From Statistics to Data Mining

Master 1
COlour in Science and Industry (COSI)
Cyber-Physical Social System (CPS2)
Saint-Étienne, France

Fabrice MUHLENBACH
https://perso.univ-st-etienne.fr/muhlfabr/
e-mail: fabrice.muhlenbach@univ-st-etienne.fr

Organization

- Theoretical part:
  - ❑ lectures: 15 hours
  - ❑ tutorials: 15 hours

- Practical part:
  - ❑ lab sessions (with R): 15 hours

- Exam:
  70% → written exam
  30% → exercises with R

Course Outline

- Basics in probabilities
  → chance experiments, random variables, moments, law of large number...
- Statistics
  → discrete and continuous distributions, estimates, maximum likelihood estimation...
- Basics in linear algebra and in convex optimization
- Linear / Polynomial / Logistic Regression
  → closed-form solution, gradient descent...
- Principal Component Analysis
- Clustering

Introduction

- Statistics

- ➢ Statistics is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of data
- ➢ Statistics is the scientific discipline that provides methods to help us make sense of data
- ➢ Statistical methods, used intelligently, offer a set of powerful tools for gaining insight into the world around us

Introduction

- Data Analysis

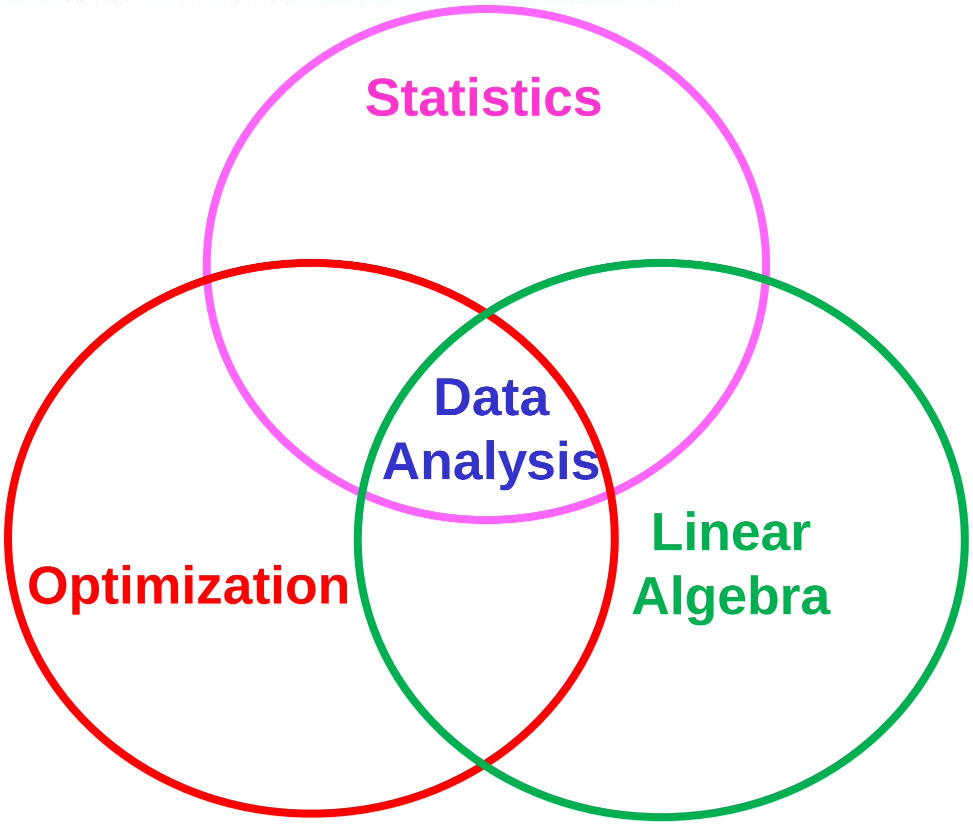
➢ Data Analysis is a process of inspecting, transforming, visualizing and modeling data with the goal of discovering useful information, suggesting conclusions, and supporting decision making

➢ The Data Analysis process can be organized into the following steps:

1. Understanding the nature of the problem and decide what to measure from a collected data set
2. Data summarization and preliminary analysis
3. Formal data analysis

Introduction

- Data Analysis



Introduction

- Pattern Recognition

- Scientific discipline whose goal is the classification of objects into a number of categories or classes
- Objects: images, signal waveforms, etc. = *patterns*
- Pattern recognition deals with the conception of automatic systems able to interpret signals of the real world
- Some application domains:
- speech recognition
- character recognition (handwritten or printed) = OCR
- vision recognition: image analysis, image segmentation
- any kind of patterns: spam, weather, plagiarism, etc.

