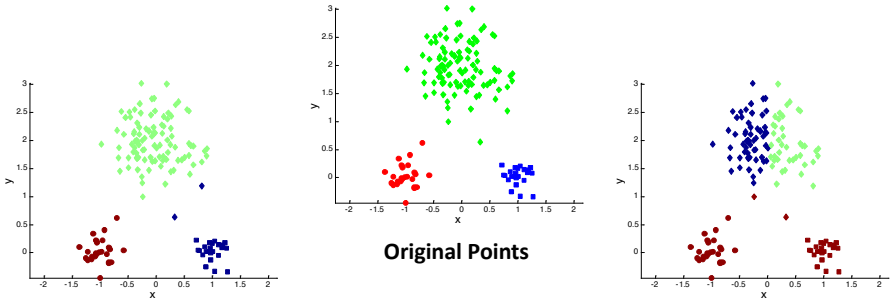
The initialization problem

46

NOVA IMS Information Management School

Clustering

- **K-means algorithm (weaknesses)**
  - The algorithm is sensitive to initial seeds.

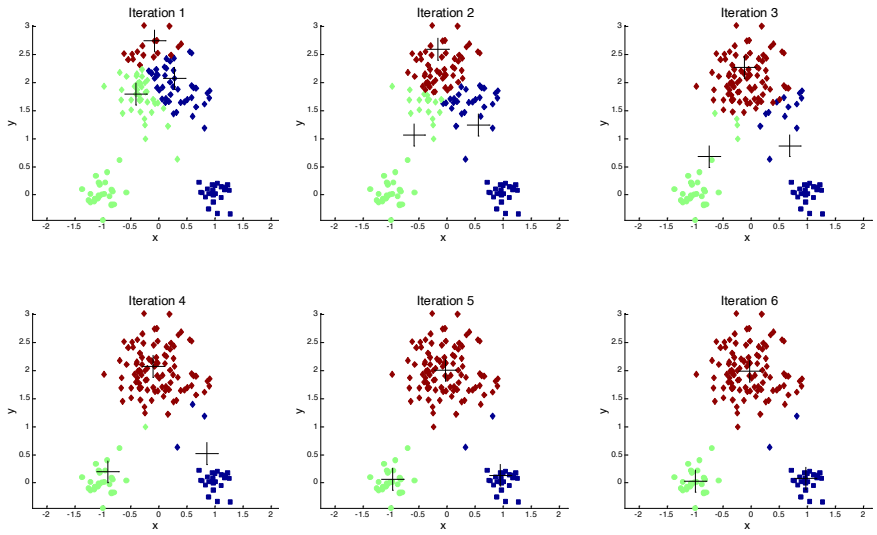


Source: Tan, Steinbach, Kumar, Introduction to Data Mining 2004

47

NOVA IMS Information Management School

Clustering

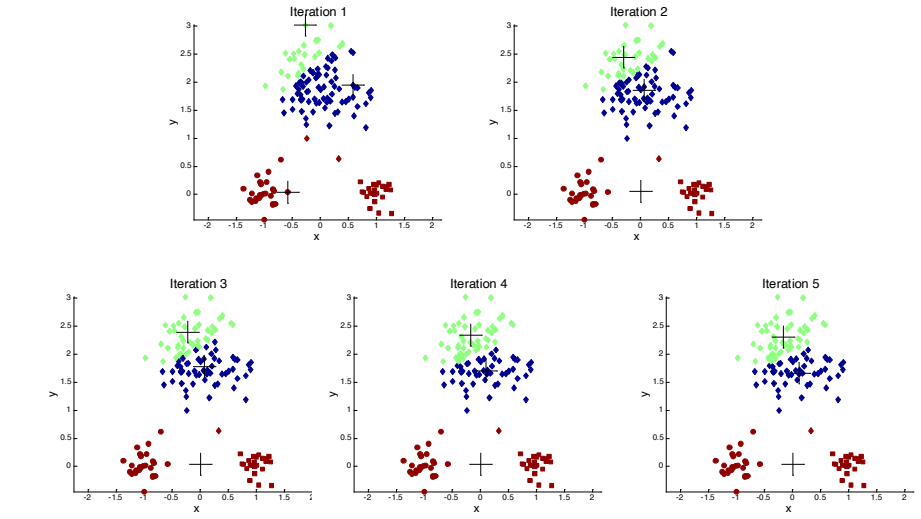


Source: Tan, Steinbach, Kumar, Introduction to Data Mining 2004

48

NOVA IMS Information Management School

Clustering



Source: Tan, Steinbach, Kumar, Introduction to Data Mining 2004

49

NOVA IMS Information Management School

Clustering

- **K-means algorithm (weaknesses)**
  - The algorithm is sensitive to initial seeds.
  - Use multiple forms of initialization;
  - Re-initialize several times;
  - Use more than one method;
  - Use a relatively large number of clusters and proceed to their regrouping by the choice of centroids.

50

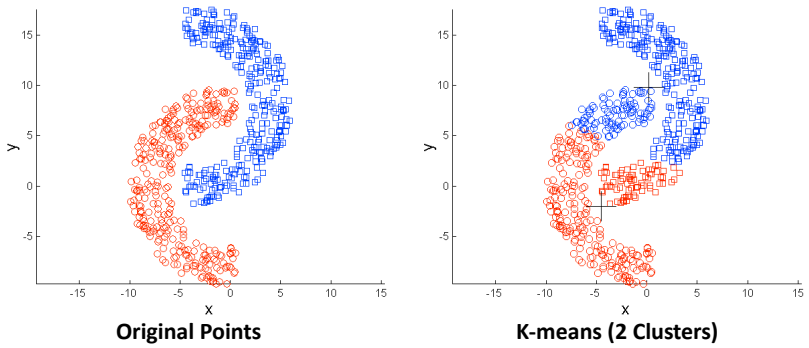
NOVA IMS Information Management School

Shape and density

51

NOVA IMS Information Management School

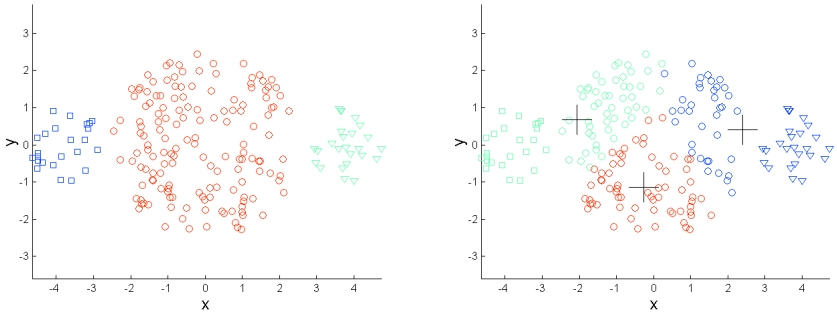
- **K-means algorithm (weaknesses)**



Original Points **K-means (2 Clusters)**

52

- **K-means algorithm (weaknesses)**
  - Have difficulties in dealing with clusters of different size and density;

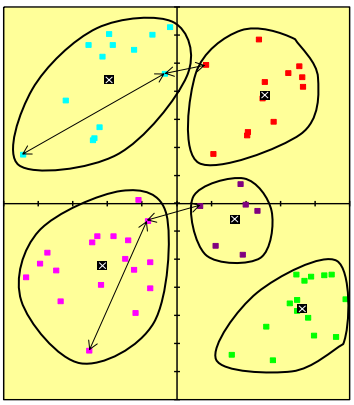


- **K-means algorithm (weaknesses)**
  - Each individual either belongs or does not belong to the cluster, having no notion of probability of belonging, in other words, there is no consideration of the quality of the representation of a particular individual in a given cluster.

NOVA IMS Information Management School

Clustering

- **K-means algorithm (weaknesses)**



Source: Fiona Cameron, Techniques for Neighbourhood Classification

55

NOVA IMS Information Management School

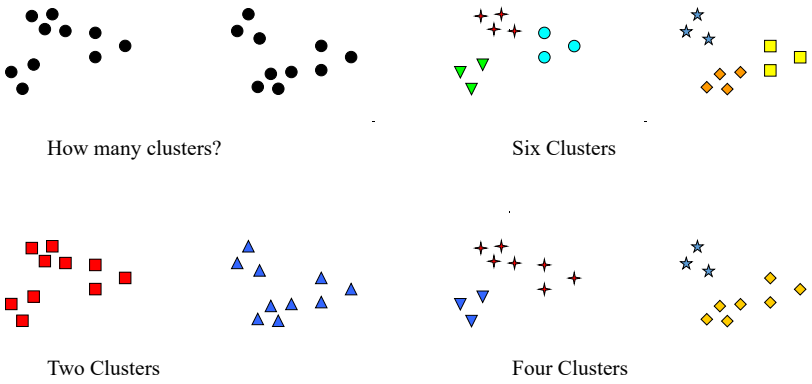
The number of clusters

56

NOVA IMS Information Management School

Clustering

- **K-means algorithm the number of clusters**



Source: Tan, Steinbach, Kumar, Introduction to Data Mining 2004

57

NOVA IMS Information Management School

Clustering

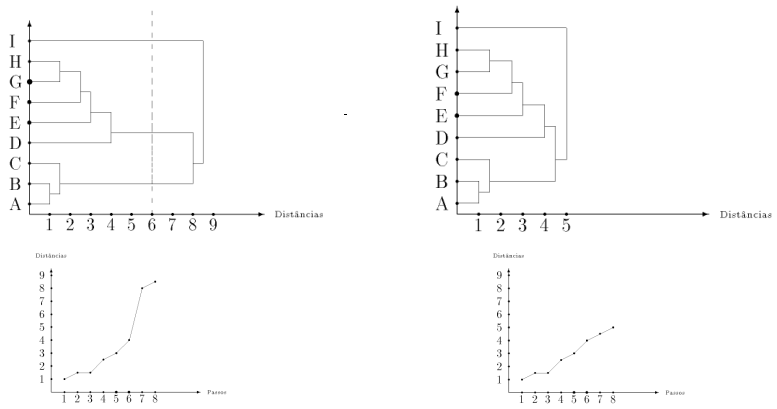
- **K-means algorithm the number of clusters**
  - This is always a difficult problem to solve, and there are no recipes to fix this.
  - One way to minimize the problem is to create various classifications with different K, and choose the best.
  - Use a hierarchical method in order to choose the number of clusters based on the dendrogram.

58

NOVA IMS Information Management School

Clustering

- **K-means algorithm the number of clusters**



59

NOVA IMS Information Management School

Clustering

- **K-means algorithm the number of clusters**
  - The choice should be guided by three fundamental criteria:
    - intra-cluster variance,
    - evaluation of the profile of the cluster (subjective),
    - operational considerations.

60

NOVA IMS Information Management School

Clustering

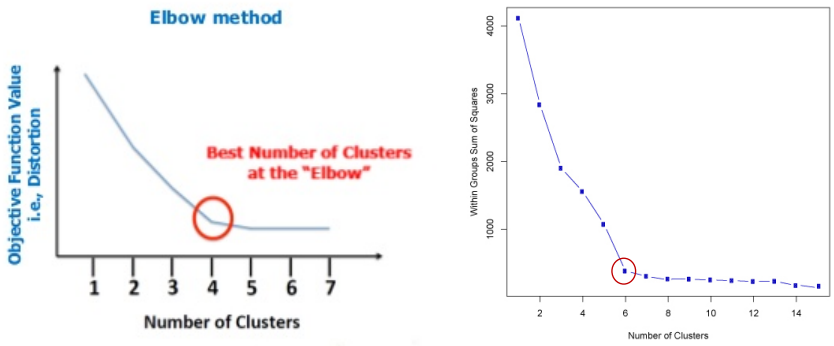
- **K-means algorithm the number of clusters**
  - Regarding the first criterion, the analysis is simple and not too subjective, since we know that the lower the intra-cluster variability the greater the cohesion of the cluster, a highly desirable feature in this type of analysis. However, as k increases, variability decreases;
  - Regarding the second criterion, the question is not as simple in the sense that it requires much more subjective assessments, which relate to the interpretation of the obtained clusters;
  - The third criterion is relatively simple in the sense that these issues are imposed on the analyst.

61

NOVA IMS Information Management School

Clustering

- **K-means algorithm the number of clusters**



62

NOVA IMS Information Management School

Clustering

- **K-means algorithm the number of clusters**
  - To test the results by varying k (number of clusters);
  - This procedure allows a series of analyzes that can instruct the choice of the number of clusters;
  - To compare the totals of the distances of the different solutions.

63

NOVA IMS Information Management School

Clustering

- **K-means algorithm the number of clusters**
  - Operational considerations are related to business environment and usually affect the decisions of the analyst:
    - A number small enough for developing a specific strategy;
    - A number of individuals large enough to be worth it to develop a specific strategy;
    - A good way to accomodate these considerations is the use of a high initial k and then proceed to the grouping of clusters.

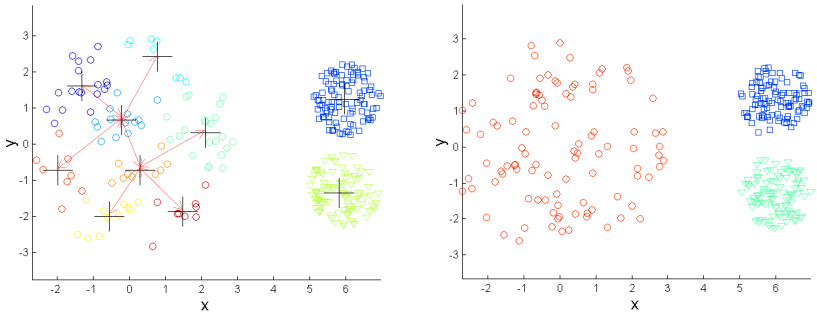
64

NOVA IMS Information Management School

Clustering

- **K-means algorithm the number of clusters**

CLUSTER	Cluster 1	Cluster 2	Cluster 3	Cluster 4
1	0	46.88621549	53.629114781	51.055735073
2	46.88621549	0	35.424488679	48.408611185
3	53.629114781	35.424488679	0	58.950971223
4	51.055735073	48.408611185	58.950971223	0



65

NOVA IMS Information Management School

Clustering

- **K-means algorithm the number of clusters**
  - This evaluation is carried out by comparing the mean values for each variable in each cluster with the mean values of the population for each variable;
  - In this case, it is particularly relevant to take into account the most important differences within the different clusters and the mean population.
  - That is why profiling is so important.

66

Data Mining

Partitioning Clustering (k-means)

Note on Profiling

17/11/2021

NOVA-IMS

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

AGENDA

- Cluster analysis
  - Clustering techniques
    - Partitioning Methods (kmeans and k-meadoids)
    - **A Note on Profiling**

2

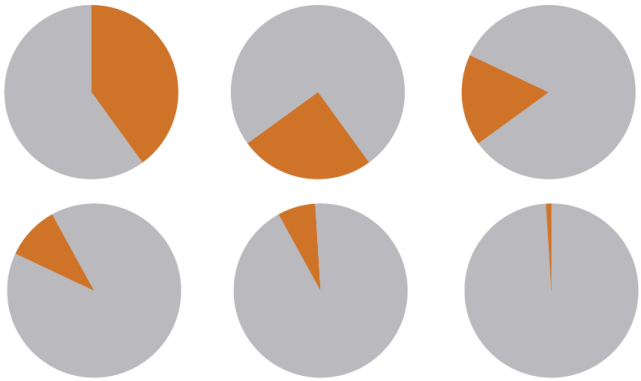
Profiling

3

Clustering

- **Profiling (size of the clusters)**

Part-to-Whole Mini-Pie Charts

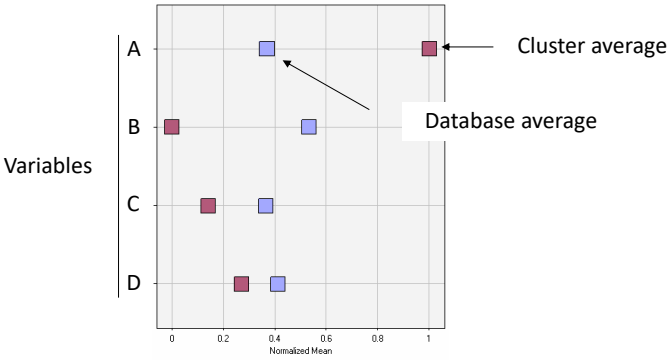


4

NOVA IMS Information Management School

Clustering

- **Profiling (comparing averages)**

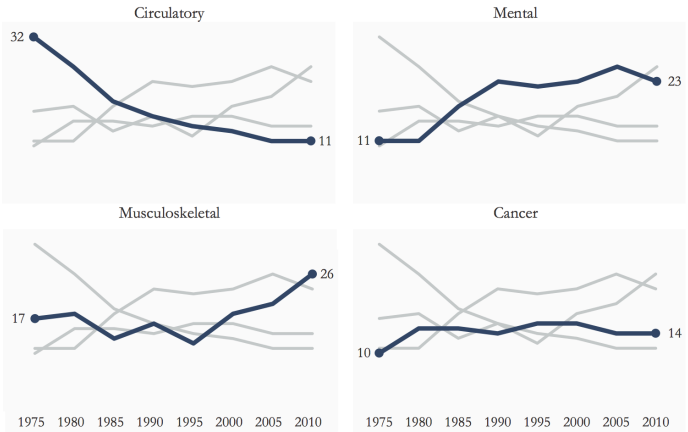


5

NOVA IMS Information Management School

Clustering

- **Profiling (comparing profiles)**



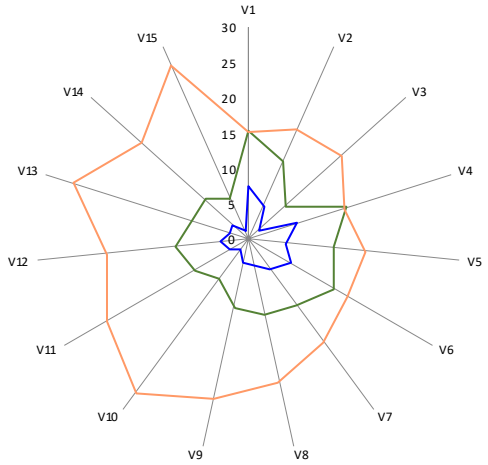
JA Schwabish "An Economist's Guide to Visualizing Data", The Journal of Economic Perspectives, 2014

