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

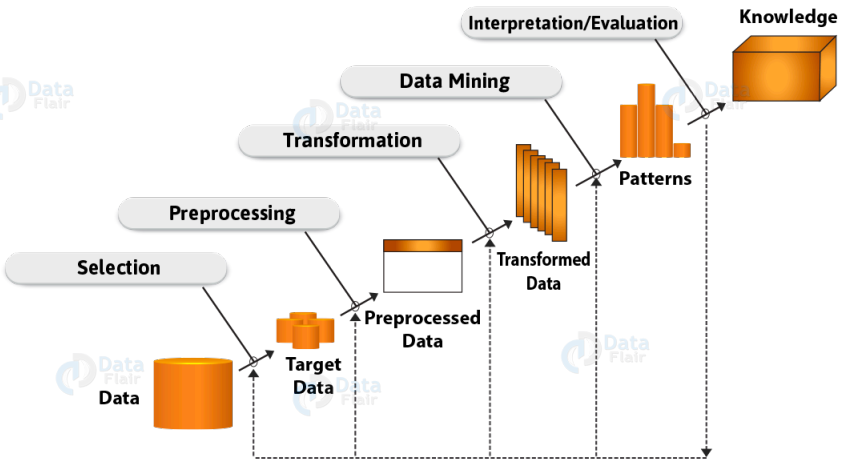
- Data Mining
  - The data mining process
  - General aspects of problem definition
  - Input space
  - The curse of dimensionality
  - Input space coverage
  - Binary and multiclass classification
  - Separability and Bayes error
  - Different types of variables
  - Spurious correlations and confounding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