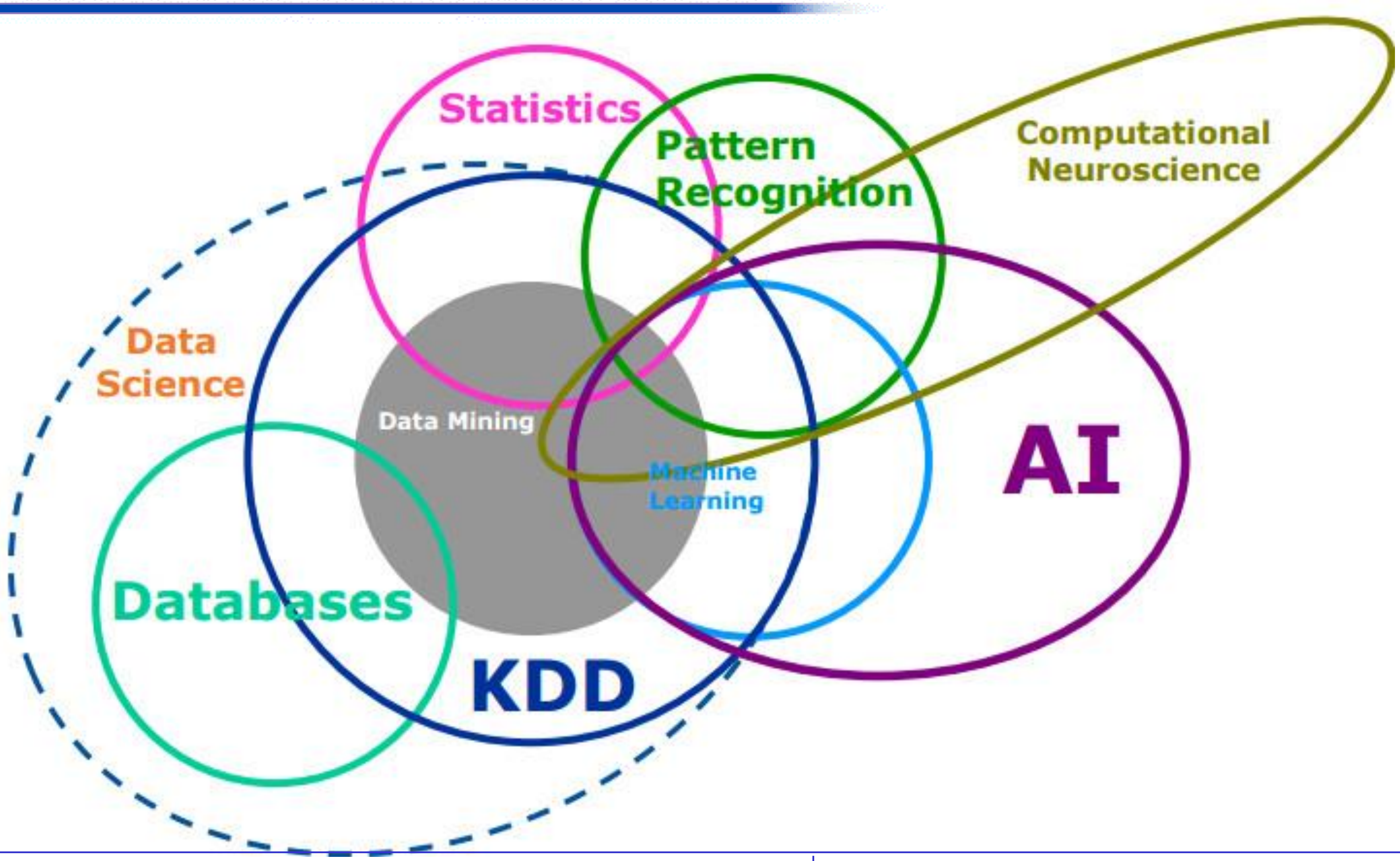
Introduction

- Data Mining

- Data Mining is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data (Fayyad *et al.*, 1996)
- Data Mining is the discovery of interesting, unexpected, or valuable structures in large data sets (Hand, 2000)
- Data mining is a process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems

Introduction

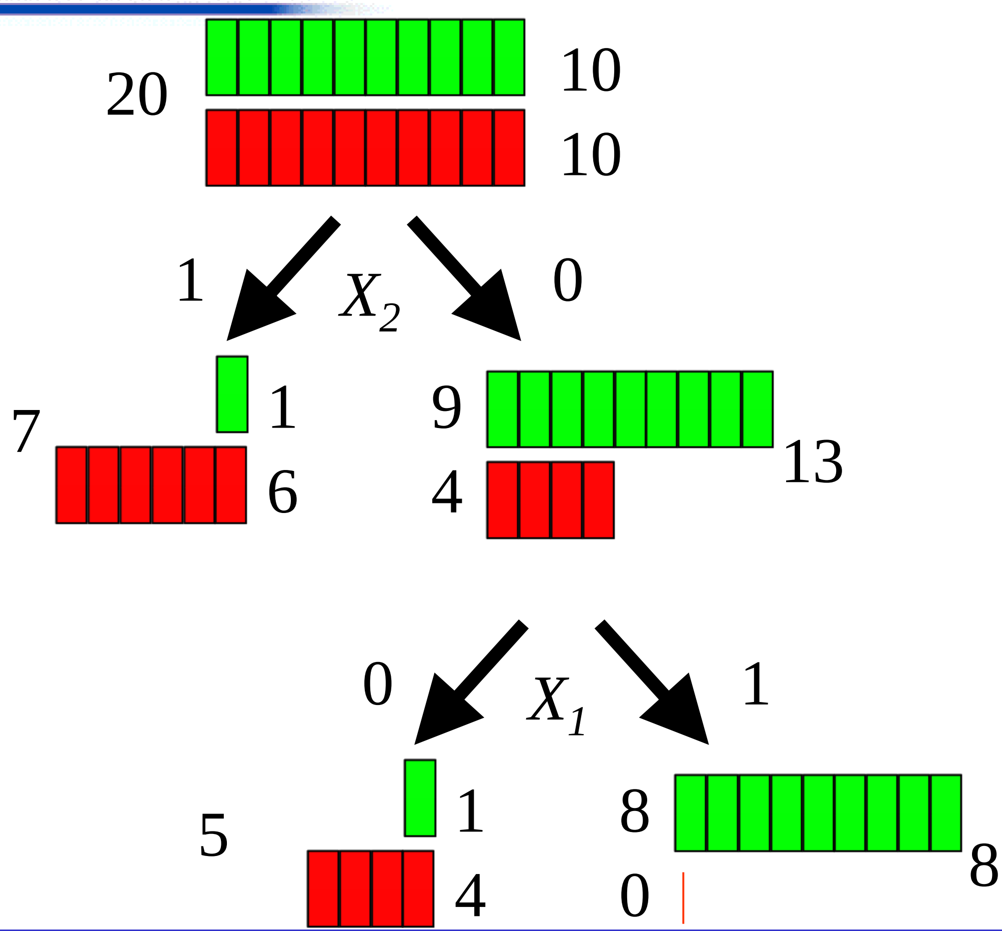
- Data Mining



Introduction

- Data Mining: some examples → Classification

Ω	C	X_1	X_2
ω_1	1	0	1
ω_2	1	0	0
ω_3	1	0	0
ω_4	1	1	0
ω_5	1	1	0
ω_6	1	1	0
ω_7	1	1	0
ω_8	1	1	0
ω_9	1	1	0
ω_{10}	1	1	0
ω_{11}	2	1	1
ω_{12}	2	0	1
ω_{13}	2	1	1
ω_{14}	2	0	1
ω_{15}	2	1	1
ω_{16}	2	1	1
ω_{17}	2	0	0
ω_{18}	2	0	0
ω_{19}	2	0	0
ω_{20}	2	0	0