6

NOVA IMS Information Management School

Clustering

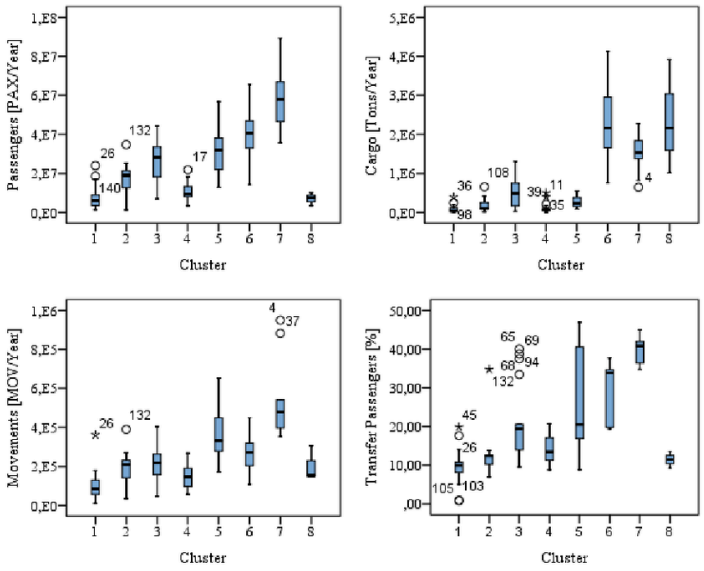
- **Profiling (comparing profiles)**



7

NOVA IMS Information Management School

Clustering

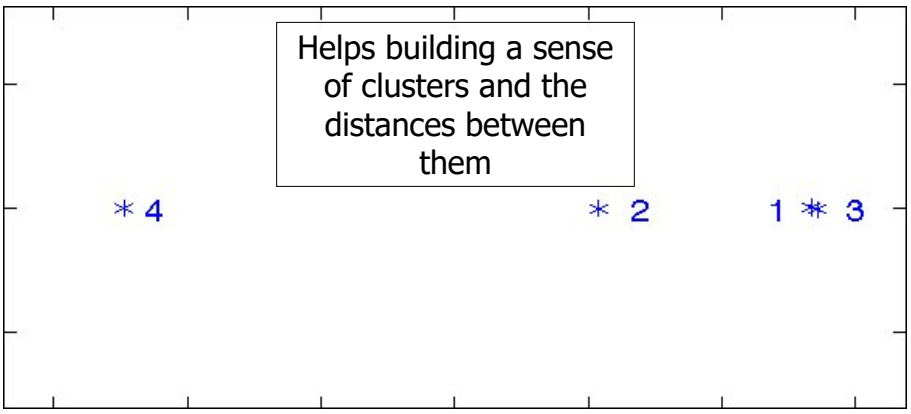


8

NOVA IMS Information Management School

Clustering

- **Profiling (multidimensional scaling)**

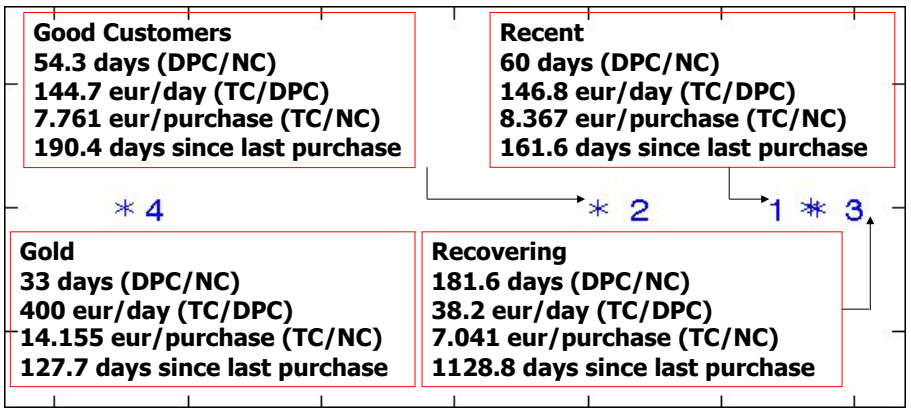


9

NOVA IMS Information Management School

Clustering

- **Profiling (multidimensional scaling)**

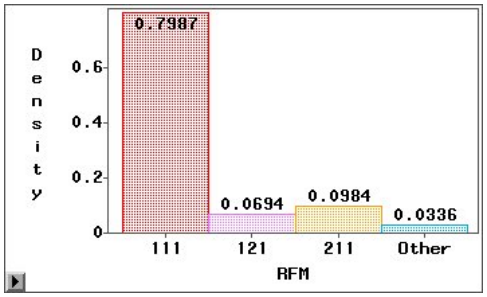


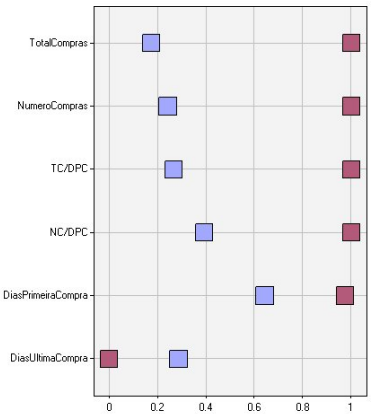
10

NOVA IMS Information Management School

Clustering

- **Profiling (comparing variables not used)**





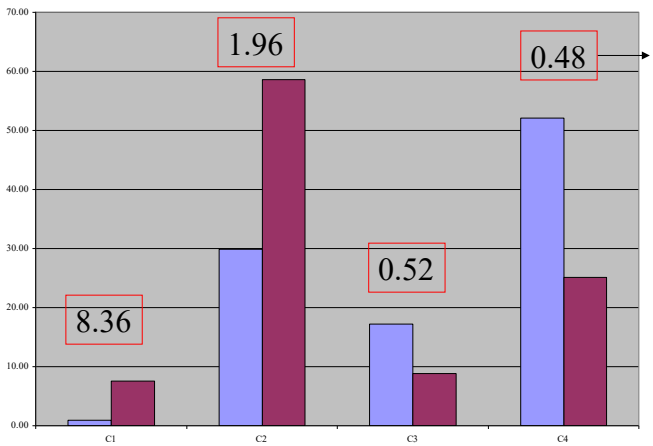
Used and unused variable distribution

11

NOVA IMS Information Management School

Clustering

- **Profiling (leverage)**



Leverage
Ratio Sales(%)/Individuals(%)

12

Clustering

- **Exploring two solutions**

		Sofist_Digital	Variedade Serviço	Preço	Status_PC	ProAtividade	Tempo Internet e RS	Sustentabilidade	Life Status	Propensao_Pfidelizacao	Switching	
k=5	_FREQ_	F1_SD	F2_VS	F3_P	F4_SPC	F5_PA	F6_IRS	F7_CC	F8_LS	F9_FI	F10_AB	
1	222	-0,40	**-0,54**	**-0,31**	-0,29	0,13	**0,65**	**-1,67**	-0,20	-0,09	-0,03	Pouco Envolvidos e pouco sensiveis ao preço
2	463	**0,70**	**0,44**	**0,38**	0,13	**-0,32**	**-0,37**	-0,19	**0,31**	**-0,58**	0,03	Social Customers
3	469	**0,58**	-0,16	0,15	**0,48**	**0,23**	0,00	0,09	**-0,22**	**0,98**	**-0,04**	Sofisticado com Propensao Fidelizacao
4	428	**-1,47**	0,27	0,25	-0,13	0,01	**-0,25**	0,29	0,15	0,15	**0,04**	Info Excluidos
5	420	0,29	-0,30	**-0,68**	**-0,39**	0,02	0,31	**0,69**	-0,16	**-0,56**	-0,02	Sustentaveis
k=6	FREQ	F1_SD	F2_VS	F3_P	F4_SPC	F5_PA	F6_IRS	F7_CC	F8_LS	F9_FI	F10_AB	
1	241	0,12	**-0,30**	-0,09	-0,20	-0,04	**1,34**	**-1,29**	-0,03	-0,13	-0,24	Pouco Envolvidos
2	134	-0,16	0,11	0,02	-0,15	-0,10	0,13	0,10	0,00	0,07	**2,84**	Switchers
3	285	0,43	**0,48**	**-1,28**	-0,16	**0,52**	**-0,28**	0,05	0,04	**0,23**	-0,07	Exigentes
4	426	**-1,42**	0,22	**0,30**	-0,20	-0,03	**-0,27**	0,10	0,01	**0,21**	-0,27	Info Excluidos
5	414	0,39	**-0,77**	0,20	**0,92**	0,19	**-0,23**	0,15	**0,21**	0,15	-0,16	Status/Conservadores
6	502	**0,63**	**0,30**	**0,34**	**-0,37**	**-0,38**	-0,10	**0,35**	**-0,19**	**-0,39**	-0,24	Social Customers

Clustering

- **Exploring two solutions with different k**

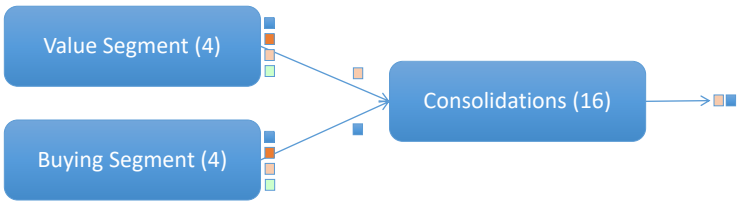
K=7 VS K=8	Pouco envolvidos	Switchers	Informados / Social Customers	Fidelizáveis Sofisticados	Info Excluidos	Status	Proativos	Desprendido	Grand Total
Pouco envolvidos	71		23	13	45	1	1	49	**203**
Switchers	2	106	4	4	2	1		3	**122**
Informados / Social Customers	4		219	68	2	20	1	64	**378**
Fidelizáveis Sofisticados	72		105	191	4	1	2	69	**444**
Status/Conservadores	7		5	39	13	157		42	**263**
Proativos	5		15	20	3		153	19	**215**
Info Excluidos			12	4	309	1	1	50	**377**
Grand Total	**161**	**106**	**383**	**339**	**378**	**181**	**158**	**296**	**2002**

K=6 VS K=8	Pouco envolvidos	Switchers	Informados / Social Customers	Fidelizáveis Sofisticados	Info Excluidos	Status	Proativos	Desprendido	Grand Total
Pouco Envolvidos	146		16	24	8	17	7	23	**241**
Switchers	2	105	6	11	1	2	4	3	**134**
Exigentes	1	1	26	56	13	27	60	101	**285**
Info Excluidos				22	350	17	11	26	**426**
Status/Conservadores	3		13	177	5	98	58	60	**414**
Social Customers	9		322	49	1	20	18	83	**502**
Grand Total	**161**	**106**	**383**	**339**	**378**	**181**	**158**	**296**	**2002**

NOVA IMS Information Management School

Clustering

- **Consolidation of two solutions**



Data Mining

Density-based Clustering (dbscan)

16/11/2021

NOVA-IMS

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

AGENDA

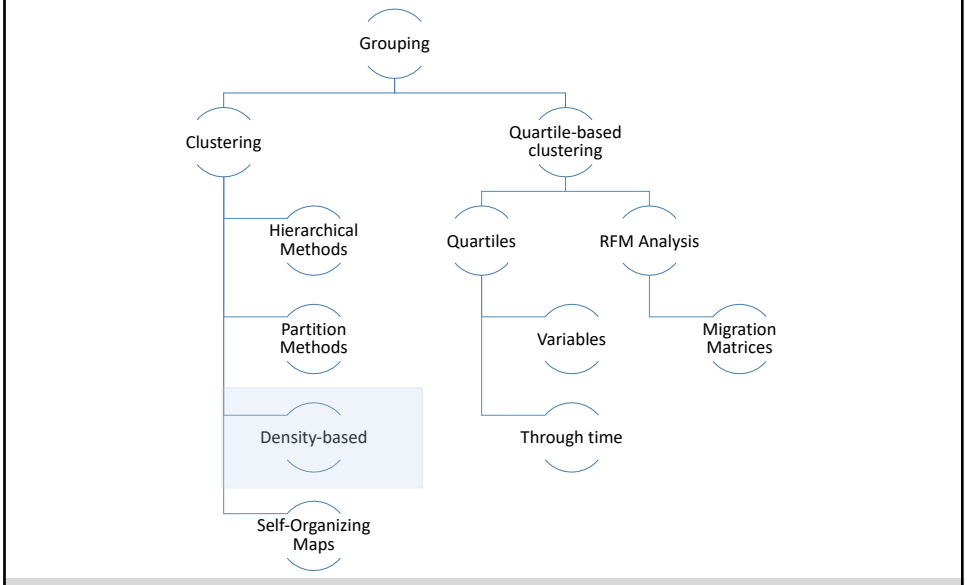
- Cluster analysis
  - Clustering techniques
    - Density-based clustering (DBscan)

2

Clustering

3

Clustering



4

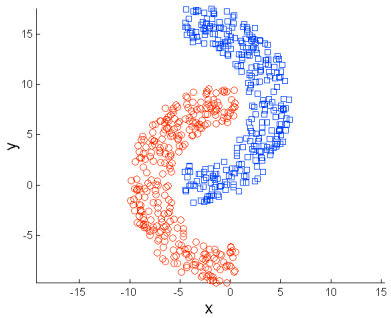
Density-Based Clustering (DBSCAN)

5

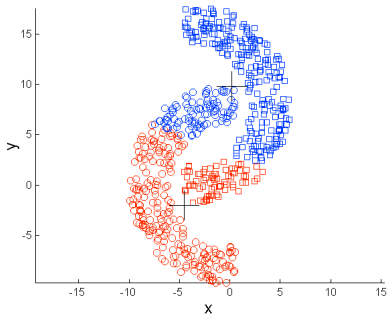
- **The idea**
  - Previous clustering methods have some limitations. They are based on a particular set of assumptions that if not true, the process yields suboptimal results.
  - That is they would likely inaccurately identify non-convex regions, and where noise or outliers are included in the clusters.

6

Density-Based Clustering (DBSCAN)



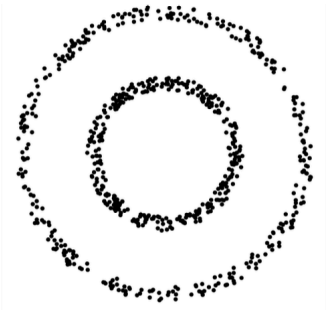
Original Points



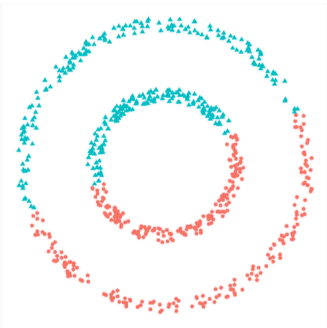
K-means (2 Clusters)

7

Density-Based Clustering (DBSCAN)



Original Points



K-means (2 Clusters)

8

Density-Based Clustering (DBSCAN)

- **The idea**
  - Density-based clustering algorithms try to **find clusters in data without assuming a particular shape**. Thus solving correctly cases like the ring example.
  - To find clusters of **arbitrary shape**, these methods model clusters as **dense regions in data space**, separated by sparse regions.

9

Density-Based Clustering (DBSCAN)

- **The idea**
  - We want to be able to cluster data like this:

10

Density-Based Clustering (DBSCAN)

- **Characteristics**
  - What do talk about when we talk about density?
  - What is our physical intuition behind density?
    - We say a point ***p*** is in a dense region if there are **many points** in the neighborhood of ***p***. In this sense, the density around a point can be measured by the number of points surrounding it.
    - To measure the density around a point ***p*** we use the tradition topological definition of neighborhood. The ε-*neighborhood* of a point ***p*** is the space within a radius $\varepsilon > 0$ centered in ***p***.

