Data Mining

S3

NOVA-IMS 2025/2026

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

General aspects of the problem definition

2

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**."

> The time it takes for a rock to reach the ground is given by the equation:
>
> $t = \sqrt{(2h/g)}$
>
> 'g' is the acceleration due to gravity (approximately 9.8 m/s² on Earth)

- Invariably, **simple models** and a **lot of data** trump more elaborate models based on less data.

3

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"**

4

Problem definition

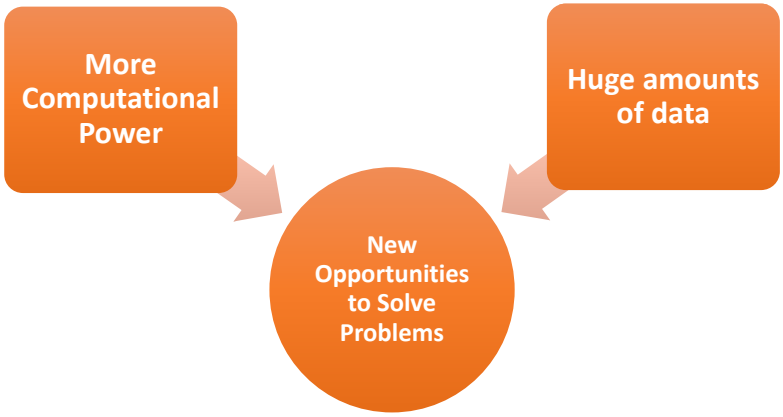
- Suppose you've constructed the **best set of features** you can, but the **classifiers you're getting are still not accurate enough**.
- What can you do now? There are **two main choices**:
  - design a **better learning algorithm**,
  - or **gather more data** (more examples, and possibly more raw 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**.)

5

Statistics vs Data Science

6

Statistics vs Data Science



7

What new about Data Mining?

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

- Traditional 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…..

8

Statistics vs Data Science

Dimension	Primary data	Secondary data
Definition	Data you collect yourself for a specific, current purpose.	Data collected by others for a different (often past) purpose.
Typical sources	Surveys, experiments, interviews, field measurements, sensors.	Government statistics, research papers, data portals, company databases, syndicated datasets, web-scraped corpora.
Control over design	Full control (sampling, instruments, definitions, timing).	Little/no control; must accept others' design choices.
Fit to your question	High: tailored to your problem and target population.	Varies: often indirect or requires redefinition/derivations.
Cost	Usually higher (money, time, staff, tooling).	Usually lower or free; licensing may apply.
Time to obtain	Longer (planning → collection → cleaning).	Faster (download/access + cleaning/understanding).
Timeliness/recency	Up-to-date by design.	May be outdated; release lags common.
Granularity	Exactly what you need (variables, frequency, detail).	Fixed by source; may lack key variables or be too aggregated.

9

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?**
    - Statistical significance answers "is there any effect?"
    - Substantive significance answers "is it big enough to matter for our goal?"

10

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.

11

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.

Fit all at once (batch): The model uses the whole dataset at each update step (or solves a closed-form) so each parameter update reflects global information.

Adjust one point at a time (online/incremental): The model updates immediately after each example, so learning is local and order-dependent, great for streams and drift.

12

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

13

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

- Nonstationarity, Selection Bias, and Dependent Observations
  - **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.

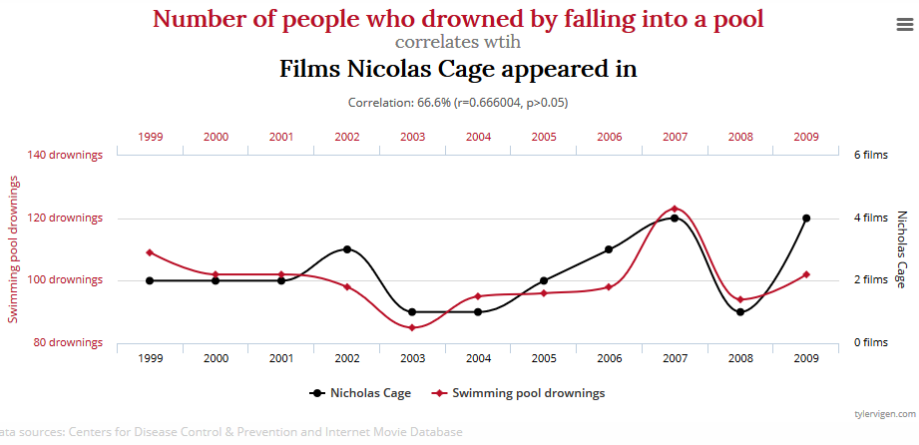
14

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.