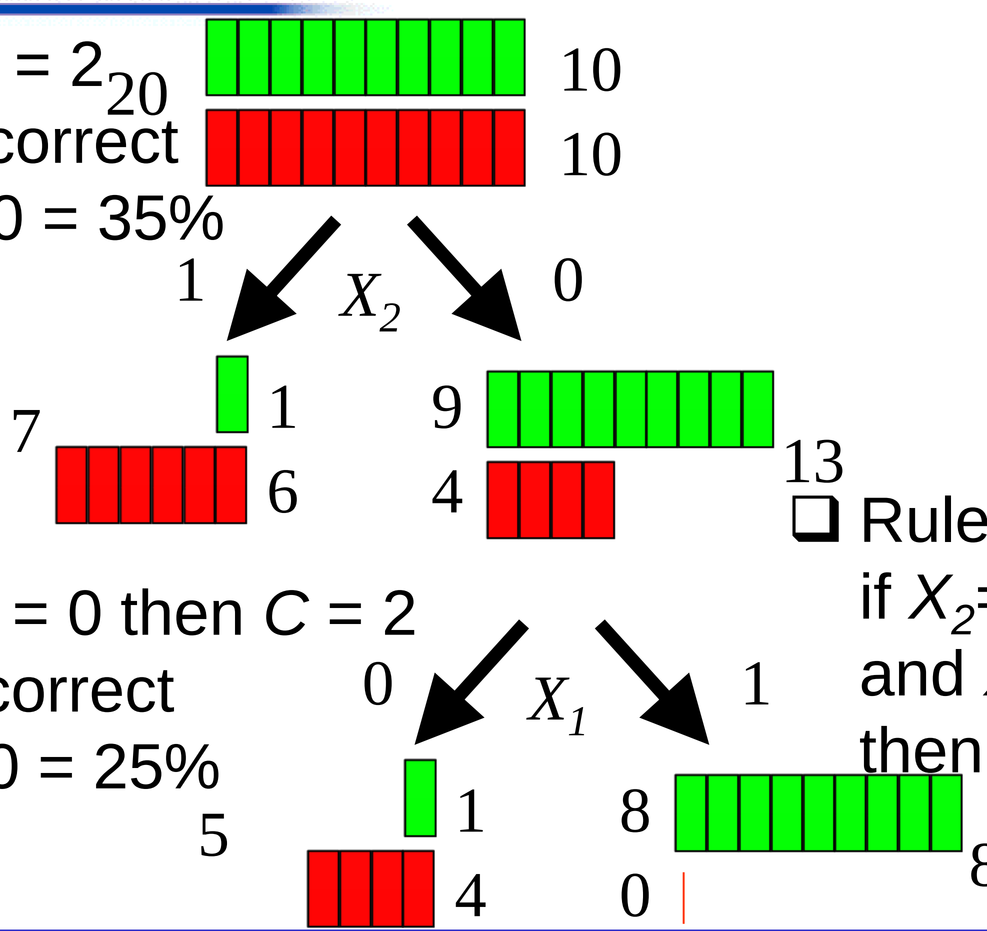
Introduction

- Data Mining: some examples → Classification

❑ Rule 1: if $X_2 = 1$ then $C = 2$

➢ the rule is 6 / 7 = 86% correct

➢ the rule represent 7 / 20 = 35% of the knowledge base



❑ Rule 2: if $X_2 = 0$ and $X_1 = 0$ then $C = 2$

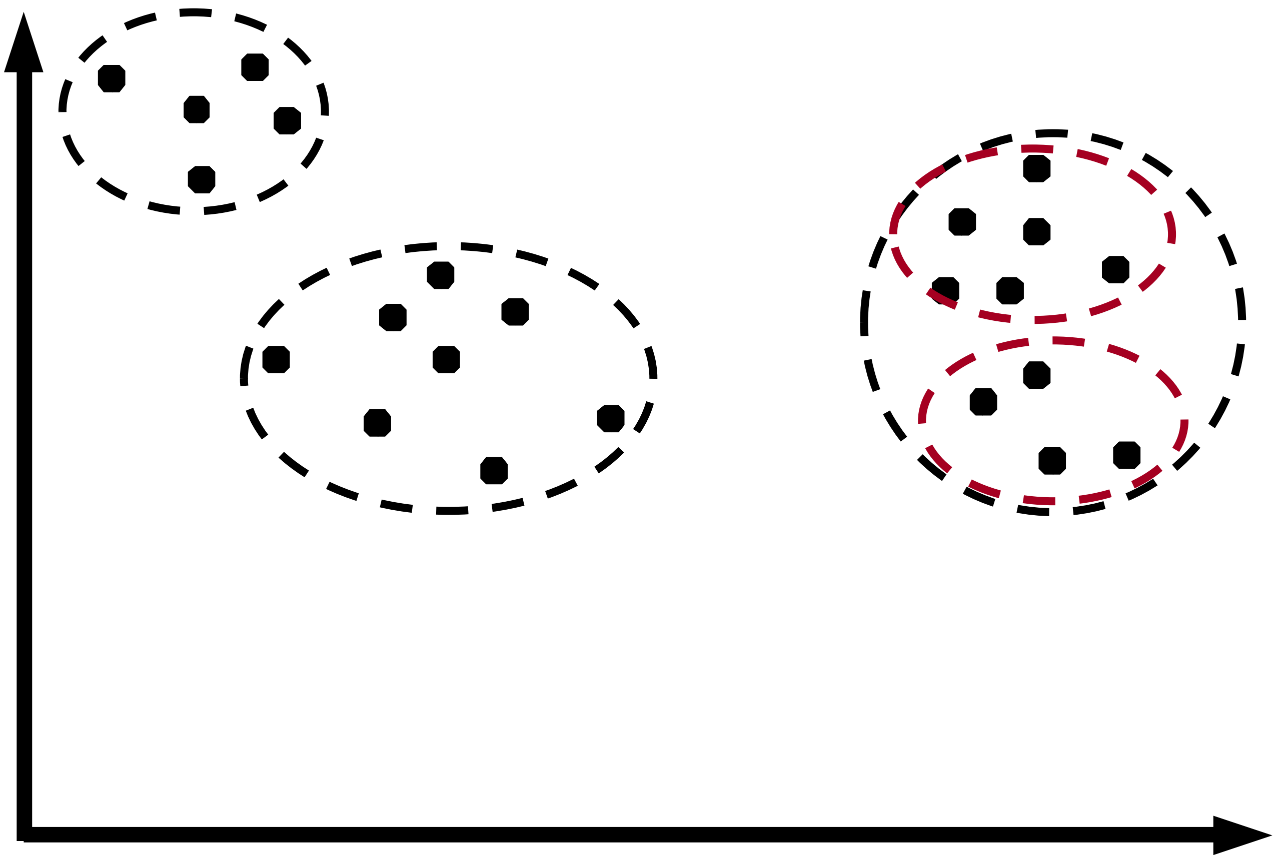
➢ the rule is 4 / 5 = 80% correct

➢ the rule represent 5 / 20 = 25% of the knowledge base

❑ Rule 3: if X_2=0 and X_1=1 then C=1

Introduction

- Data Mining: some examples → Clustering



Introduction

- Data Mining: some examples → Clustering

lion

whale

horse

cow

seal

dolphin

royal python

crocodile

Hermann's tortoise

leatherback sea turtle