11

Density-Based Clustering (DBSCAN)

- **DBSCAN:** Density-Based Clustering Based on Connected Regions with High Density
  - DBSCAN takes two input parameters:
    - ε - the radius defining the neighborhood
    - *MinPts* - the minimum of points in ε -neighborhood

12

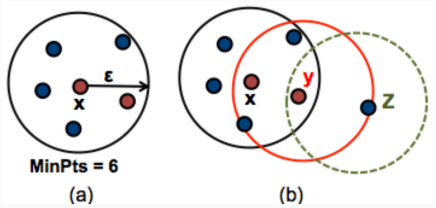
Density-Based Clustering (DBSCAN)

- **DBSCAN:** Density-Based Clustering Based on Connected Regions with High Density
  - DBSCAN importante concepts:
    - If ***z*** is a point that have at least *MinPts* in its ε -neighborhood is called **core point**.
    - ***x*** is **border point**, if the number of its neighbors is less than *MinPts*, but it belongs to the ε -neighborhood of some core point ***z***
    - If a point is neither a **core** nor a **border** point, then it is called a **noise point** or an **outlier**.

13

Density-Based Clustering (DBSCAN)

- **DBSCAN:** Density-Based Clustering Based on Connected Regions with High Density



  - Assuming MinPts = 6.
    - **x** is a **core point** because ε -neighborhood (x) = 6,
    - **y** is a **border point** because ε –neighborhood (y) < *MinPts*, but it belongs to the ε -neighborhood of the core point x.
    - **z** is a **noise point**.

14

Density-Based Clustering (DBSCAN)

- **DBSCAN:** Density-Based Clustering Based on Connected Regions with High Density
  - We define 3 terms, required for understanding the DBSCAN algorithm:
    - **Direct density reachable**: A point "A" is directly density reachable from another point "B" if:
      - i) "A" is in the ε –neighborhood of "B" and
      - ii) "B" is a core point.

15

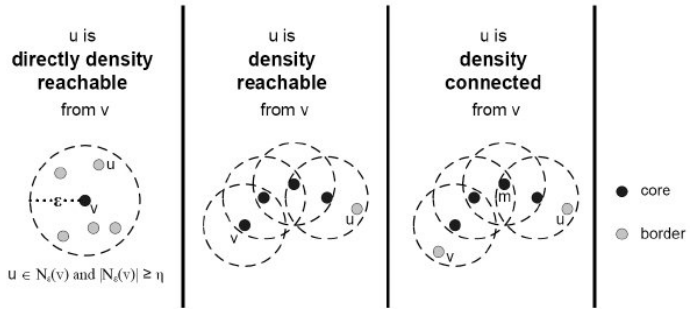
Density-Based Clustering (DBSCAN)

- **DBSCAN:** Density-Based Clustering Based on Connected Regions with High Density
  - We define 3 terms, required for understanding the DBSCAN algorithm:
    - **Density reachable**: A point "A" is density reachable from "B" if there are a set of core points leading from "B" to "A.
    - **Density connected**: Two points "A" and "B" are density connected if there are a core point "C", such that both "A" and "B" are density reachable from "C".

16

Density-Based Clustering (DBSCAN)

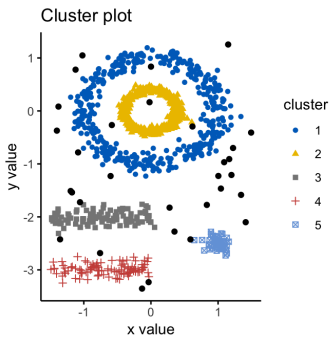
- **DBSCAN:** Density-Based Clustering Based on Connected Regions with High Density



17

Density-Based Clustering (DBSCAN)

- **DBSCAN:** Density-Based Clustering Based on Connected Regions with High Density
  - A **density-based cluster** is defined as a **group of density connected points**.



18

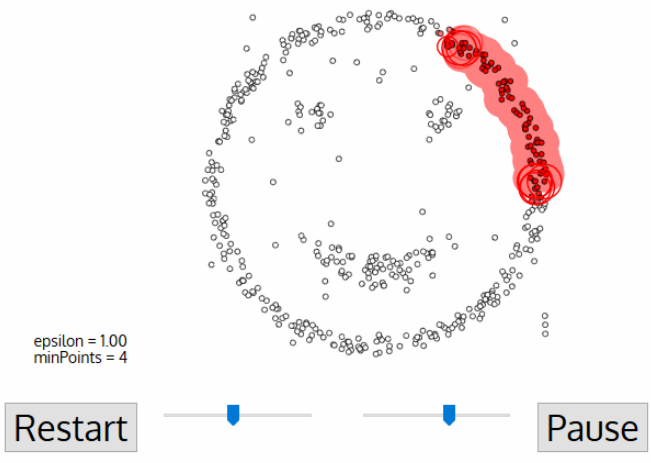
Density-Based Clustering (DBSCAN)

- **Algorithm**

1. DBSCAN begins with an arbitrary data point that has not been visited. The ε –neighborhood of this point is extracted (all points which are within the ε distance are neighborhood points).
2. If there are a sufficient number of points (according to *MinPts*) within this neighborhood then the clustering process starts and the current data point becomes the first point in the new cluster. Otherwise, the point will be labeled as noise. In both cases that point is marked as "visited".
3. For this first point in the new cluster, the points within its ε distance neighborhood also become part of the same cluster. This procedure of making all points in the ε –neighborhood belong to the same cluster is then repeated for all of the new points that have been just added to the cluster group.
4. This process of steps 2 and 3 is repeated until all points in the cluster are determined i.e all points within the ε –neighborhood of the cluster have been visited and labeled.

19

Density-Based Clustering (DBSCAN)



https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/

20

Density-Based Clustering (DBSCAN)

- **Advantages**
  - It does not require a pre-set number of clusters at all.
  - It identifies outliers as noises (not affected by),
  - it can finds arbitrarily sized and arbitrarily shaped clusters quite well.
- **Disadvantages**
  - It doesn't perform well when the clusters are of varying density
  - Setting of the distance threshold ε and *MinPts* for identifying the neighborhood points will vary from cluster to cluster when the density varies
  - Doesn't work well in high-dimensional spaces

21

Questions?

22

NOVA IMS Information Management School

Density-Based Clustering (DBSCAN)

- **DBSCAN:** Density-Based Clustering Based on Connected Regions with High Density

1. For each point $\boldsymbol{x_i}$, compute the distance between $\boldsymbol{x_i}$ and the other points. Finds all neighbor points within ε –neighborhood of the starting point $\boldsymbol{x_i}$. Each point, with a neighbor count greater than or equal to *MinPts*, is marked as **core point** or visited.
2. For each **core point**, if it's not already assigned to a cluster, create a new cluster. Find recursively all its density connected points and assign them to the same cluster as the **core point**.
3. Iterate through the remaining unvisited points in the dataset.

Those points that do not belong to any cluster are treated as outliers or noise.

Data Mining

Mean-shift algorithm

16/11/2021

NOVA-IMS

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

AGENDA

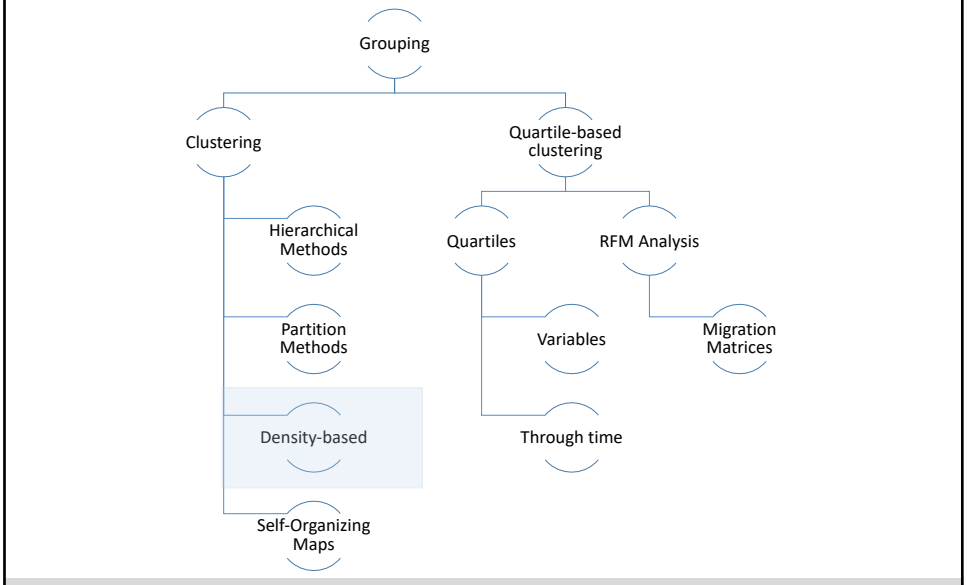
- Cluster analysis
  - Clustering techniques
    - Density-based clustering (mean shift clustering)

2

Clustering

3

Clustering



4

Mean-Shift Clustering

5

- **Characteristics**
  - Mean shift clustering is a sliding-window-based algorithm that attempts to find dense areas of data points.
  - It is a centroid-based algorithm meaning that the goal is to locate the center points of each group/class,
    - Works by updating candidates for center points to be the mean of the points within the sliding-window.
    - These candidate windows are then filtered in a post-processing stage to eliminate near-duplicates, forming the final set of center points and their corresponding groups.

6

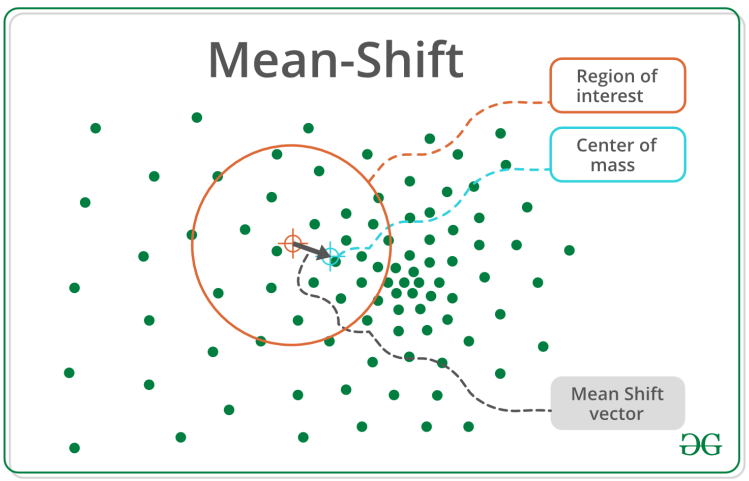
Mean-Shift Clustering

- **Algorithm**

1. Begin with a circular sliding window centered at a point C (randomly selected) and having radius r as the kernel.
2. At every iteration, the sliding window is shifted towards regions of higher density by shifting the center point to the mean of the points within the window (will gradually move towards areas of higher point density).
3. We continue shifting the sliding window according to the mean until there is no direction at which a shift can accommodate more points inside the kernel.
4. This process of steps 1 to 3 is done with many sliding windows until all points lie within a window. When multiple sliding windows overlap the window containing the most points is preserved. The data points are then clustered according to the sliding window in which they reside.

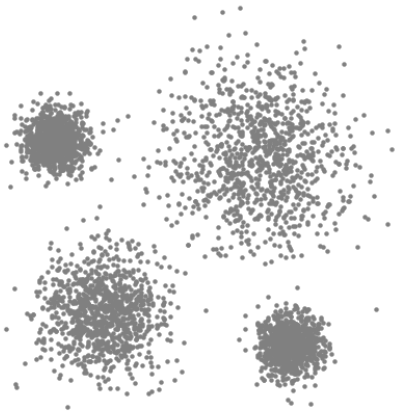
7

Mean-Shift Clustering



8

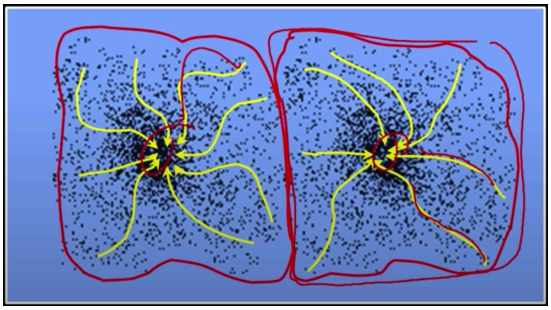
Mean-Shift Clustering



9

Mean-Shift Clustering

- **Characteristics**
  - **Cluster**: all the data points in the attraction basin of a mode
  - **Attraction basin**: the region where all trajectories lead to the same mode



10

11

Data Mining

Notes on Clustering Strategies

24/11/2021

NOVA-IMS

Fernando Lucas Bação

bacao@isegi.unl.pt

http://www.isegi.unl.pt/fbacao

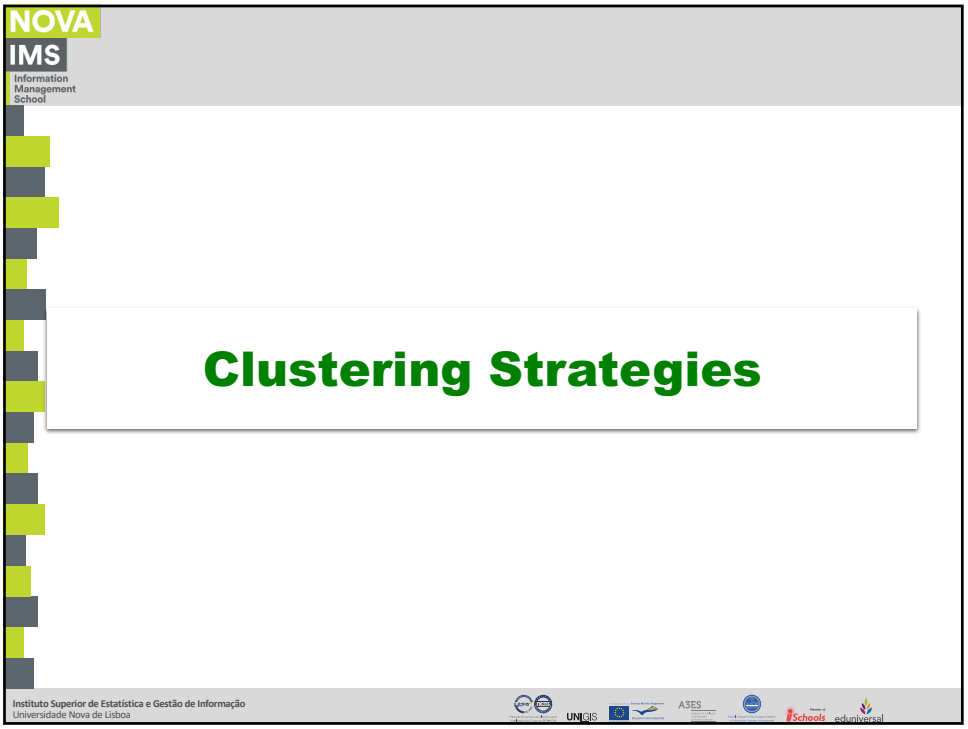
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

1

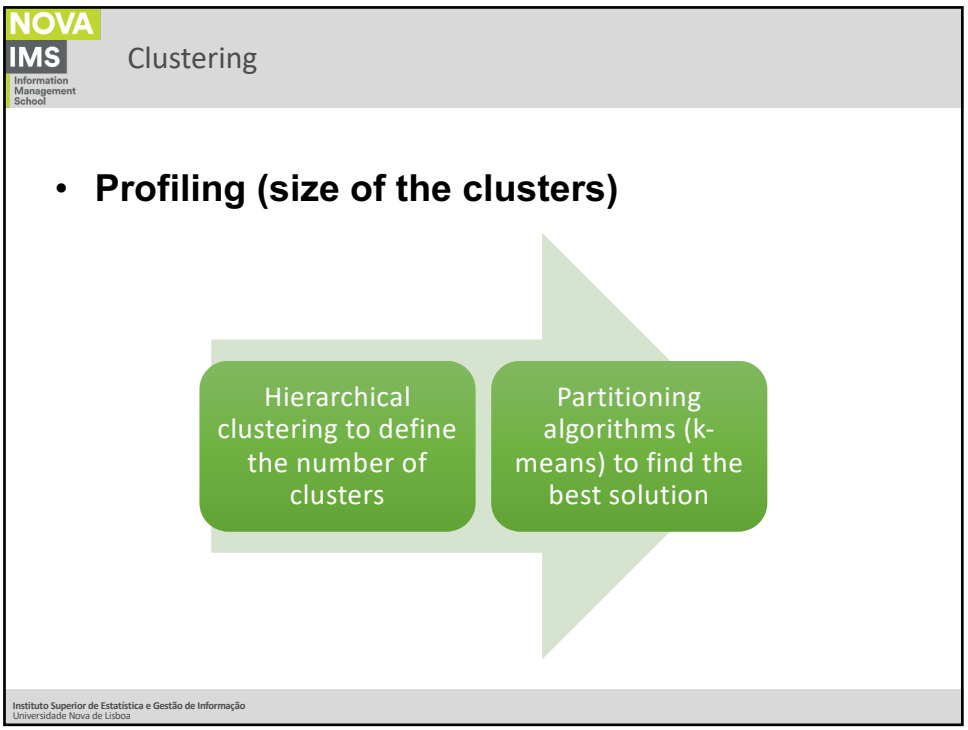
AGENDA

- Cluster analysis
  - Clustering Strategies
  - Reintroducing the Outliers or New Observations