15

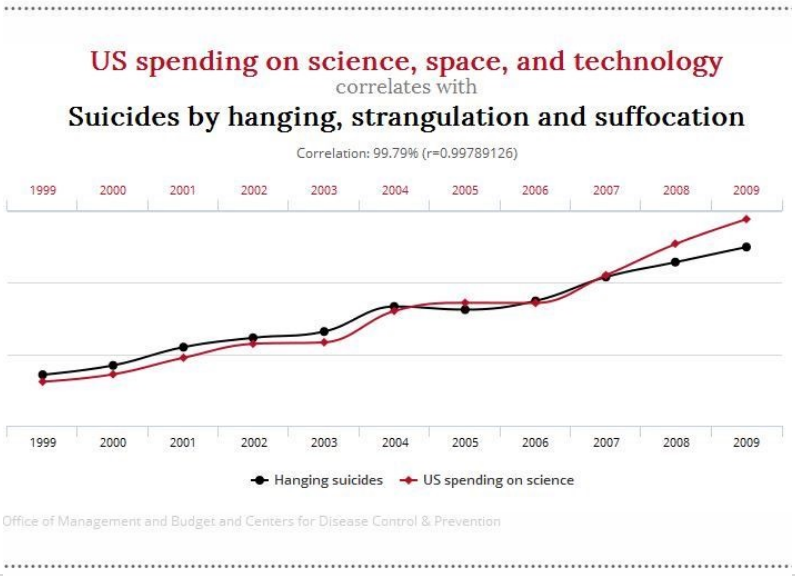
What new about Data Mining?
(Data Mining: Statistics and More? David J. Hand)



16

What new about Data Mining?

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



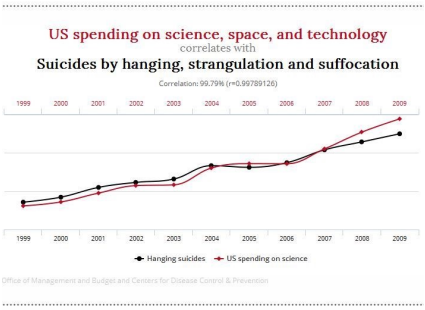
17

Spurious Correlations and Confounding Variables

18

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.



19

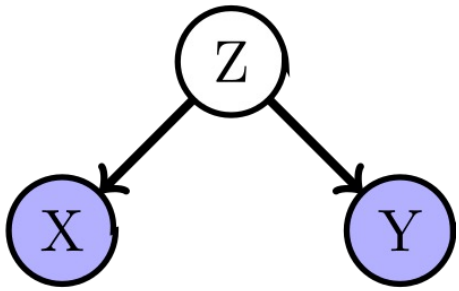
Causality and correlation

- **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**.