2

The Data Mining Process

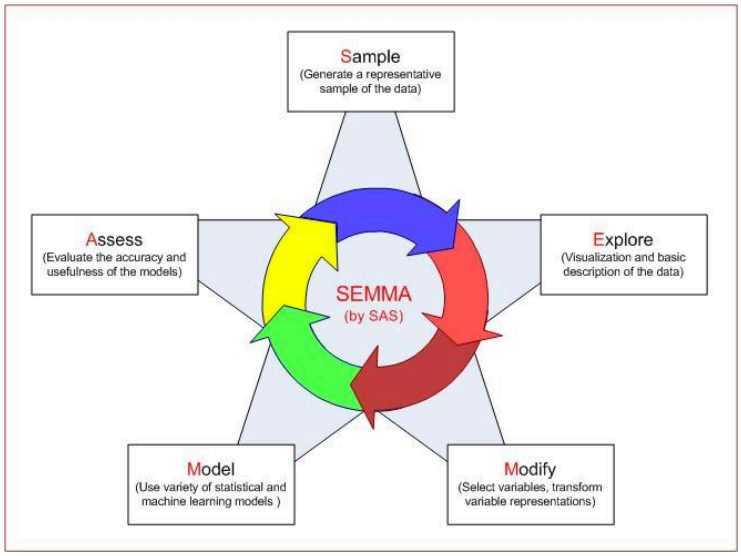
3

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

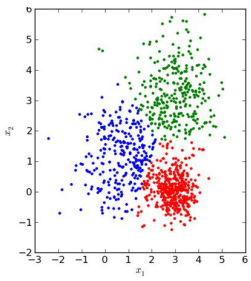
Input Space

11

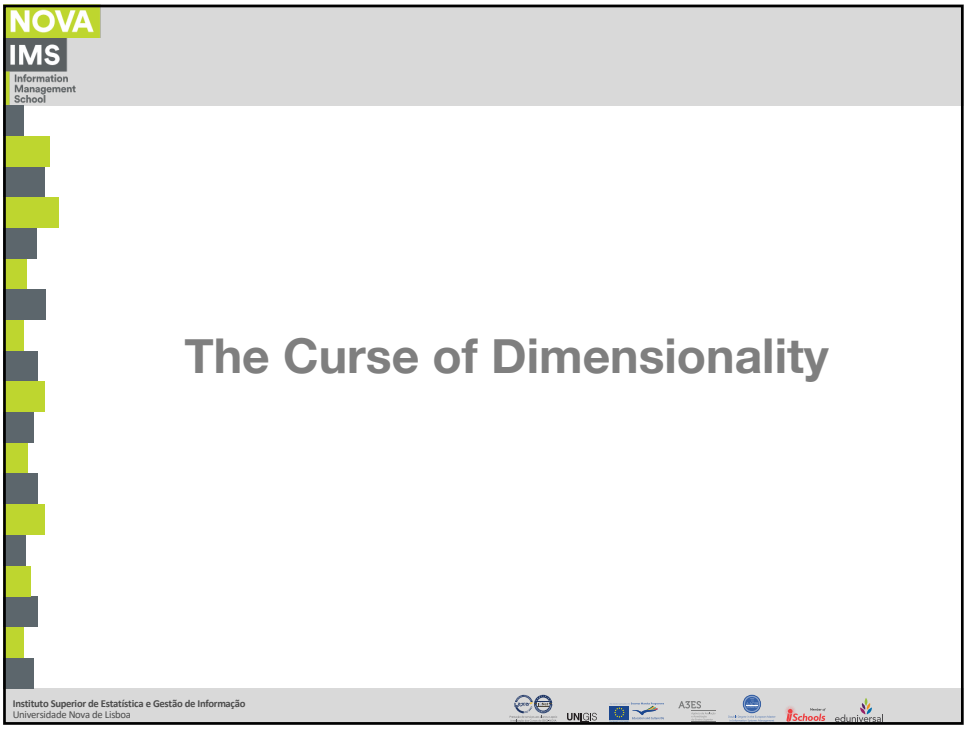
NOVA IMS Information Management School

Problem definition

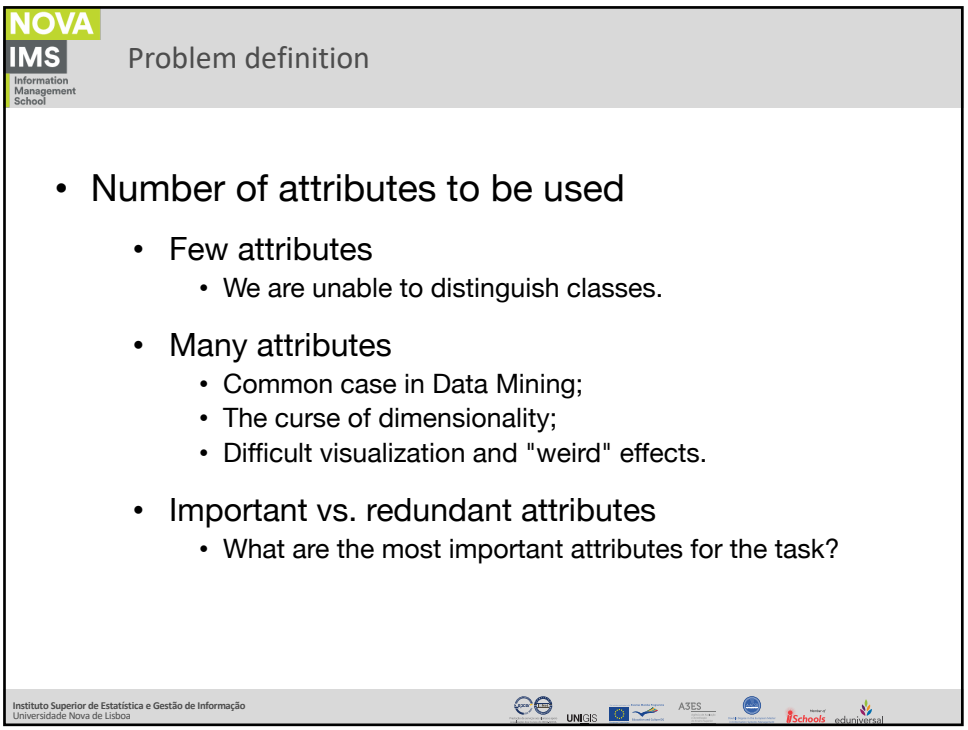
- **Input Space**
  - The input space is defined by the input feature vectors.
  - Where the algorithms will try to find a solution to the problem