2


NOVA
IMS
Information Management School
Clustering Strategies
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

3


NOVA
IMS
Information Management School
Clustering
• Profiling (size of the clusters)
Hierarchical clustering to define the number of clusters
Partitioning algorithms (k-means) to find the best solution
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

4

NOVA IMS Information Management School

Clustering

- **Profiling (size of the clusters)**

Partitioning algorithm (k-means) with a large number of clusters

Hierarchical algorithm to find the appropriate number of clusters

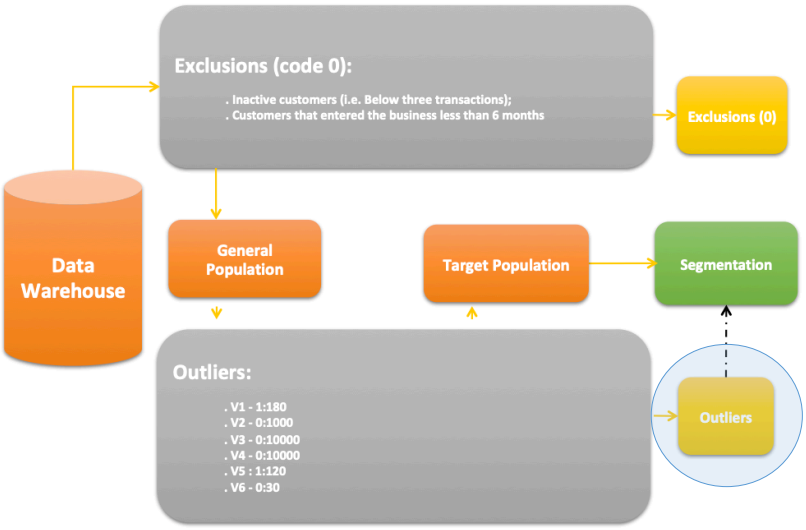
5

NOVA IMS Information Management School

Reintroducing the Outliers or New Observations

6

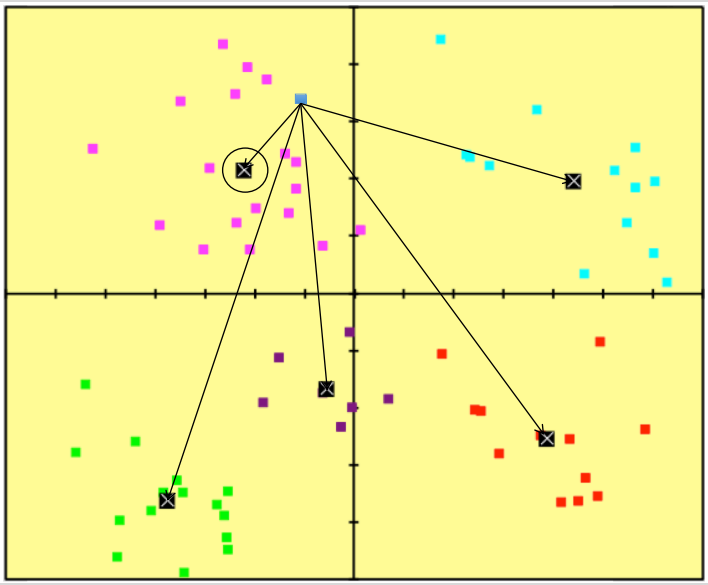
Clustering



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

7

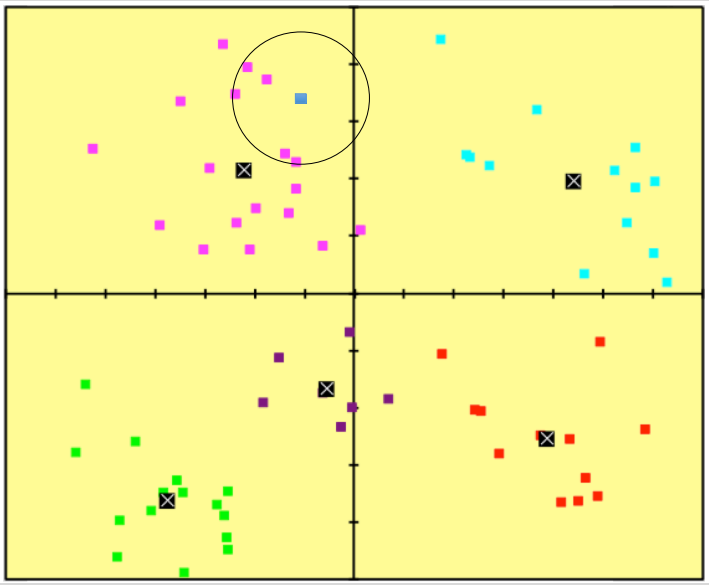
Going from clustering into classification



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

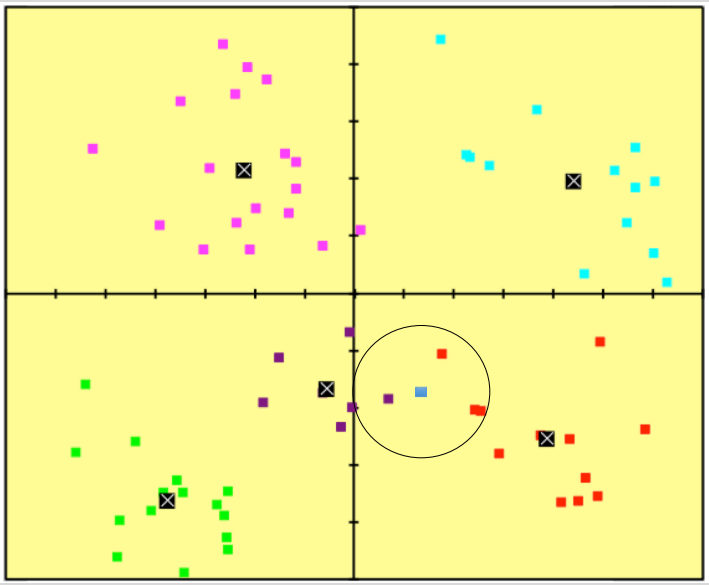
8

Going from clustering into classification



9

Going from clustering into classification



10

Data Mining

Semi-supervised classification

Nearest Neighbors

24/11/2021

NOVA-IMS

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

AGENDA

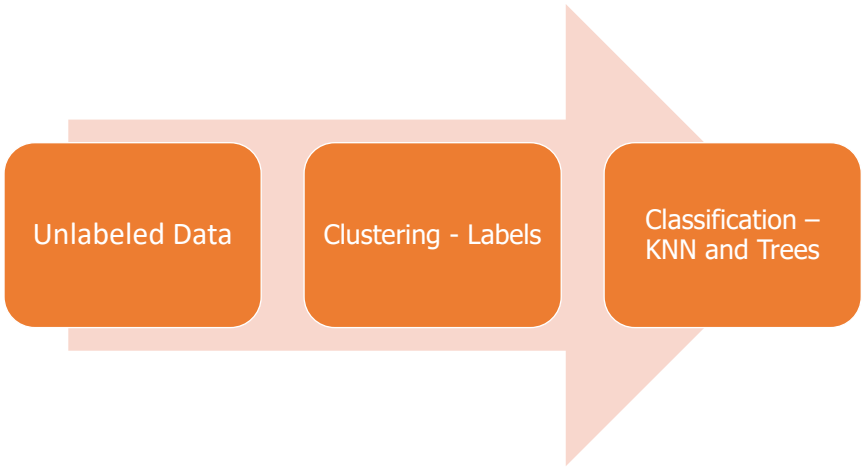
- Cluster analysis
  - Semi-supervised classification
    - Nearest Neighbors

2

Semi-supervised classification

3

Going from clustering into classification



4

Going from clustering into classification

5

k-nearest neighbors (k-NN)

6

k-nearest neighbors

- Instance based classification:
  - Simplest form of learning;
  - Training instances are searched for instances that most closely resembles new instance;
  - The instances themselves represent the knowledge;
  - Also called instance-based learning
  - Similarity function defines what's "learned"

k-nearest neighbors

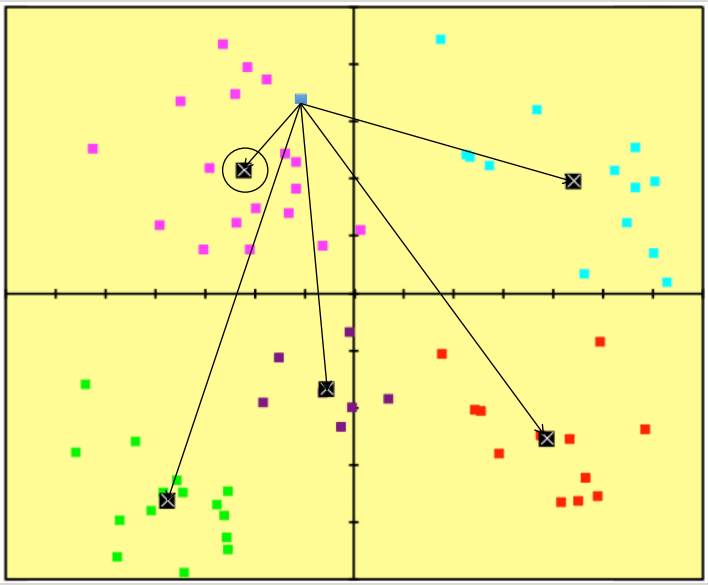
- Requires three things:
  1. The set of stored records (with labels)
  2. A distance metric to compute distance between records (can use Euclidean distance)
  3. The value of k, the number of nearest neighbors to retrieve

k-nearest neighbors

- To classify an unknown record:
  1. Compute distance to other training records
  2. Identify k nearest neighbors
  3. Use class labels of nearest neighbors to determine the class label of unknown record (e.g., by taking majority vote)

9

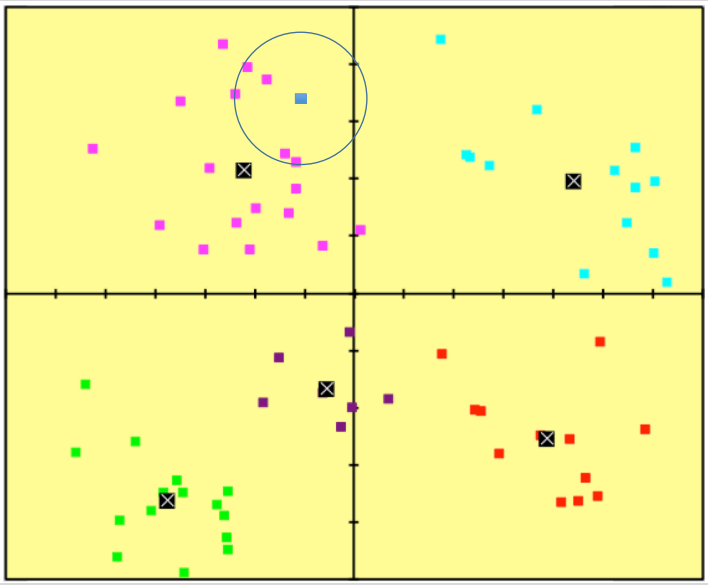
Going from clustering into classification



10

NOVA IMS Information Management School

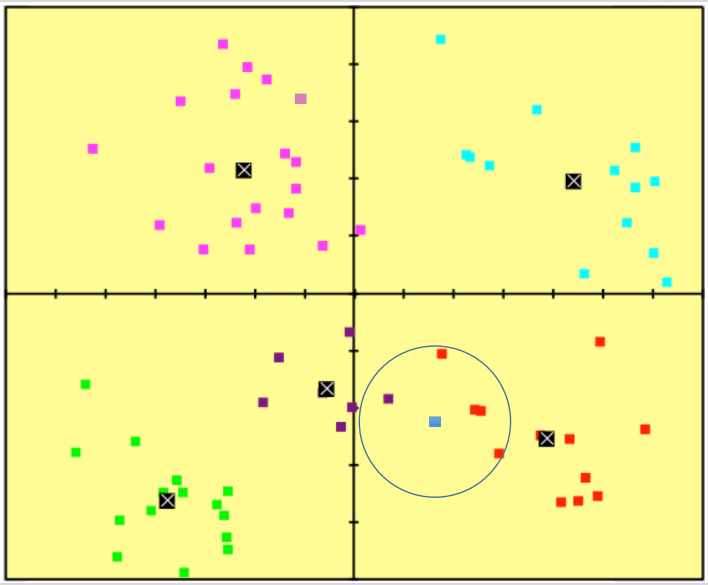
Going from clustering into classification



11

NOVA IMS Information Management School

Going from clustering into classification



12

k-nearest neighbors

- Compute distance between two points:
  - Euclidean distance

$$d(p,q) = \sqrt{\sum_i (p_i - q_i)^2}$$

  - Determine the class from nearest neighbor list
    - Take the majority vote of class labels among the k-nearest neighbors
    - Weigh the vote acording to distance
      - weight factor, w = 1/d2

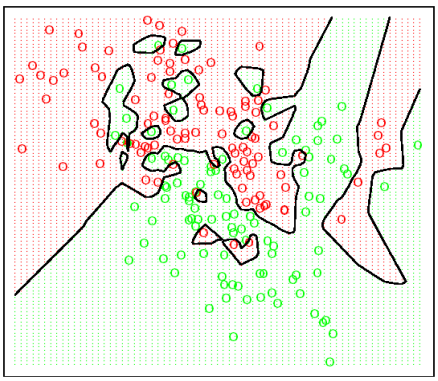
k-nearest neighbors

- k-nn frontiers (and the number *k*):
  - Large k
    - Smooth frontiers
    - Unable to detect small variations
  - Small k
    - Very sensitive to outliers
    - Crisp frontiers

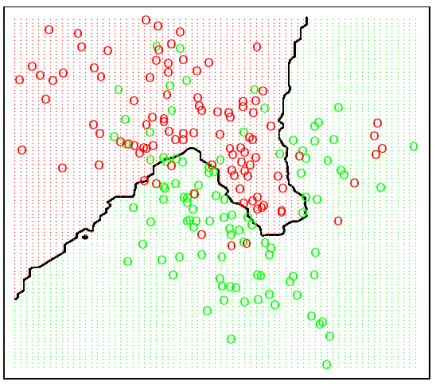
NOVA IMS Information Management School

k-nearest neighbors

1-nn



15-nn



Data Mining

Semi-supervised classification

Classification trees

24/11/2021

NOVA-IMS

Fernando Lucas Bação

bacao@isegi.unl.pt

http://www.isegi.unl.pt/fbacao

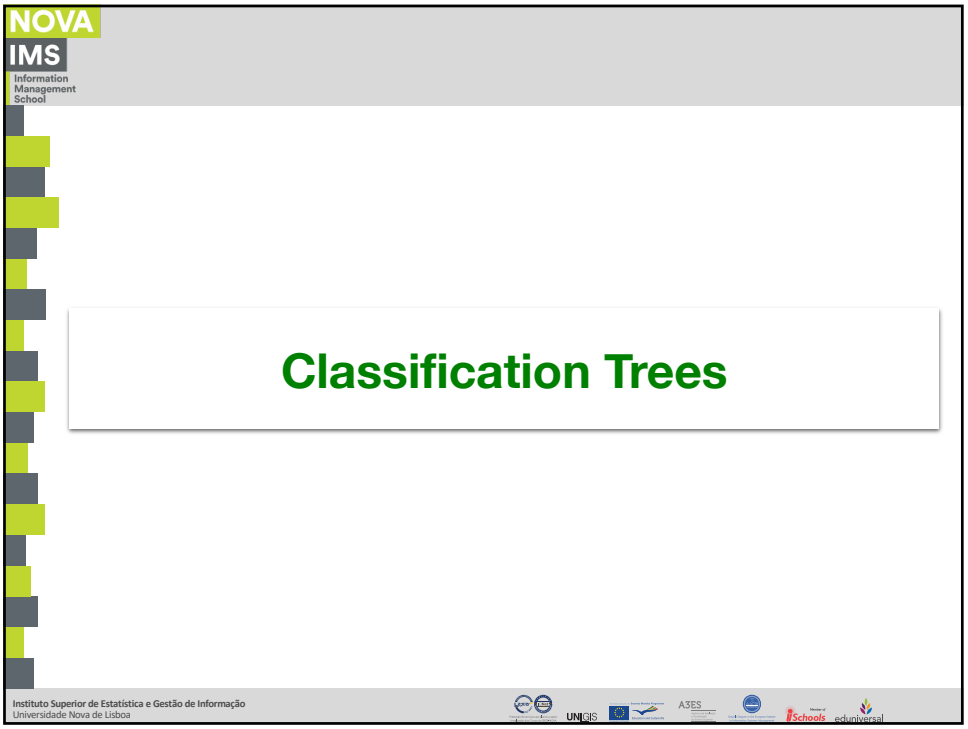
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