20

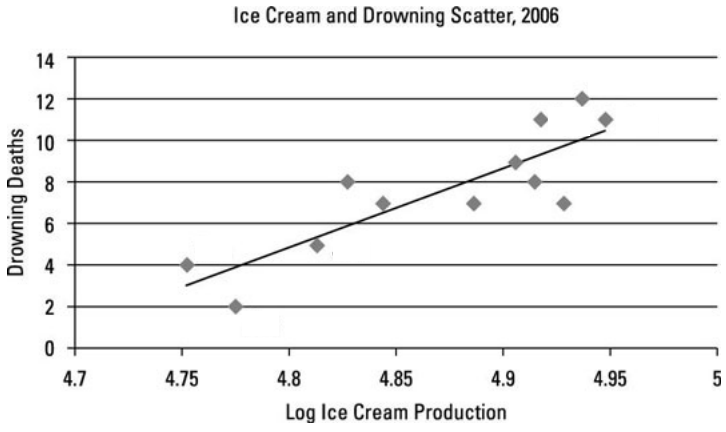
Problem definition

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



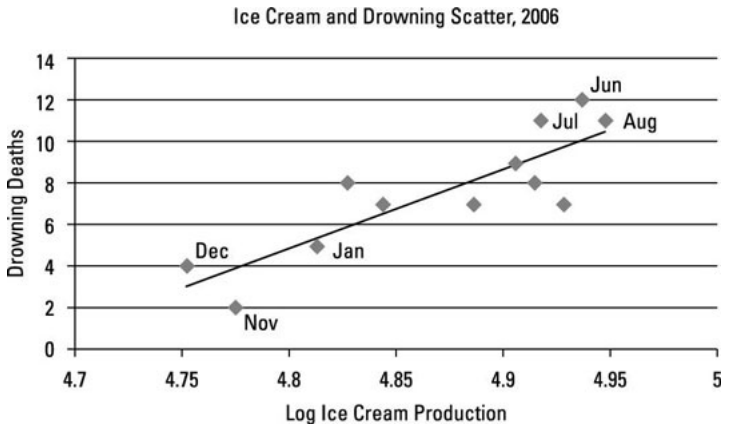
21

Causality and correlation



22

Causality and correlation

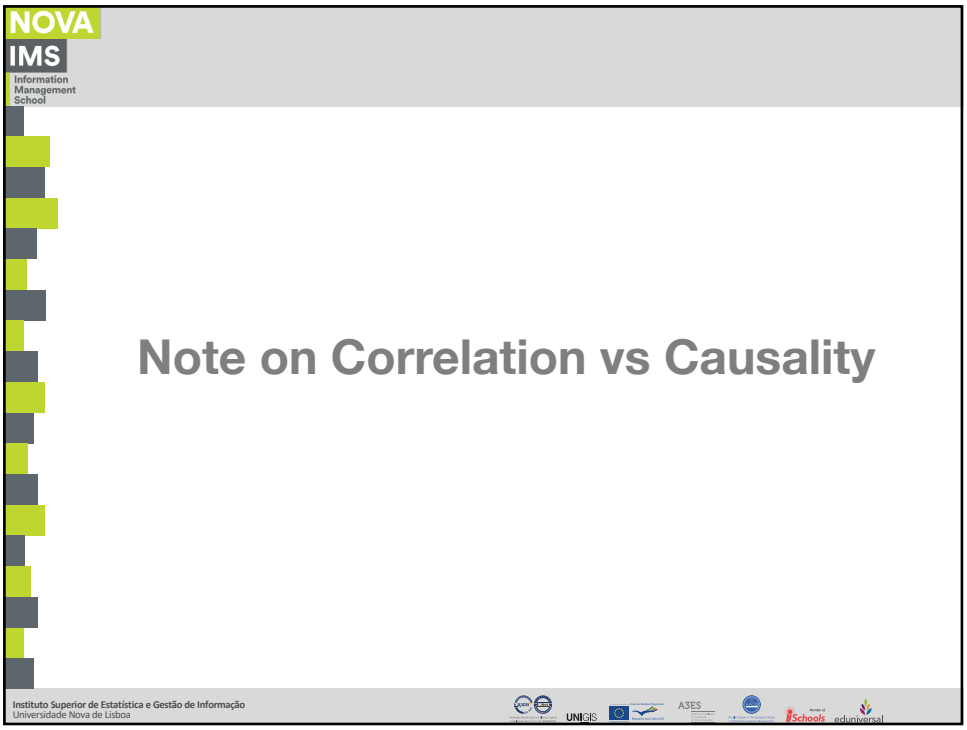


23

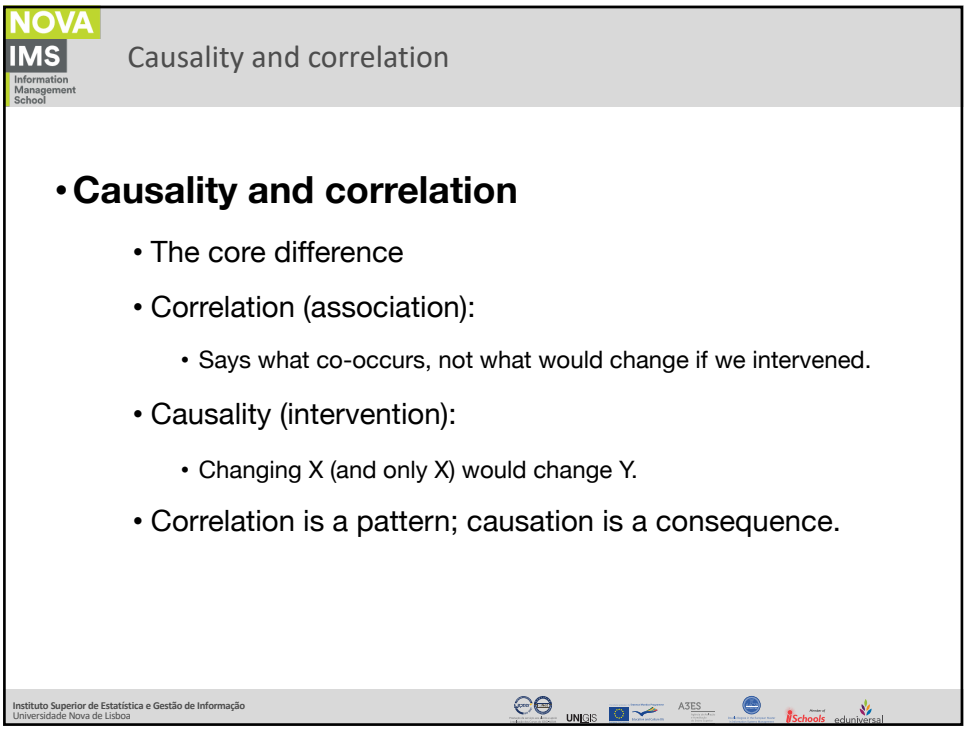
Causality and correlation

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

24


NOVA
IMS
Information Management School
Note on Correlation vs Causality
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

25


NOVA
IMS
Information Management School
Causality and correlation
• Causality and correlation
• The core difference
• Correlation (association):
• Says what co-occurs, not what would change if we intervened.
• Causality (intervention):
• Changing X (and only X) would change Y.
• Correlation is a pattern; causation is a consequence.
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