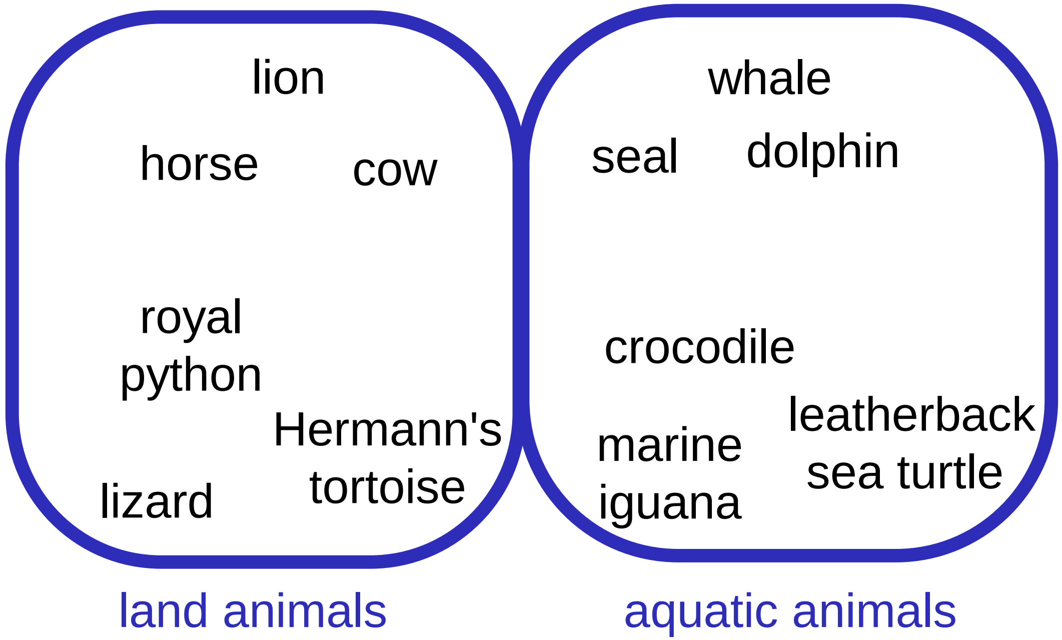
marine iguana

lizard

➢ Machine learning: unsupervised learning

Introduction

- Data Mining: some examples → Clustering



➢ Machine learning: unsupervised learning

Introduction

- Data Mining: some examples → Clustering

lion
whale
horse
cow
seal
dolphin

mammals

royal python
crocodile
Hermann's tortoise
leatherback sea turtle
marine iguana
lizard

reptiles

➢ Machine learning: unsupervised learning

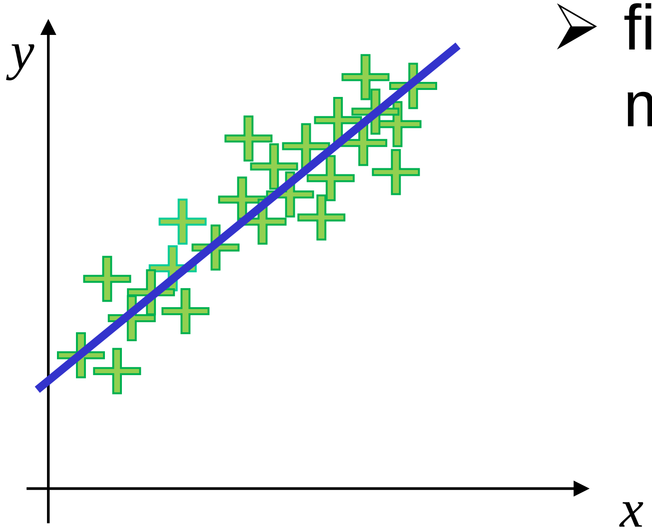
Introduction

- Data Mining: some examples → Regression

➢ Machine learning: supervised learning

➢ objective: learn (predict) a particular variable (known as the "class" variable) based on other variables