1

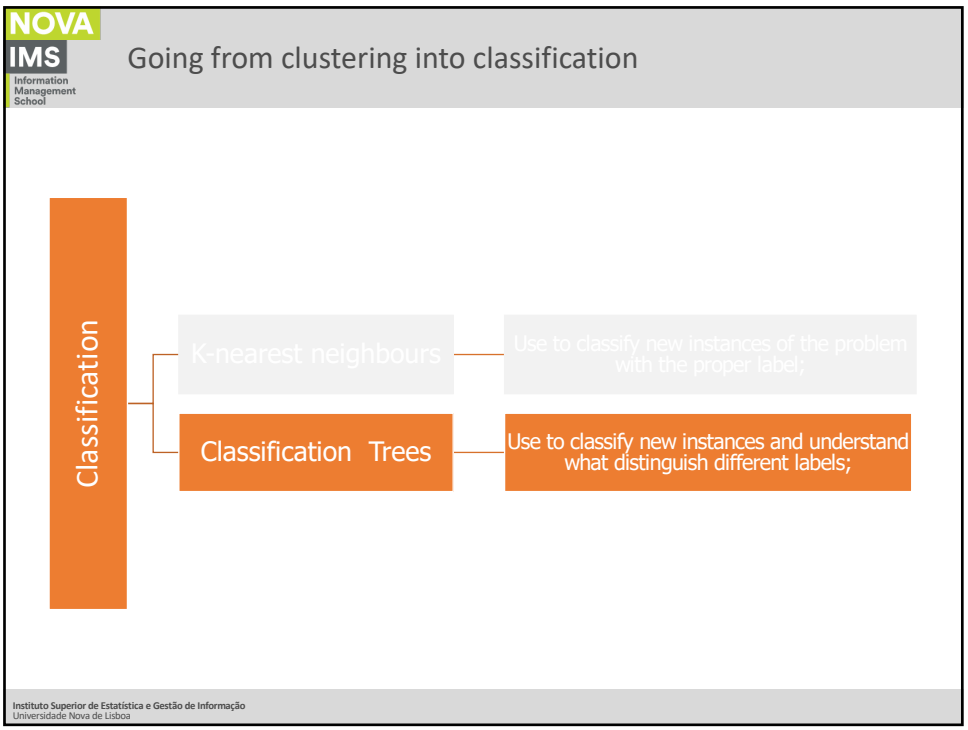
AGENDA

- Cluster analysis
  - Semi-supervised classification
    - Classification Trees

2


NOVA
IMS
Information Management School
Classification Trees
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

3


NOVA
IMS
Information Management School
Going from clustering into classification
Classification
K-nearest neighbours
Use to classify new instances of the problem with the proper label;
Classification Trees
Use to classify new instances and understand what distinguish different labels;
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

4

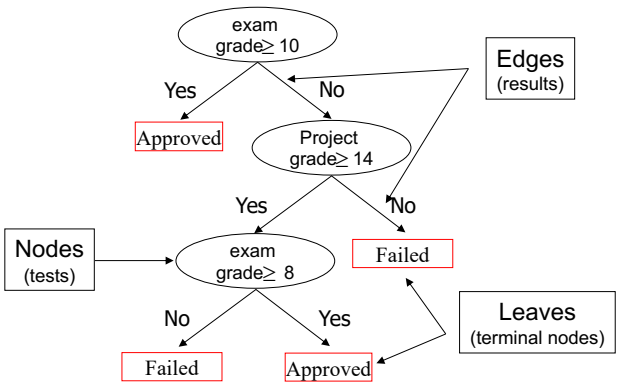
Classification Trees

- **Classification Trees:**
  - Classification tree are typically considered to be classification and regression tools
  - One of its most important advantages relates with the simplicity of the interpretation of its results
  - Thus, the end result of a classification tree can easily be expressed in English or SQL.

5

Classification Trees

- A classification tree is a decisional algorithm
- It can be seen as a way of storing knowledge
- The objective is to discriminate between Class
- Obtain leaves as pure as possible
- If possible each leave should represent only individuals from a specific class



6

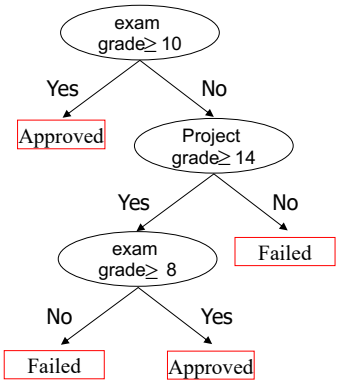
Classification Trees

$$aprovado \Leftrightarrow (exame \geq 10)$$

$$aprovado \Leftrightarrow (exame < 10) \wedge (projecto \geq 14) \wedge (exame \geq 8)$$

$$reprovado \Leftrightarrow (exame < 10) \wedge (projecto < 14)$$

$$reprovado \Leftrightarrow (exame < 10) \wedge (projecto > 14) \wedge (exame < 8)$$



7

Classification Trees

- **Classification Trees (strenghts):**
  - Interpretation
    - We can easily understand the reasons behind a specific classification decision
  - May use different types of data
    - Interval, ordinal, nominal, etc.
  - Insensitive to scale factors
    - Variables measured in different scales may be used without any type of normalization

8

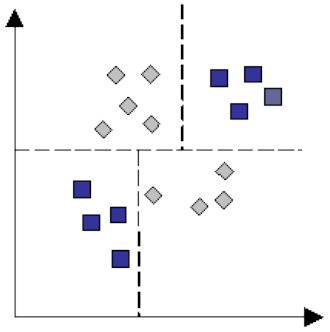
Classification Trees

- **Classification Trees (strenghts):**
  - Automatically defines the most relevant variables
    - These are the variables used at the top of the tree
  - Can be adapted to a regression
    - Each leave becomes a linear model

9

Classification Trees

- **Classification Trees (weaknesses):**
  - Boundaries are linear and perpendicular to the variables axys
  - Sensitive to small perturbations in the data

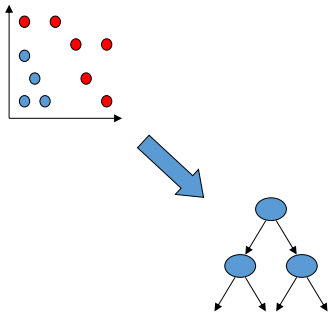


From Gahegan and West
http://divcom.otago.ac.nz/SIRC/GeoComp/GeoComp98/61/gc_61.htm

10

Classification Trees

- **Classification Trees:**
  - Build (induce) a tree from data
  - Problems:
    - What to do?
      - Which variable to use?
      - What partition to use?
      - Which node to split?
      - How many edges per node?
      - When to stop?



11

Classification Trees

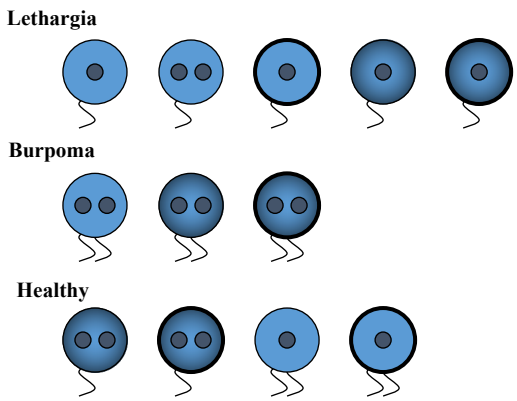
- **Classification Trees:**
  - ID3, C4.5 e C5 [Quinlan 86,93]
    - Iterative Dichotomizer 3
  - CART
    - Classification and regression trees [Breiman 84]
  - CHAID [Hartigan 75]
    - Used in SPSS and SAS…
  - Muitas (mesmo muitas) outras variantes...
    - In SAS you can choose different parameters to build your tree.

12

Worked-Example
General Idea

Langley, P: 1996, Elements of Machine Learning, Morgan and Kaufmann Publishers.

13



14

NOVA IMS Information Management School

Classification Trees

Table

# Nucleus	# Tails	Color	Membrane	*Class*
1	1	Light	Thin	*Lethargia*
2	1	Light	Thin	*Lethargia*
1	1	Light	Thick	*Lethargia*
1	1	Dark	Thin	*Lethargia*
1	1	Dark	Thick	*Lethargia*
2	2	Light	Thin	*Burpoma*
2	2	Dark	Thin	*Burpoma*
2	2	Dark	Thick	*Burpoma*
2	1	Dark	Thin	*Healthy*
2	1	Dark	Thick	*Healthy*
1	2	Light	Thin	*Healthy*
1	2	Light	Thick	*Healthy*

15

NOVA IMS Information Management School

Classification Trees

- **Classification Trees:**
  - Measure to discriminate the attribute

$$f(A) = \frac{1}{n}\sum_{i=1}^{|A|} C_i$$

  - n is the total number of examples and C_i the number of examples correctly classified based on the most frequent class.
  - This is a measure of “dominance” or “purity”

16

NOVA IMS Information Management School

Classification Trees

Table

# Nucleus	1	2
Lethargia	4	1
Burpoma	0	3
Healthy	2	2

Discrimination:

(4 + 3) / 12 = 0.58

# Nucleus	# Tails	Color	Membrane	Class
1	2	Light	Thin	Lethargia
2	1	Light	Thin	Lethargia
1	2	Light	Thick	Lethargia
1	1	Dark	Thin	Lethargia
1	1	Dark	Thick	Lethargia
2	2	Light	Thin	Burpoma
2	2	Dark	Thin	Burpoma
2	2	Dark	Thick	Burpoma
2	1	Dark	Thin	Healthy
2	1	Dark	Thick	Healthy
1	2	Light	Thin	Healthy
1	2	Light	Thick	Healthy

17

NOVA IMS Information Management School

Classification Trees

Table

# Tails	1	2
Lethargia	5	0
Burpoma	0	3
Healthy	2	2

Discrimination:

(5 + 3) / 12 = 0.67

# Nucleus	# Tails	Color	Membrane	Class
1	1	Light	Thin	Lethargia
2	1	Light	Thin	Lethargia
1	1	Light	Thick	Lethargia
1	1	Dark	Thin	Lethargia
1	1	Dark	Thick	Lethargia
2	2	Light	Thin	Burpoma
2	2	Dark	Thin	Burpoma
2	2	Dark	Thick	Burpoma
2	1	Dark	Thin	Healthy
2	1	Dark	Thick	Healthy
1	2	Light	Thin	Healthy
1	2	Light	Thick	Healthy

18

NOVA IMS Information Management School

Classification Trees

Table

Color	Light	Dark
Lethargia	3	2
Burpoma	1	2
Healthy	2	2

Discrimination:

(3 + 2) / 12 = 0.41

# Nucleus	# Tails	Color	Membrane	Class
1	1	Light	Thin	Lethargia
2	1	Light	Thin	Lethargia
1	1	Light	Thick	Lethargia
1	1	Dark	Thin	Lethargia
1	1	Dark	Thick	Lethargia
2	2	Light	Thin	Burpoma
2	2	Dark	Thin	Burpoma
2	2	Dark	Thick	Burpoma
2	1	Dark	Thin	Healthy
2	1	Dark	Thick	Healthy
1	2	Light	Thin	Healthy
1	2	Light	Thick	Healthy

19

NOVA IMS Information Management School

Classification Trees

Table

Membrane	Thin	Thick
Lethargia	3	2
Burpoma	2	1
Healthy	3	1

Discrimination:

(3 + 2) / 12 = 0.41

# Nucleus	# Tails	Color	Membrane	Class
1	1	Light	Thin	Lethargia
2	1	Light	Thin	Lethargia
1	1	Light	Thick	Lethargia
1	1	Dark	Thin	Lethargia
1	1	Dark	Thick	Lethargia
2	2	Light	Thin	Burpoma
2	2	Dark	Thin	Burpoma
2	2	Dark	Thick	Burpoma
2	1	Dark	Thin	Healthy
2	1	Dark	Thick	Healthy
1	2	Light	Thin	Healthy
1	2	Light	Thick	Healthy

20

NOVA IMS Information Management School

Classification Trees

Choice: # Tails

# Nucleus	1	2
Lethargia	4	1
Burpoma	0	3
Healthy	2	2

0.58

Color	Light	Dark
Lethargia	3	2
Burpoma	1	2
Healthy	2	2

0.41

# Tails	1	2
Lethargia	5	0
Burpoma	0	3
Healthy	2	2

0.67

Membrane	Thin	Thick
Lethargia	3	2
Burpoma	2	1
Healthy	3	1

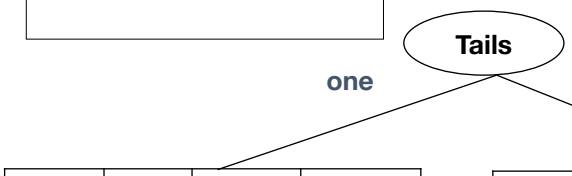
0.41

21

NOVA IMS Information Management School

Classification Trees

Initial Partition



# Nucleus	Color	Membrane	Class
1	Light	Thin	Lethargia
2	Light	Thin	Lethargia
1	Light	Thick	Lethargia
1	Dark	Thin	Lethargia
1	Dark	Thick	Lethargia
2	Dark	Thin	Healthy
2	Dark	Thick	Healthy

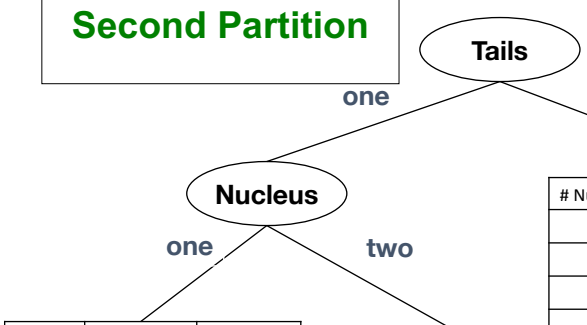
# Nucleus	Color	Membrane	Class
2	Light	Thin	Burpoma
2	Dark	Thin	Burpoma
2	Dark	Thick	Burpoma
1	Light	Thin	Healthy
1	Light	Thick	Healthy

22

NOVA IMS Information Management School

Classification Trees

Second Partition



# Nucleus	Color	Membrane	Class
2	Light	Thin	Burpoma
2	Dark	Thin	Burpoma
2	Dark	Thick	Burpoma
1	Light	Thin	Healthy
1	Light	Thick	Healthy

Color	Membrane	Class
Light	Thin	Lethargia
Light	Thick	Lethargia
Dark	Thin	Lethargia
Dark	Thick	Lethargia

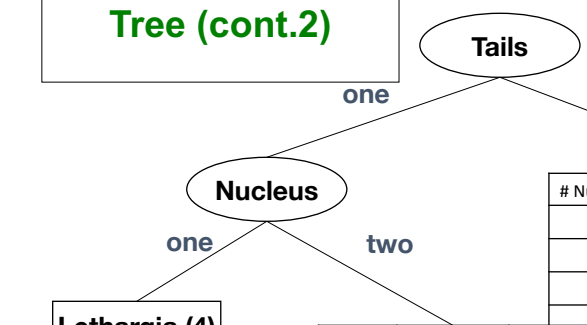
Color	Membrane	Class
Light	Thin	Lethargia
Dark	Thin	Healthy
Dark	Thick	Healthy

23

NOVA IMS Information Management School

Classification Trees

Tree (cont.2)



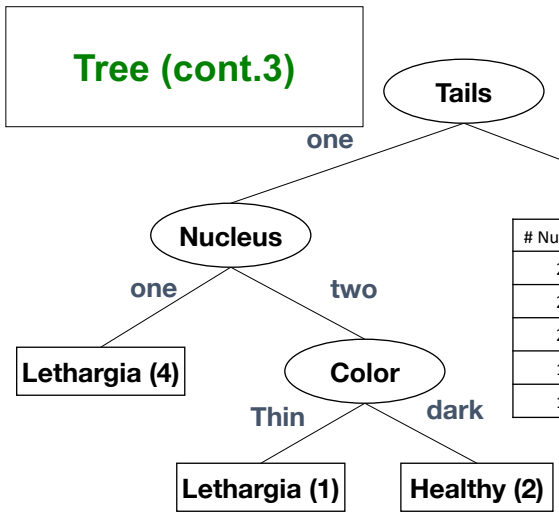
# Nucleus	Color	Membrane	Class
2	Light	Thin	Burpoma
2	Dark	Thin	Burpoma
2	Dark	Thick	Burpoma
1	Light	Thin	Healthy
1	Light	Thick	Healthy

Color	Membrane	Class
Light	Thin	Lethargia
Dark	Thin	Healthy
Dark	Thick	Healthy

24

Classification Trees

Tree (cont.3)