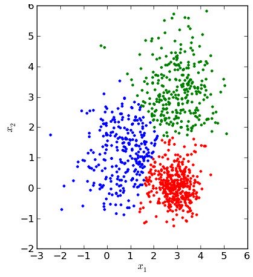
26

Input Space

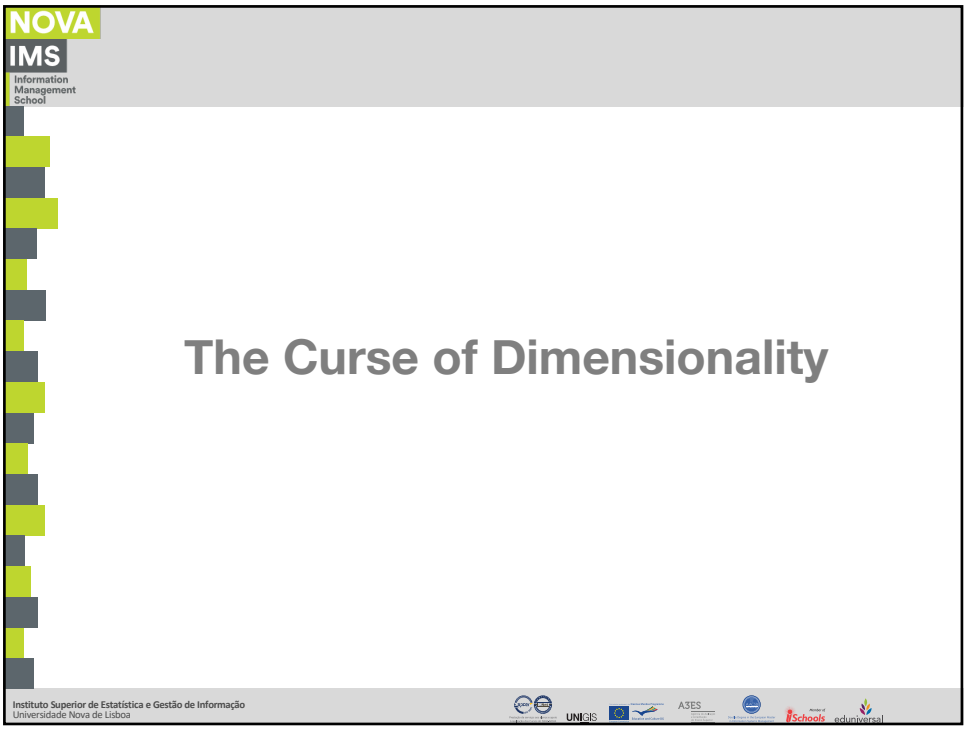
27

Problem definition

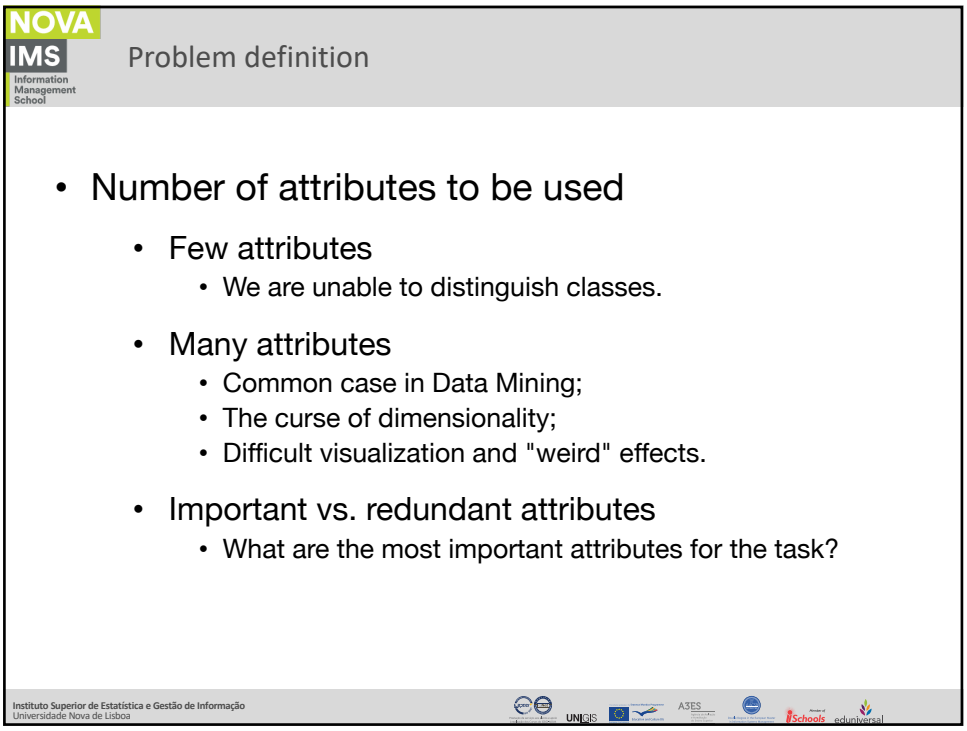
- **Input Space**
  - The input space is defined by the input feature vectors.
  - Where the algorithms will try to find a solution to the problem



28


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

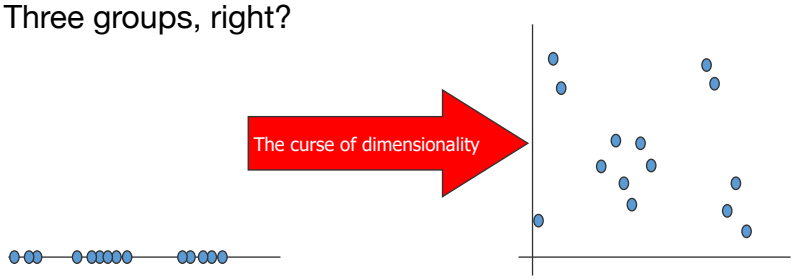
29


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

30

Problem definition

Three groups, right?

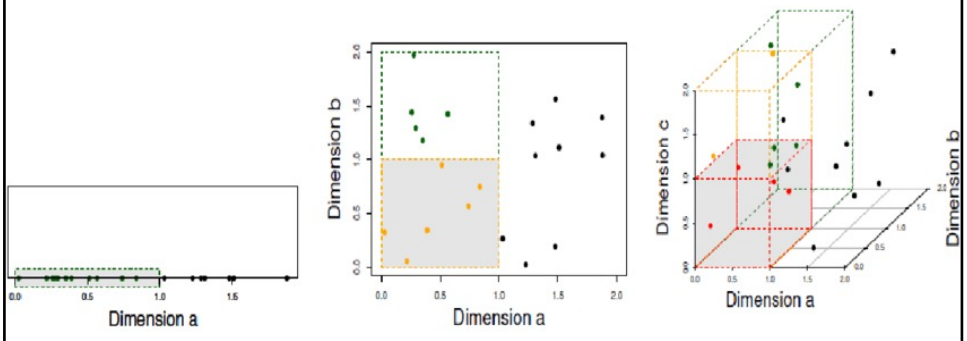


Well… not exactly.