12


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

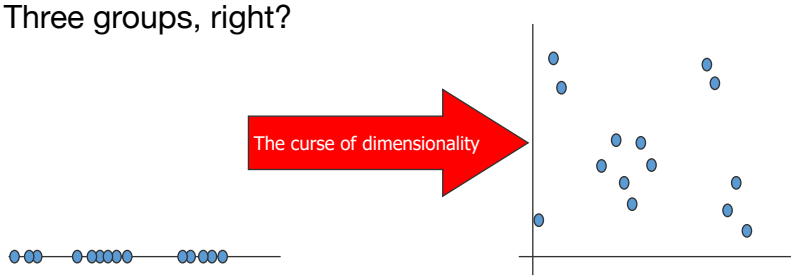
13


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?

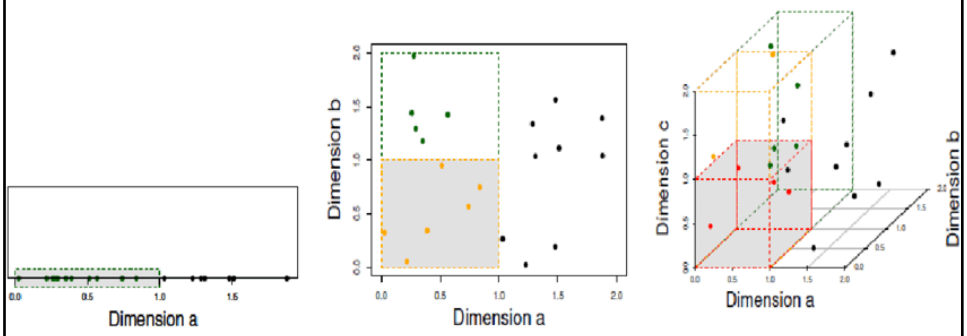


Well… not exactly.

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

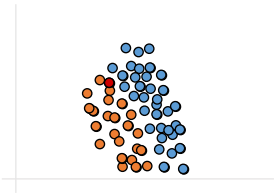
- The curse of dimensionality
  - **Generalizing correctly becomes exponentially harder** as the dimensionality (number of features) of the examples grows, because a fixed-size training set covers a dwindling fraction of the input space.
  - With a **dimension of 100 and a huge training set of a trillion 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