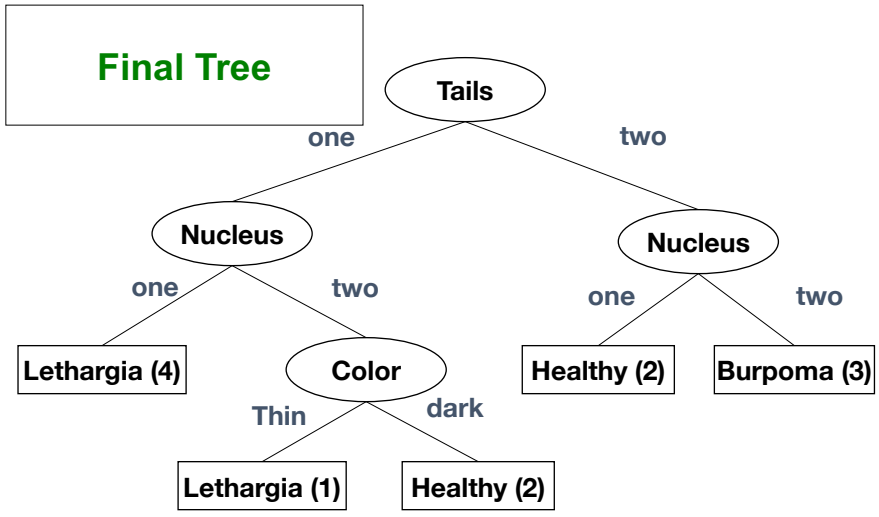


# Nucleus	Color	Membrane	Class
2	Light	Thin	Burpoma
2	Dark	Thin	Burpoma
2	Dark	Thick	Burpoma
1	Light	Thin	Healthy
1	Light	Thick	Healthy

25

Classification Trees

Final Tree



26

The description of the three Classs

$$(tails = 1) \wedge (nucleous = 1)$$
$$\vee$$
$$(tails = 1) \wedge (nucleous = 2) \wedge (color = light) \rightarrow Lethargic$$

$$(tails = 2) \wedge (nucleous = 1)$$
$$\vee$$
$$(tails = 1) \wedge (nucleous = 2) \wedge (color = dark) \rightarrow Healthy$$

$$(tails = 2) \wedge (nucleous = 2) \rightarrow Burpoma$$

- **Classification Trees:**
  - In each level it divides the set into alternative partitions.
    - Using a measure of quality selects the best partition.
  - The process is repeated for each element of the partition.
  - Stops when a given criteria is reached

Classification Trees

- **Classification Trees:**
  - It assumes the existence of a target variable "Class" meaning the examples were previously classified.
  - Each node specifies a unique attribute which is used as test.
  - N – node N
  - ASET – Atribute Set
  - ISET – Instance Set

29

Classification Trees

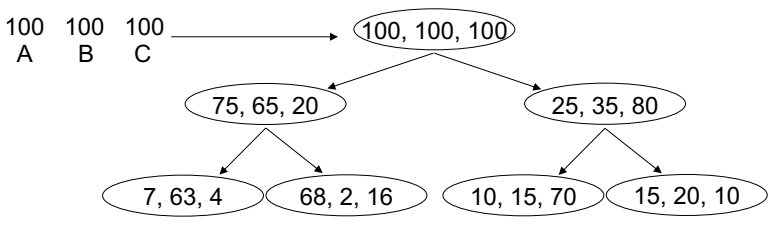
```
If Se the ISET is empty then the terminal node N is an unknown class
    if not
        If all the examples of ISET are of the same class
        then the terminal node N has the name of the class
            if not
                For each attribute A of the set of attribute ASET
                    Evaluate A acColording to its capability to discriminate a class
                    Select the attribute B which has the best discriminate value
                    For each value V of the best attribute B
                        Create a new node C from node N
                        Place the par attribute value (B, V) in C
                        Let JSET be the set of examples of ISET with value V in B
                        Let KSET be the set of attributes of ASET with B removed
                        DDT(C, KSET, JSET)
```


30

Worked-Example
Tree Accuracy

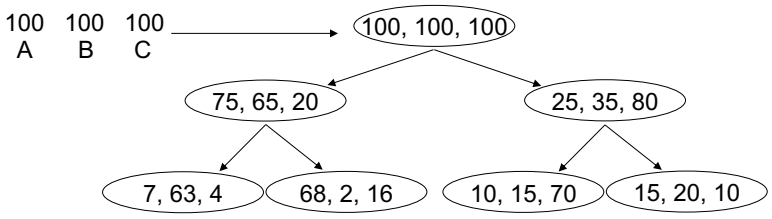
31

Quality of the results
Error rate



32

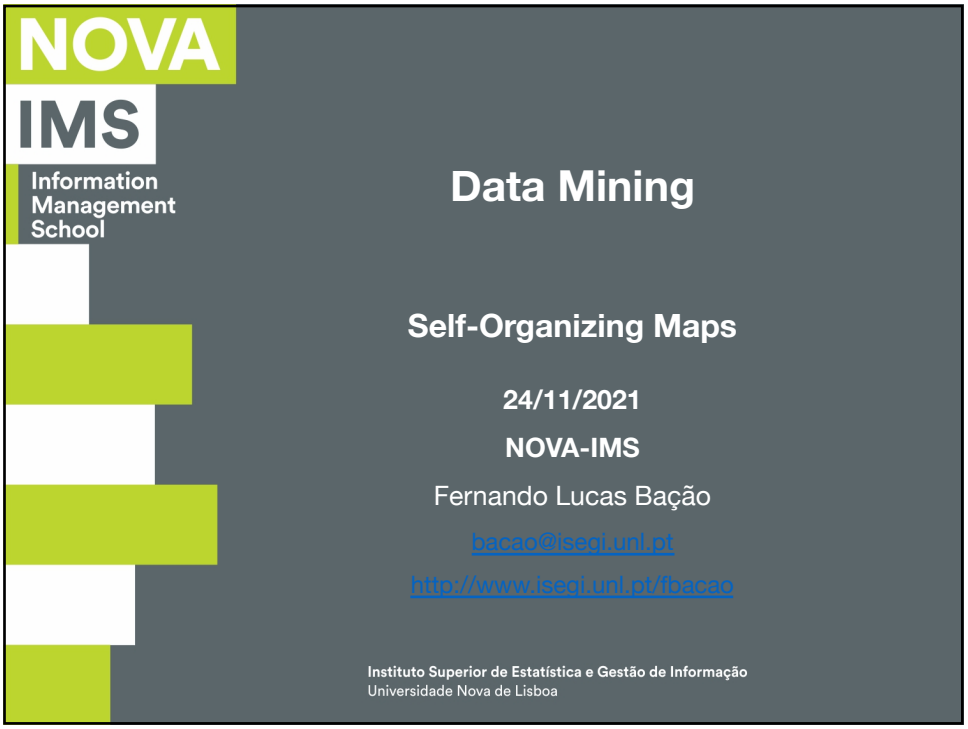
Quality of the results
Error rate



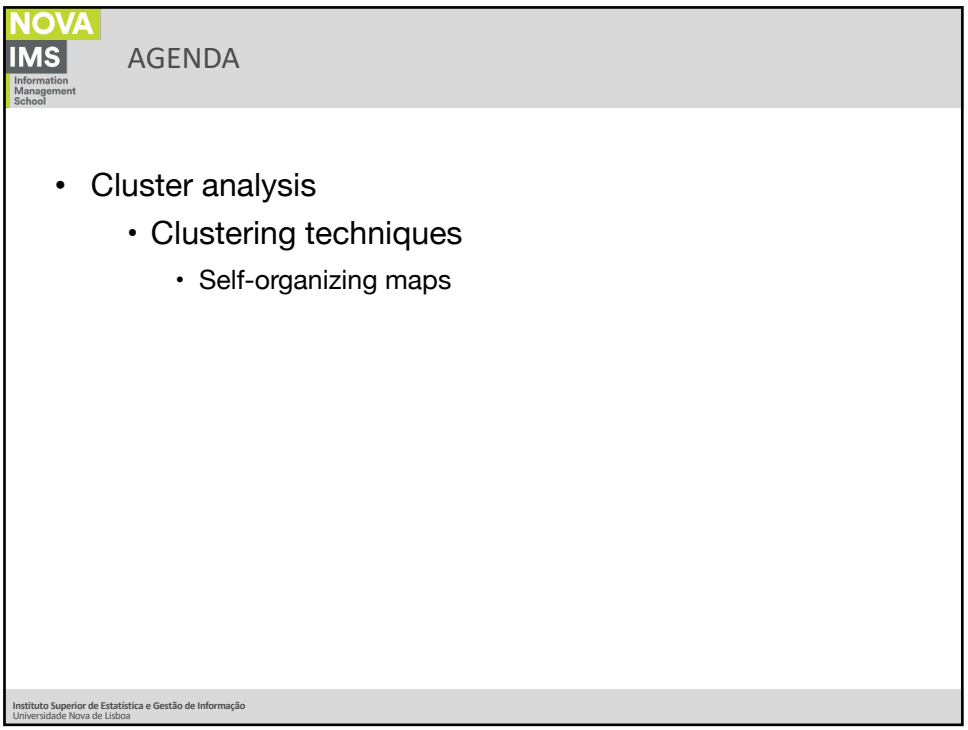
Error Calculation

$$TEA = \frac{(74*14{,}9\% + 86*20{,}9\% + 95*26{,}3\% + 45*55{,}5\%)}{300} = 26{,}3\%$$

33


NOVA
IMS
Information
Management
School
Data Mining
Self-Organizing Maps
24/11/2021
NOVA-IMS
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

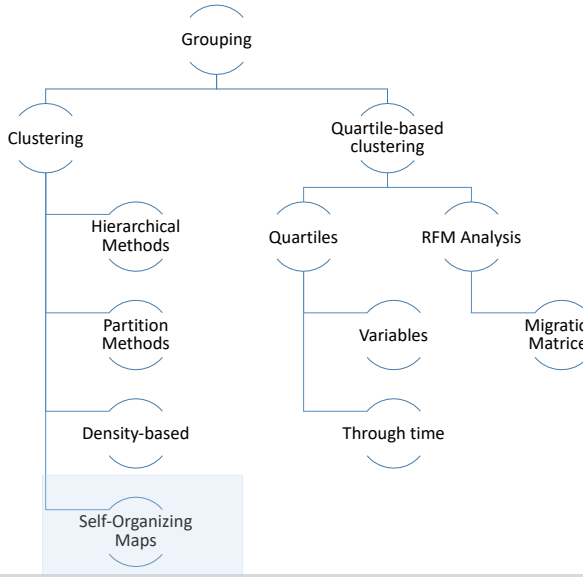

NOVA
IMS
Information
Management
School
AGENDA
• Cluster analysis
• Clustering techniques
• Self-organizing maps
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

2

Clustering

3

Clustering



4

Self-Organizing Maps

5

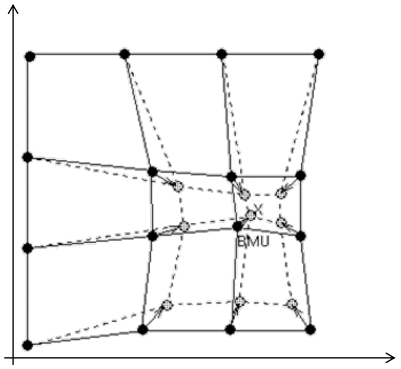
Self-Organizing Maps

- Unsupervised neural networks;
- Closely related to clustering;
- The inputs are connected to a two-dimensional (it may have several dimensions) matrix of units (neurons);
- Each unit is connected to its neighbors.
- What is its use?
  - Multidimensional data visualization;
  - Cluster detection;
  - Market segmentation;
  - Outlier detection;
  - Solve TSP, robot control, alarm detection, etc., etc., etc.

6

Self-Organizing Maps

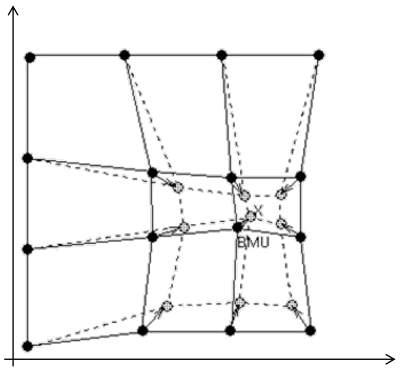
- Each neuron is a **vector in the input space**, just as the data patterns;
- During training, neurons are **pulled** to the positions of the input data, **dragging** with them their neighbors in the output space;
- The map can be seen as a **rubber sheet**, stretched and twisted, so that it passes in (or at least near) the data patterns.



7

Self-Organizing Maps

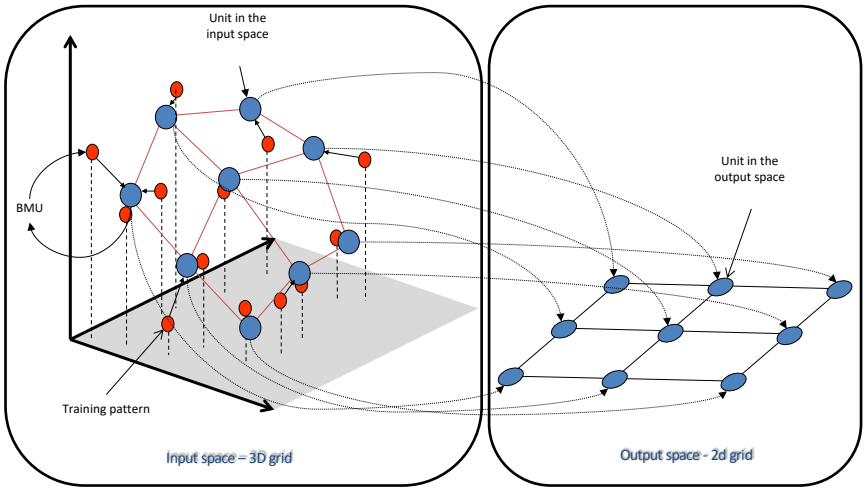
- Input patterns are compared with all neurons and the **closest is considered to be the winning neuron.**
- We consider that the input pattern is **mapped to the winning neuron**.
- The **winner is updated** (so that it resembles even more the data pattern that it represents), and its neighbors are also updated a little.
- There is always a slight difference between the data and the neurons that represent them. That difference is the **quantization error**.



8

NOVA IMS Information Management School

Self-Organizing Maps



9

NOVA IMS Information Management School

Self-Organizing Maps

- **Comparisons with biology**
  - Biological systems have to use some kind of self-organization and adaptation.
  - There is evidence of:
    - Layered structure in the brain;
    - Those layers seem to spatially organize the information;
    - Similar “Concepts” are mapped to adjacent areas;
    - Experimental work with animals points to an organization similar to SOM in the visual cortex.

10

The SOM Algorithm

11

- **SOM Algorithm:**
  - How does SOM processes data?

Video

12

NOVA IMS Information Management School

Self-Organizing Maps

- **SOM Algorithm:**
  - What happens in the Input space?
  - If we only have two variables, optimization will look like this

SOM 1-dimensional

SOM 2-dimensional

13

NOVA IMS Information Management School

Self-Organizing Maps

- **SOM Algorithm:**
  - What happens in the output space

Color Demo

  - Suppose we want to group cells according to their rgb code (red, green and blue)
  - Each individual (color) represents a particular combination of rgb intensities
  - In this demo, we can see how the output space is transformed as individuals are presented to the network

14

NOVA IMS Information Management School

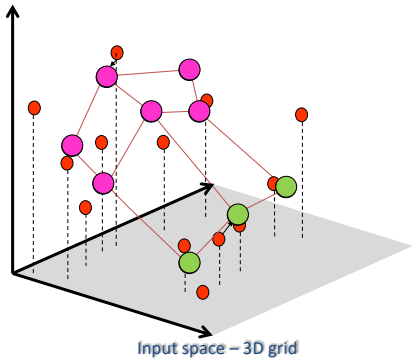
Self-Organizing Maps

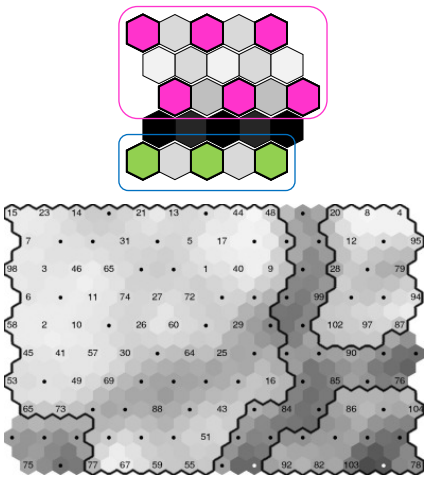
- **SOM Algorithm:**
  - Key outputs of the SOM:
    - U-Matrices;
    - Component plans;
    - Hit Plots.