➢ case of a numeric variable → regression



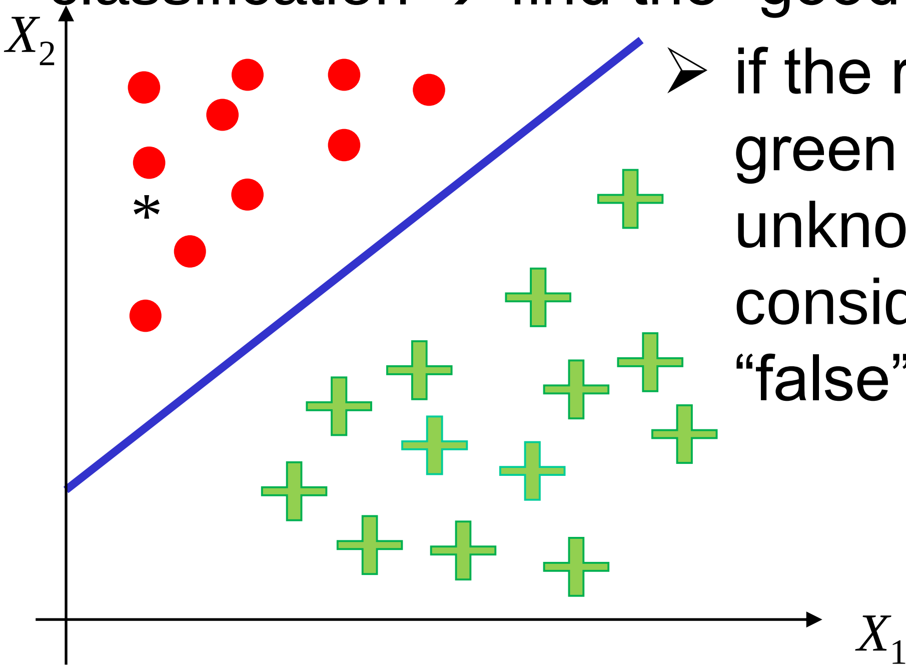
➢ find the values of a and b of a model such as $y = a\,x + b$

Introduction

- Data Mining: some examples → Classification

➢ case of a categorical variable → find the answer to a “yes” or “no” question, or a category of the class variable