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

31

Problem definition



32

Problem definition

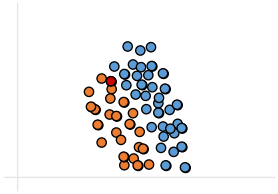
- The curse of dimensionality
  - **Generalizing correctly becomes exponentially harder** as the dimensionality (number of features) of the examples grows, because a fixed-size training set covers a dwindling fraction of the input space.
  - With a **dimension of 100 and a huge training set of a trillion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

33

Input Space Coverage

34

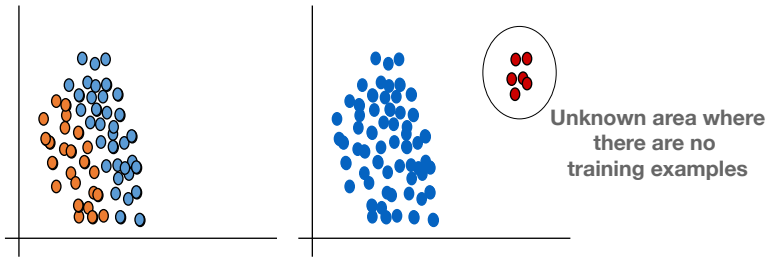
Problem definition



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

35

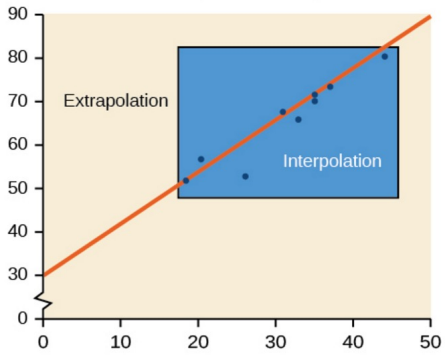
Space coverage



- 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).

36

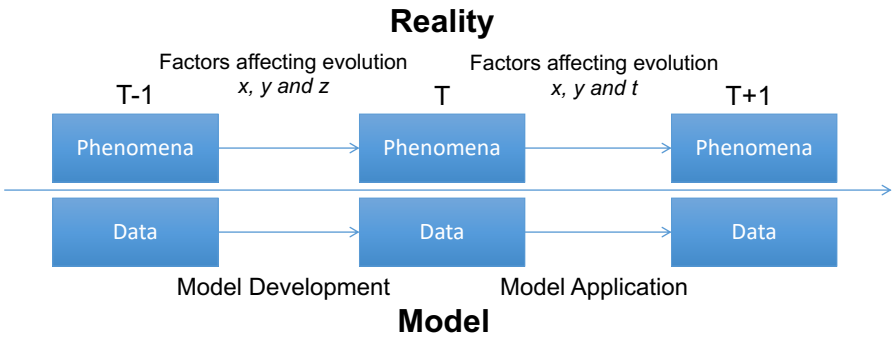
Extrapolation vs Interpolation



- **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.

37

General aspects of data collection

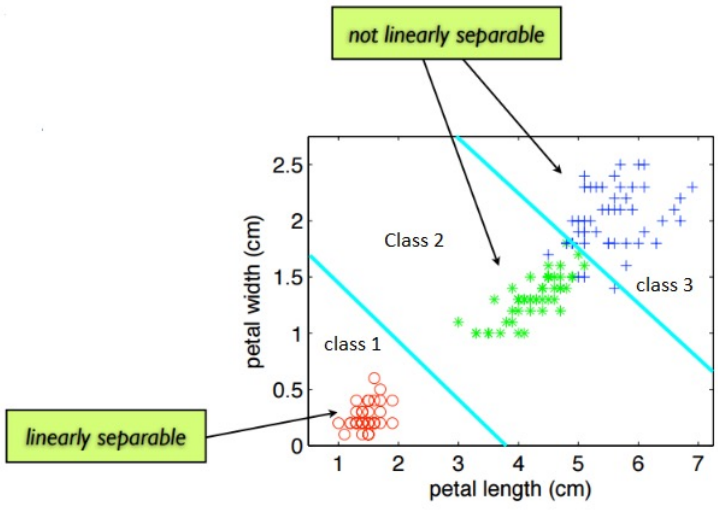


38

Separability and Bayes Error

39

Separation and error



40

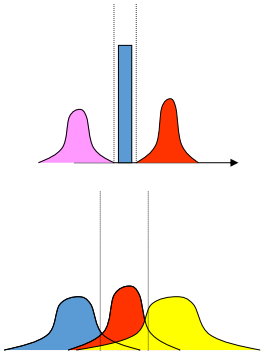
Problem definition

Separable

- Ø error possible

Not separable

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



41

Different Types of Variables

42

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

43

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.

44

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

45

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.

46

47