15

NOVA IMS Information Management School

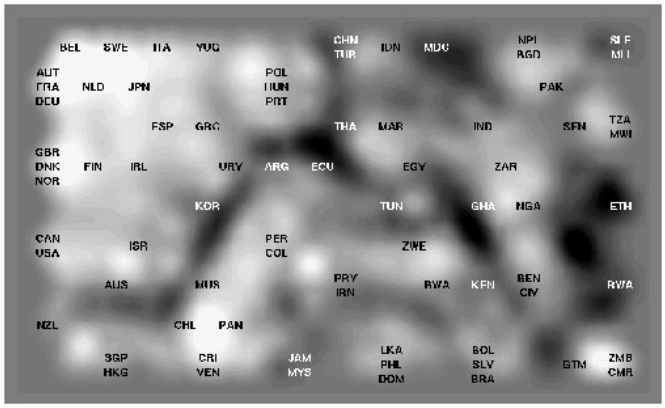
Self-Organizing Maps – U-Matrix





16

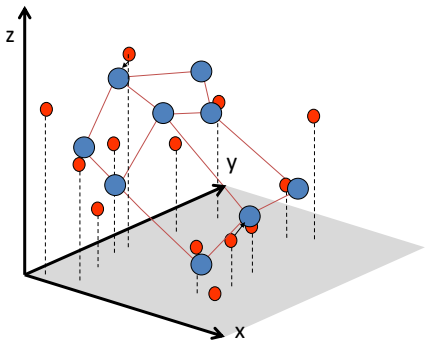
Self-Organizing Maps – U-Matrix



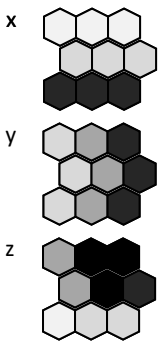
Kaski, S. and Kohonen, T. 1996. Exploratory Data Analysis by the Self-Organizing Map: Structures of Welfare and Poverty in the World. In: Apostolos, P. N., Refenes, Y. A., Moody, J. and Weigend, A. (eds.) Neural Networks in Financial Engineering. Singapore: World Scientific, 498-507.

17

Self-Organizing Maps – Component Planes

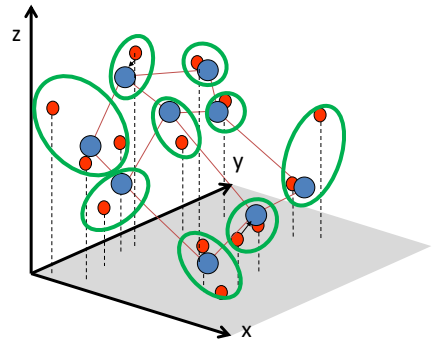


Input space – 3D grid

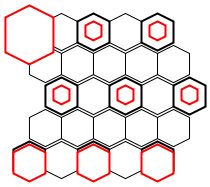


18

Self-Organizing Maps – Hit Maps

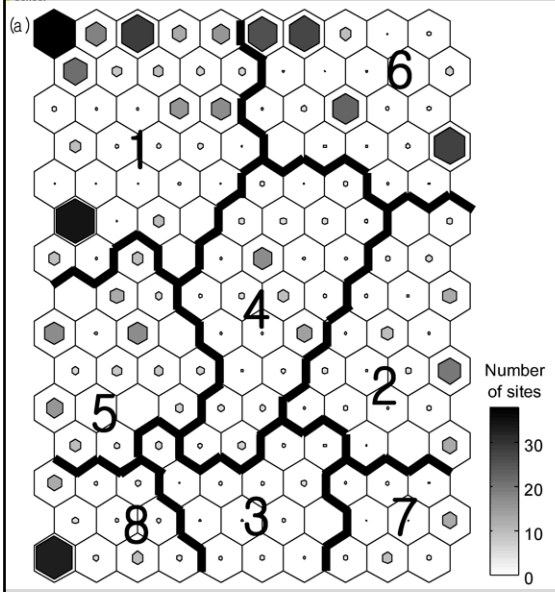


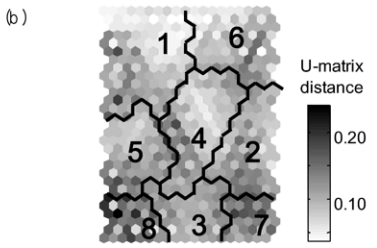
Input space – 3D grid

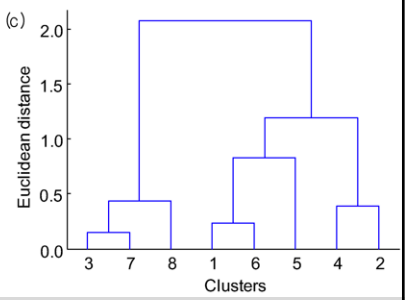


19

Self-Organizing Maps – Interpreting the U-Matrix







20

NOVA IMS Information Management School

Self-Organizing Maps

- **SOM Algorithm:**

Step 0: Randomly initialize the weights w_{ij}
Set the neighborhood topological parameters
Set the learning rate

Step 1: While stop condition false, do steps 2-7

Step 2: For each input vector x, do steps 3-5

Step 3: For each *j*, execute:

$$D(j) = \sum (w_{ij} - x_i)^2$$

Step 4: Find the unit that minimizes D(*j*)

Step 5: For every *j* unit within the predefined and for all the *i*:

$$w_{iJ}\ (\text{new}) = w_{iJ}\ (\text{old}) + \alpha\ [x_i - w_{iJ}\ (\text{old})]$$

Step 6: Update the learning rate

Step 7: Update (reduce) the radius of the topological neighborhood

NOVA IMS Information Management School

Self-Organizing Maps

- **SOM Algorithm:**

Vectors to classify:

(1, 1, 0, 0); (0, 0, 0, 1); (1, 0, 0, 0); (0, 0, 1, 1)

Maximum number of clusters to form:

m = 2

Learning rate:

$\alpha(0) = .6$, $\alpha(t+1) = .5\ \alpha(t)$

NOVA IMS Information Management School

Self-Organizing Maps

- **SOM Algorithm:**

Step 0: Initial matrix of weights

(0.2, 0.6, 0.5, 0.9);

(0.8, 0.4, 0.7, 0.3);

Initial radius: R = 0

Initial learning rate:

$$\alpha(0) = .6$$

23

NOVA IMS Information Management School

Self-Organizing Maps

- **SOM Algorithm:**

Step 1: Initialize training

Step 2: first vector (1, 1, 0, 0);

Step 3:

$D(1) = (.2-1)^2 + (.6-1)^2 + (.5-0)^2 + (.9-0)^2$

$= 1.86;$

$D(2) = (.8-1)^2 + (.4-1)^2 + (.7-0)^2 + (.3-0)^2$

$= 0.98$

Vectors of weights:

(0.2, 0.6, 0.5, 0.9)

(0.8, 0.4, 0.7, 0.3)

24

- **SOM Algorithm:**

Step 4: The input vector is closest to unit 2, therefore

j = 2 = (0.8, 0.4, 0.7, 0.3)

Step 5: The weights of the winning unit are adjusted

$$w_{i2}\text{ (new)} = w_{i2}\text{ (old)} + .6\,[x_i - w_{i2}\text{ (old)}]$$

$$= .4\,w_{i2}\text{ (old)} + .6x_i$$

NOVA IMS Information Management School
Self-Organizing Maps

- **SOM Algorithm:**

Step 4: The input vector is closest to unit 2, therefore

j = 2 = (0.8, 0.4, 0.7, 0.3)

Step 5: The weights of the winning unit are adjusted

$$w_{i2}\text{ (new)} = w_{i2}\text{ (old)} + .6\,[x_i - w_{i2}\text{ (old)}]$$

$$= .4\,w_{i2}\text{ (old)} + .6x_i$$

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

NOVA IMS Information Management School

Self-Organizing Maps

- **SOM Algorithm:**

Thus, the weight matrix is adjusted:

.2	**.92**
.6	**.76**
.5	**.28**
.9	**.12**

27

NOVA IMS Information Management School

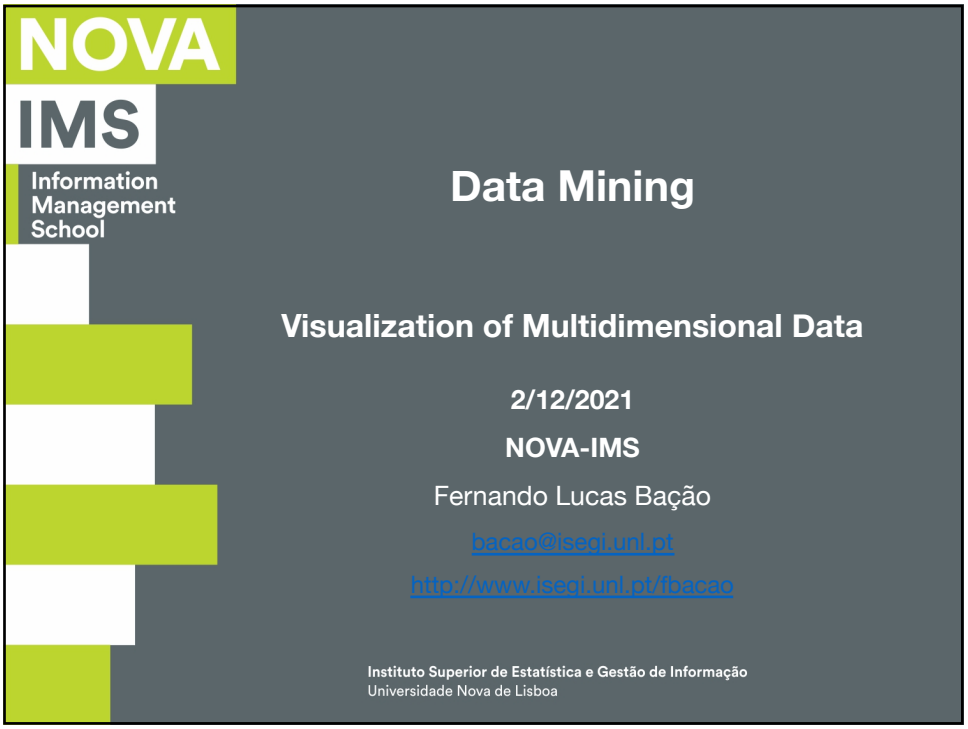
Self-Organizing Maps

- **SOM Algorithm:**

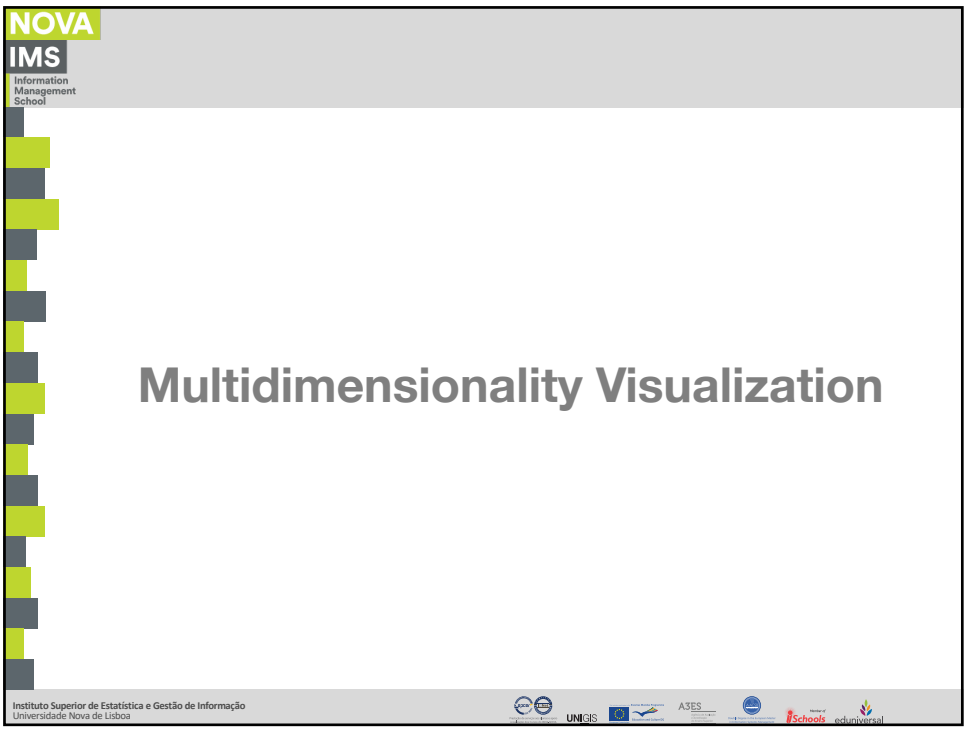
Thus, the weight matrix is adjusted:

.08	**.92**
.24	**.76**
.20	**.28**
.96	**.12**

28

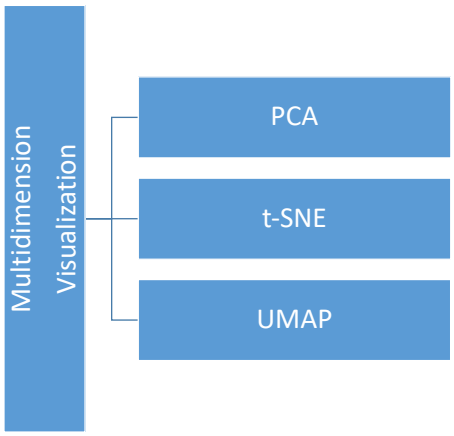

NOVA
IMS
Information
Management
School
Data Mining
Visualization of Multidimensional Data
2/12/2021
NOVA-IMS
Fernando Lucas Bação
bacao@isegi.unl.pt
http://www.isegi.unl.pt/fbacao
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

1


NOVA
IMS
Information
Management
School
Multidimensionality Visualization
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

2

Multidimensionality Visualization



3

Principal Component Analysis

PCA

4

NOVA IMS Information Management School

Principal Component Analysis

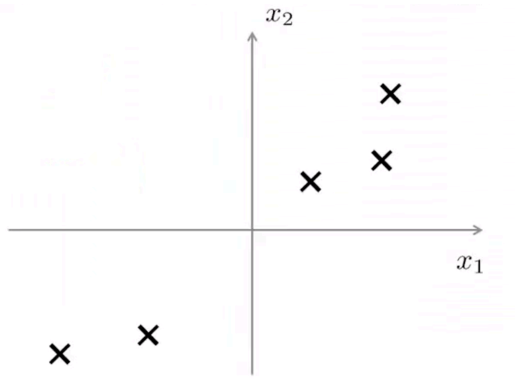
- **Size Reduction of the Input Space:**
  - Principal Component Analysis
    - A procedure that uses an **orthogonal transformation** to convert a set of observations of possibly correlated variables into a set of values of **linearly uncorrelated variables called principal components**.
    - The number 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.

5

NOVA IMS Information Management School

Principal Component Analysis

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

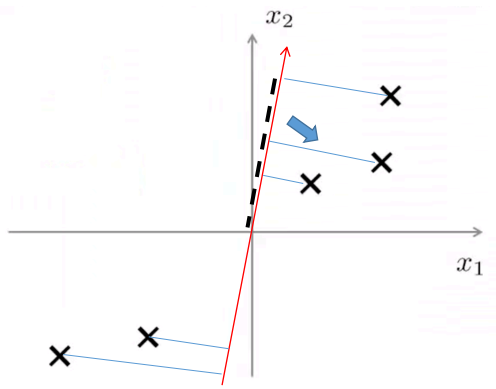


6

NOVA IMS Information Management School

Principal Component Analysis

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

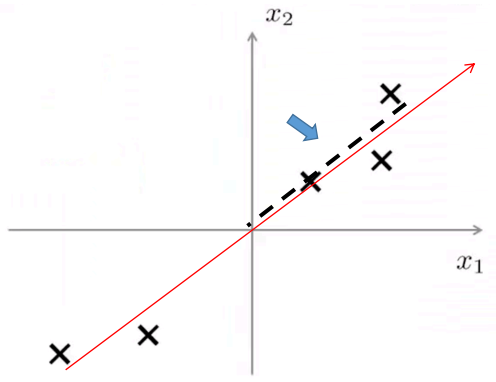


7

NOVA IMS Information Management School

Principal Component Analysis

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

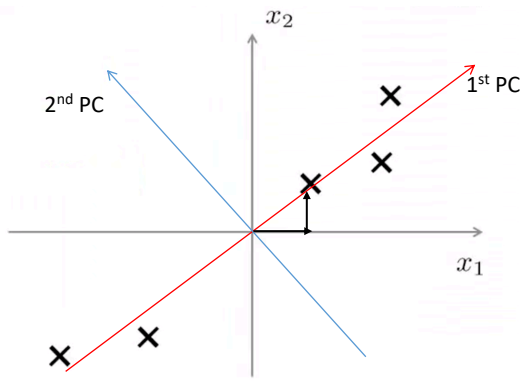


8

NOVA IMS Information Management School

Principal Component Analysis

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

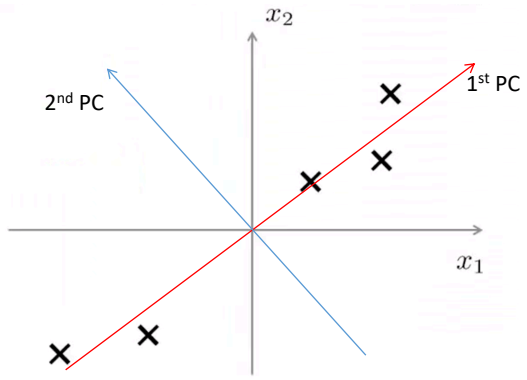


9

NOVA IMS Information Management School

Principal Component Analysis

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

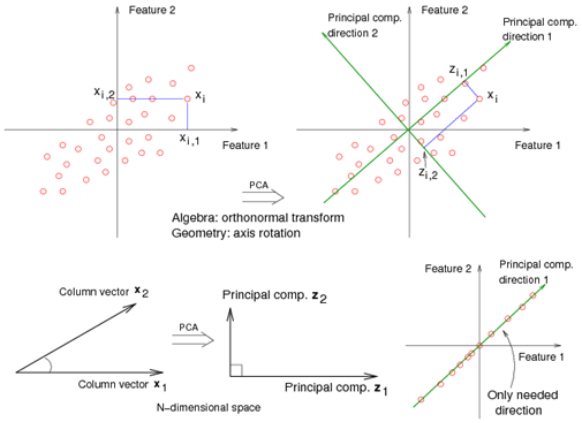


10

NOVA IMS Information Management School

Principal Component Analysis

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

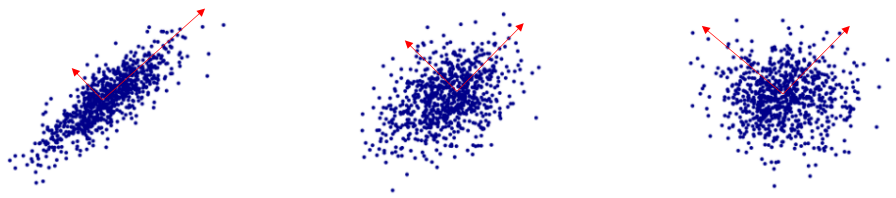


11

NOVA IMS Information Management School

Principal Component Analysis

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



12

t-distributed Stochastic Neighbor Embedding

t-SNE

13

- **t-SNE:**
  - t-distributed stochastic neighbor embedding (t-SNE) is a statistical method for visualizing high-dimensional data by **giving each datapoint a location in a two or three-dimensional map**.
  - It is a **nonlinear dimensionality reduction technique** well-suited for embedding high-dimensional data for visualization in a low-dimensional space of two or three dimensions.
  - Specifically, it **models each high-dimensional object** by a two- or three-dimensional point in such a way that **similar objects are modeled by nearby points** and dissimilar objects are modeled by distant points with high probability.