➢ classification → find the “good” class value

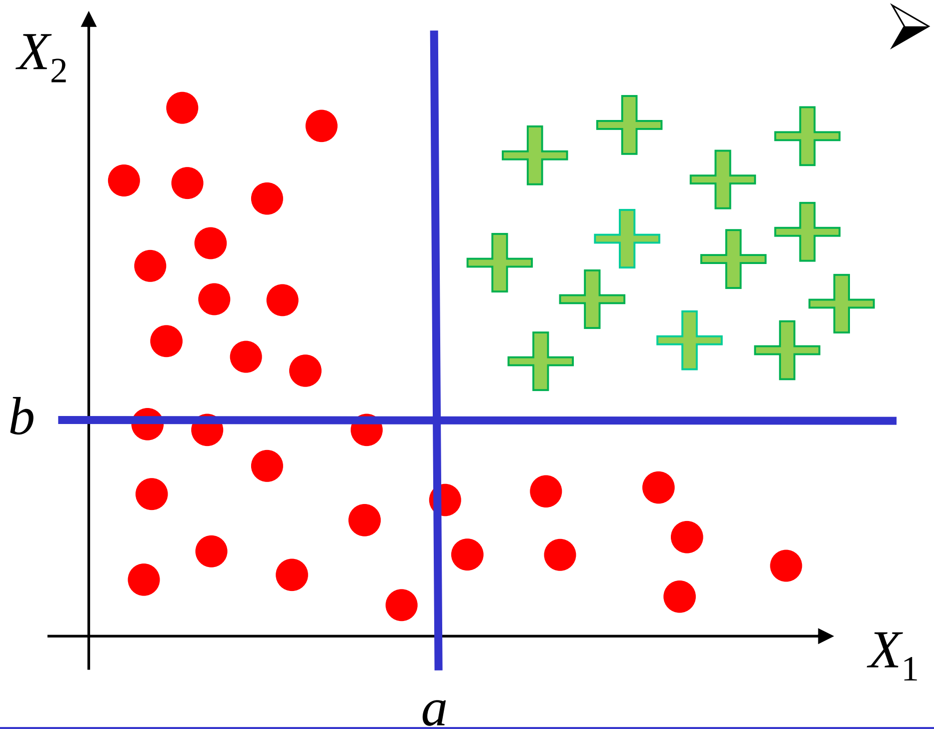


➢ if the red dots are “false” and the green crosses are “true”, the unknown example (star) will be considered to be an example of “false” (similar to the red dots)

Introduction

- Data Mining: some examples → Classification

➢ Rule-based model → decision tree



➢ Use of thresholds (a and b)

➢ if $X_1 > a$ and $X_2 > b$,
then Y is true else Y is false

Introduction

- Data Mining: some examples → Association rules

➢ Example: market basket analysis

Transaction 1	🍎 🍺 🍚 🍗
Transaction 2	🍎 🍺 🍚
Transaction 3	🍎 🍺
Transaction 4	🍎 🍐
Transaction 5	🍼 🍺 🍚 🍗
Transaction 6	🍼 🍺 🍚
Transaction 7	🍼 🍺
Transaction 8	🍼 🍐

ID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}
...	...