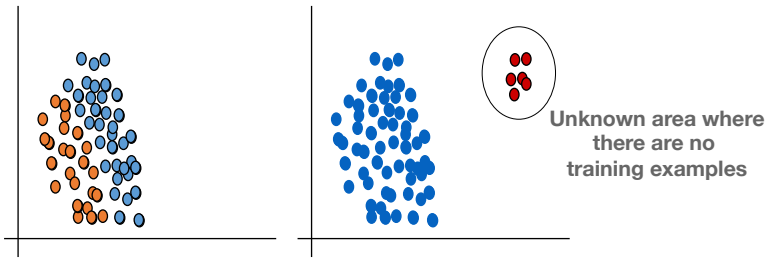
Problem definition



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

19

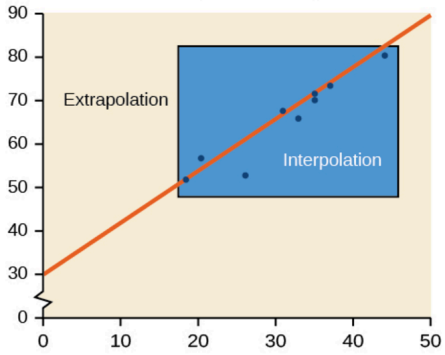
Space coverage



- 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

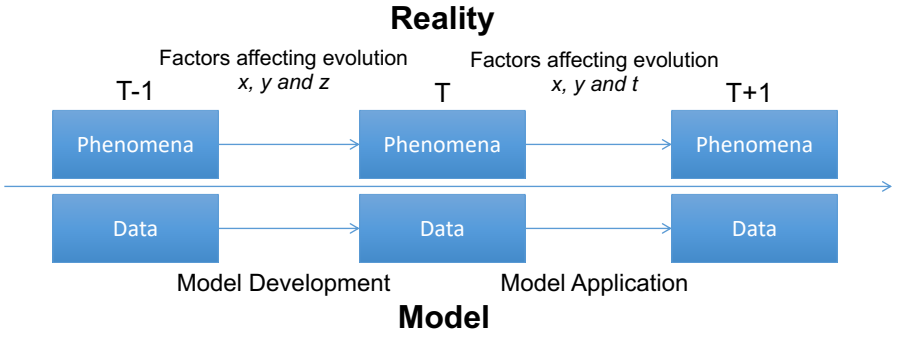
Extrapolation vs Interpolation



- **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

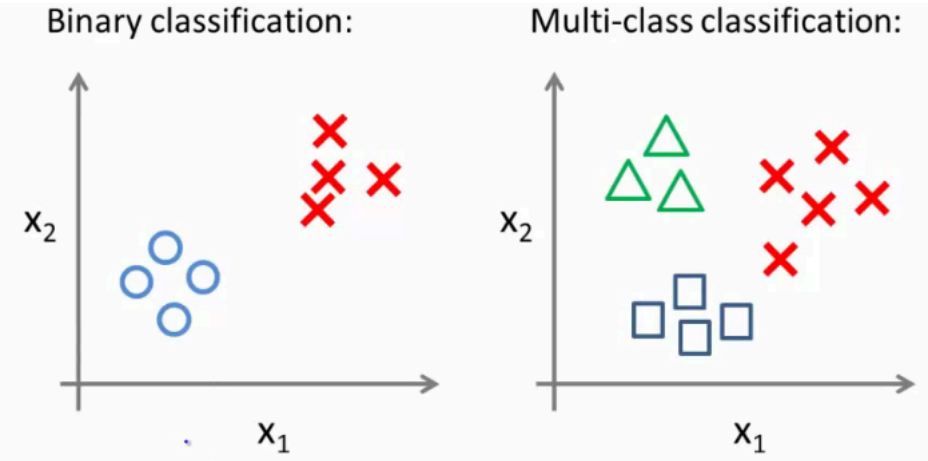
NOVA IMS Information Management School

Binary and Multiclass Classification

23

NOVA IMS Information Management School

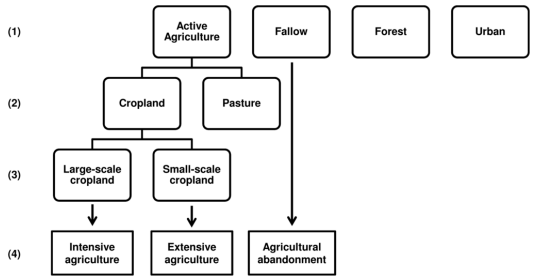
Binary and Multi-class



24

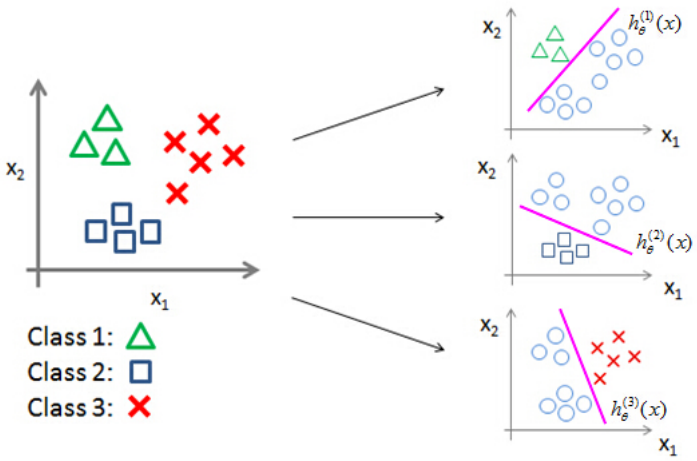
Granularity

- **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

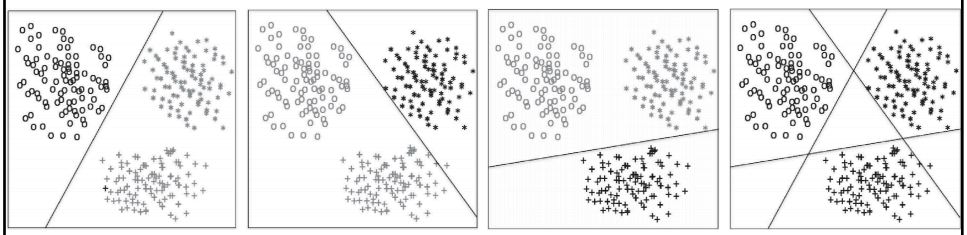


26

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.