14

NOVA IMS Information Management School

t-SNE

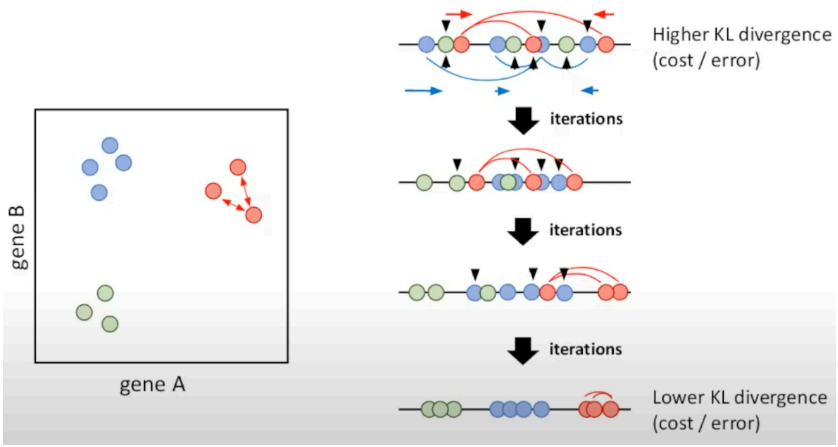
- **t-SNE:**
  - The t-SNE algorithm comprises two stages:
    - First, t-SNE **constructs a probability distribution over pairs of high-dimensional objects** in such a way that similar objects are assigned a higher probability while dissimilar points are assigned a lower probability.
    - Second, t-SNE defines a **similar probability distribution over the points in the low-dimensional map**, and it minimizes the Kullback–Leibler divergence (KL divergence) between the two distributions with respect to the locations of the points in the map.
  - While the original algorithm uses the Euclidean distance between objects as the base of its similarity metric, this can be changed as appropriate.

15

NOVA IMS Information Management School

t-SNE

- **t-SNE:**

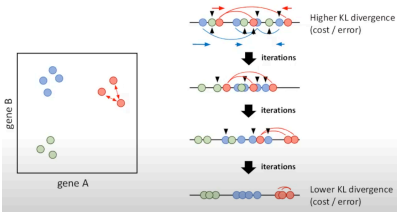


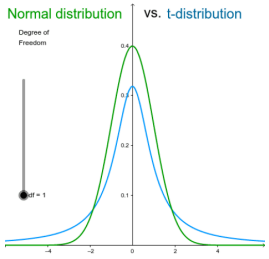
16

NOVA IMS Information Management School

t-SNE

- **t-SNE:**
  - Optimizing the KL divergence is a measure of how one probability distribution is different from a second, reference probability distribution





17

NOVA IMS Information Management School

t-SNE

- **t-SNE:**
  - t-SNE has been used for **visualization in a wide range of applications**, including genomics, computer security research, natural language processing, music analysis, cancer research, bioinformatics, etc
  - While t-SNE plots often seem to display clusters, **the visual clusters can be influenced strongly by the chosen parameterization** and therefore a good understanding of the parameters for t-SNE is necessary.
  - Interactive exploration may thus be necessary to choose parameters and validate results.

18

NOVA IMS Information Management School

t-SNE

- **t-SNE:**
  - Perplexity is the main t-SNE parameter, "perplexity,"
    - Basically defines how to balance attention between local and global aspects of your data.
    - Perplexity is a measure for information that is defined as 2 to the power of the Shannon entropy.
    - In t-SNE, the perplexity may be viewed as a knob that sets the number of effective nearest neighbors.
    - It is comparable with the number of nearest neighbors k that is employed in many manifold learners.
    - The original paper says, "The performance of SNE is fairly robust to changes in the perplexity, and typical values are between 5 and 50."

19

NOVA IMS Information Management School

t-SNE

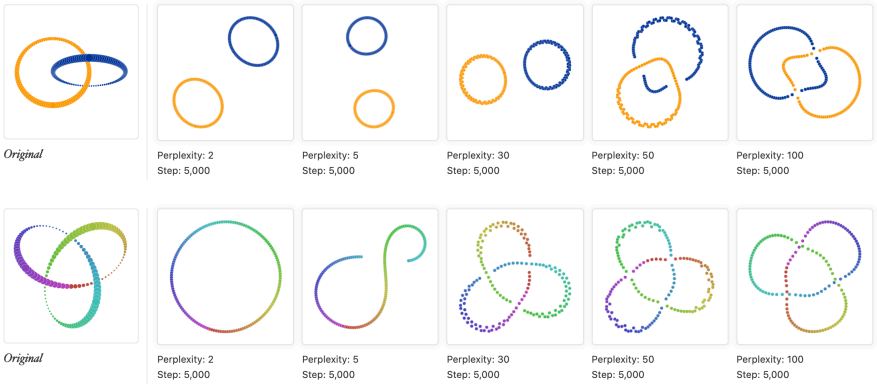
- **t-SNE:**
  - Perplexity is the main t-SNE parameter, "perplexity,"

20

NOVA IMS Information Management School

t-SNE

- **t-SNE:**
  - Perplexity is the main t-SNE parameter, "perplexity,"



21

NOVA IMS Information Management School

Uniform Manifold Approximation and Projection

UMAP

22

NOVA IMS Information Management School

UMAP

- **UMAP:**
  - Uniform Manifold Approximation and Projection (UMAP) is a dimension reduction technique that can be used for visualisation similarly to t-SNE.
  - The algorithm is founded on three assumptions about the data
    - The data is uniformly distributed on Riemannian manifold (a topological space that locally resembles Euclidean space near each point);
    - The manifold is locally connected.
    - From these assumptions it is possible to model the manifold with a fuzzy topological structure. The embedding is found by searching for a low dimensional projection of the data that has the closest possible equivalent fuzzy topological structure.
    - For a more detailed explanation see https://youtu.be/nq6iPZVUxZU

23

NOVA IMS Information Management School

UMAP

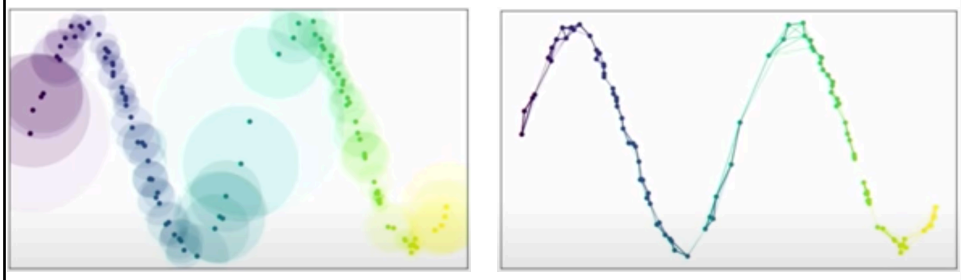
- **UMAP:**
  - Uniform Manifold Approximation and Projection (UMAP) is a dimension reduction technique that can be used for visualisation similarly to t-SNE.
  - Two phases in the UMAP algorithm
    - The first phase consists of constructing a fuzzy topological representation in the original space;
    - The second phase is simply optimizing (through stochastic gradient descent) the low dimensional representation to have as close a fuzzy topological representation as possible as measured by cross entropy.

24

NOVA IMS Information Management School

UMAP

- **UMAP:**
  - The first phase consists of constructing a fuzzy topological representation;



25

NOVA IMS Information Management School

UMAP

- **UMAP:**
  - The second phase is simply optimizing (through stochastic gradient descent) the low dimensional representation to have as close a fuzzy topological representation as possible as measured by cross entropy.

Get the clumps right

$$\sum_{a \in A} \mu(a) \log\left(\frac{\mu(a)}{\nu(a)}\right) + (1-\mu(a)) \log\left(\frac{1-\mu(a)}{1-\nu(a)}\right)$$

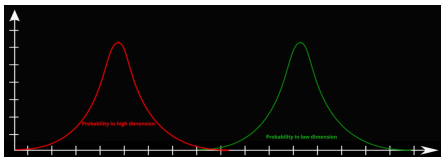
Get the gaps right

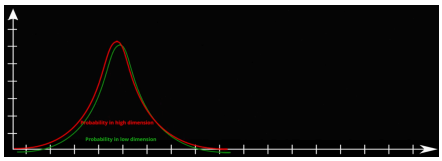
26

NOVA IMS Information Management School

UMAP

- **UMAP:**
  - The second phase is simply optimizing (through stochastic gradient descent) the low dimensional representation to have as close a fuzzy topological representation as possible as measured by cross entropy.





27

NOVA IMS Information Management School

UMAP

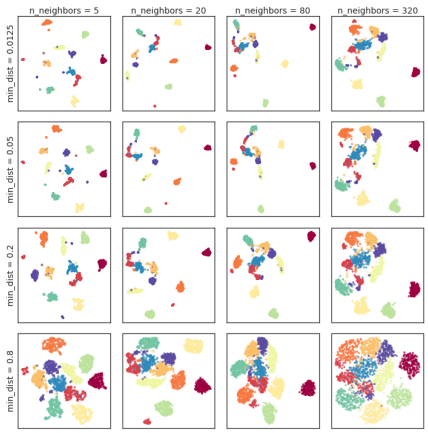
- **UMAP:**
  - Uniform Manifold Approximation and Projection (UMAP) is a dimension reduction technique that can be used for visualisation similarly to t-SNE.
  - Parameters
    - Number of nearest neighbors - controls how UMAP balances local versus global structure in the data;
    - Minimum distance - controls how tightly UMAP is allowed to pack points together. It provides the minimum distance apart that points are allowed to be in the low dimensional representation

28

NOVA IMS Information Management School

UMAP

- **UMAP:**



29

NOVA IMS Information Management School

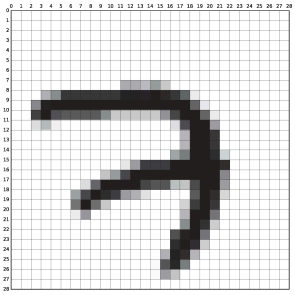
Comparison Between the 3 Methods

30

NOVA IMS Information Management School

Multidimensionality Visualization

- **Mnist Digits:**
  - 28x28 image (784 dimensions)
  - grayscale images of handwritten single digits
  - 60,000 examples, and a test set of 10,000 examples



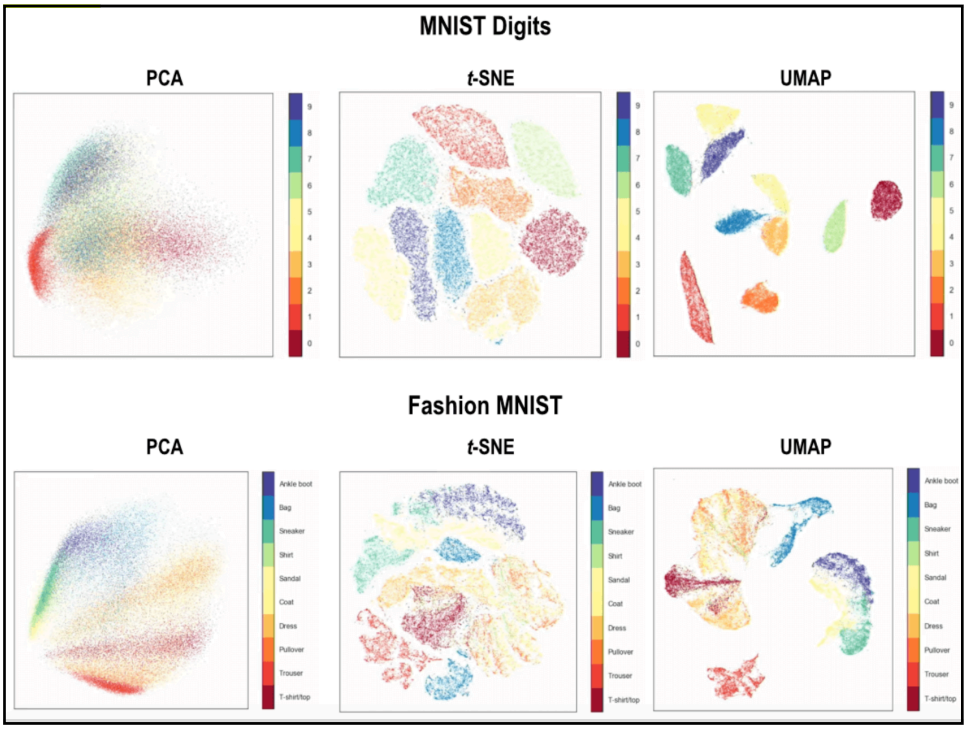
(a) MNIST sample belonging to the digit '7'.

(b) 100 samples from the MNIST training set.

NOVA IMS Information Management School

Multidimensionality Visualization

- **Mnist Fashion:**
  - 28x28 image (784 dimensions)
  - grayscale images of handwritten single digits
  - The dataset has 60,000 images



33

NOVA IMS Information Management School

Multidimensionality Visualization

- **Mnist Digits:**
  - 3D woolly mammoth skeleton
  - https://pair-code.github.io/understanding-umap/

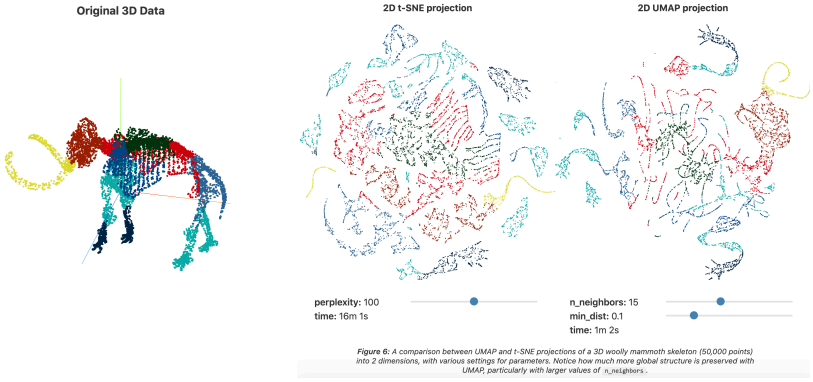


Figure 6: A comparison between UMAP and t-SNE projections of a 3D woolly mammoth skeleton (50,000 points) into 2 dimensions, with various settings for parameters. Notice how much more global structure is preserved with UMAP, particularly with larger values of `n_neighbors`.

34

NOVA IMS Information Management School

UMAP

- **UMAP:**

	t-SNE	UMAP
COIL20	20 seconds	7 seconds
MNIST	22 minutes	98 seconds
Fashion MNIST	15 minutes	78 seconds
GoogleNews	4.5 hours	14 minutes

35

Questions?

36