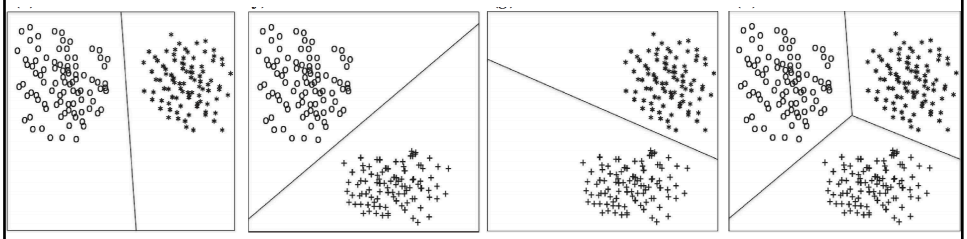


27

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

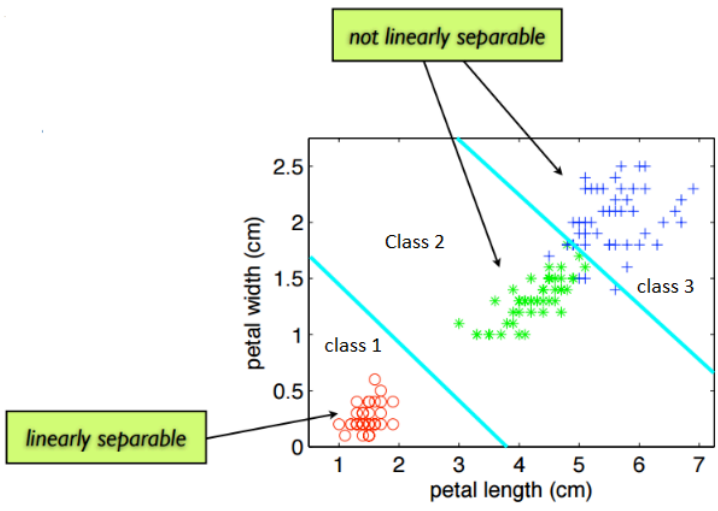
NOVA IMS Information Management School

Separability and Bayes Error

29

NOVA IMS Information Management School

Separation and error



30

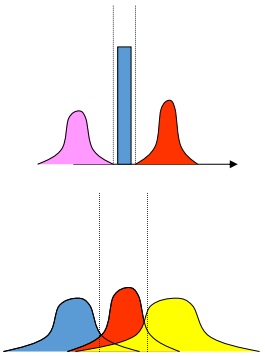
Problem definition

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

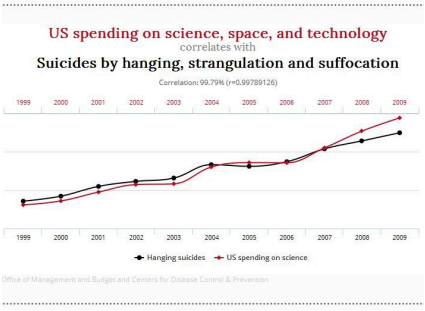
Spurious Correlations and Confounding Variables

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

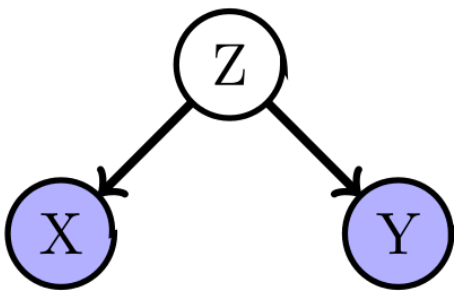
Causality and correlation

- **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