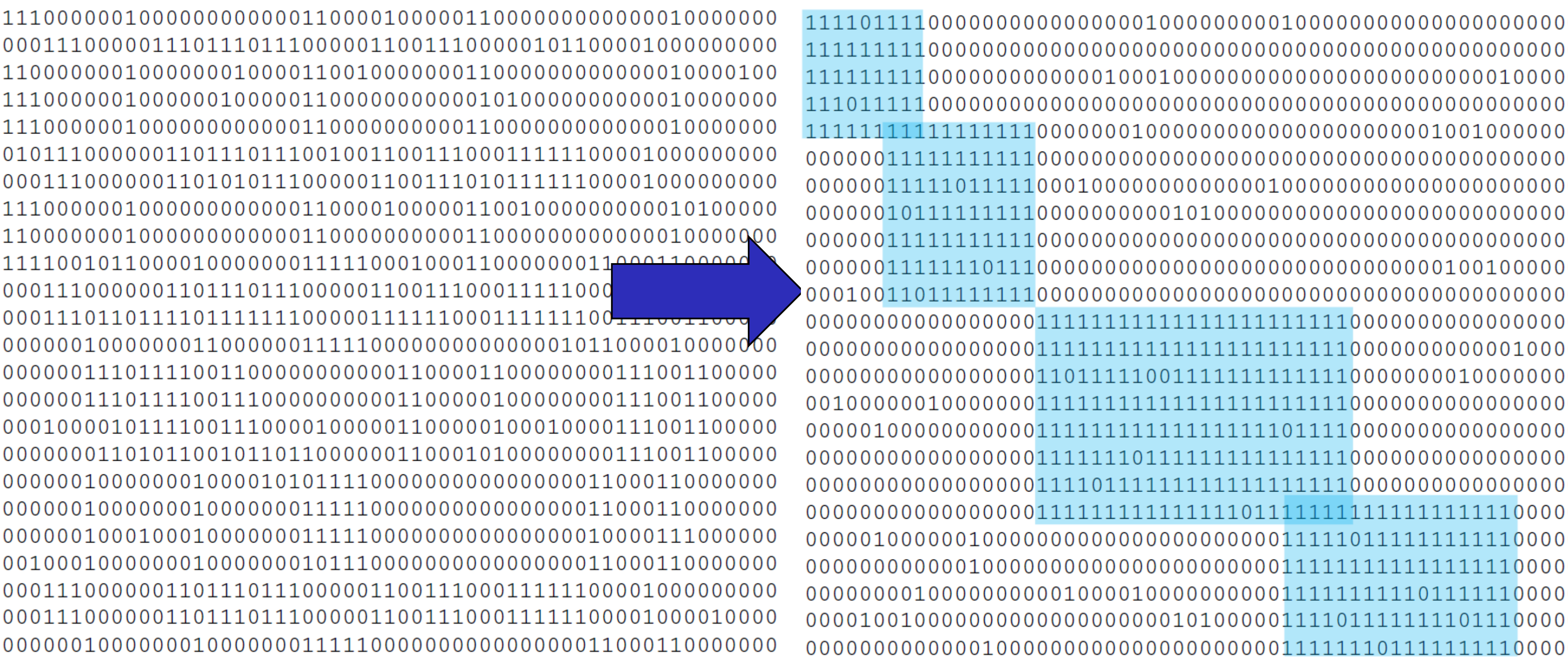
market basket transactions

{Diapers, Beer} Example of a frequent itemset

{Diapers} → {Beer} Example of an association rule

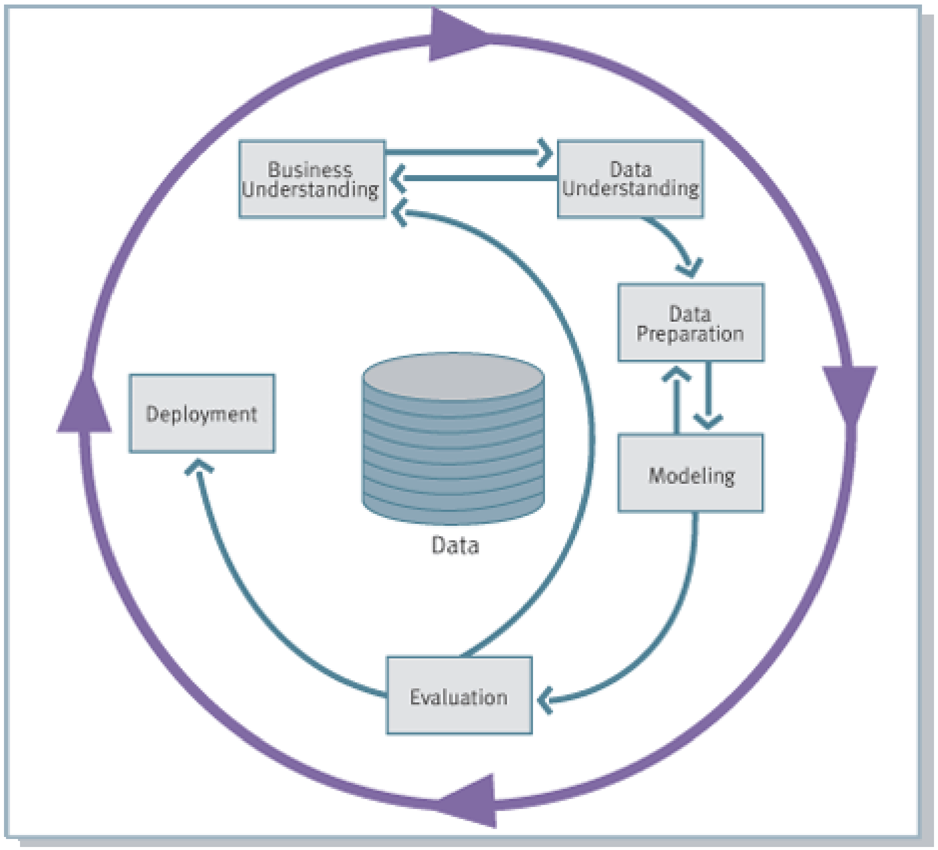
Introduction

- Data Mining: some examples → Pattern Mining



Introduction

- Data Mining process



Introduction

- Data Mining methods and techniques

➢ Descriptive methods:

- ❑ factor analysis → principal component analysis
- ❑ cluster analysis → partitioning / hierarchal / neural clustering
- ❑ link detection → search for association rules

➢ Predictive methods:

- ❑ logical rule-based models → decision trees
- ❑ models based on mathematical functions → neural networks, parametric or non-parametric models (regression)
- ❑ prediction without model → probabilistic analysis (k-NN)

Introduction

- Machine Learning

- ➢ As a broad subfield of artificial intelligence, machine learning is concerned with the design and development of algorithms and techniques that allow computers to improve their performances by learning
- ➢ A machine learning technique can be:
- ❑ Supervised: the classes of patterns are a priori known
- ❑ Unsupervised (= Clustering)
- ❑ Semi-supervised: only few labeled examples

Introduction

- History of Data Mining and Machine Learning techniques

- ➢ 1700s: Bayes' theorem
- ➢ 1800s: regression analysis
- ➢ 1950s: neural networks, clustering, genetic algorithms
- ➢ 1960s: decision trees
- ➢ 1980s: support vector machines
- ➢ 1990s: association rules learning
- ➢ 2000s: domain-specific data mining → Web mining, social network analysis, text mining, sequential pattern mining...
- ➢ 2010s: deep learning

Introduction

- Classification and Clustering

➢ Classification (or supervised learning):

❑ **Context:** the training data are labeled
A label characterizes a class of objects that share similar features (e.g. female vs. male)

❑ **Task:** assign unknown objects (patterns) into the correct class

➢ Clustering (or unsupervised learning):

❑ **Context:** no training data, with class labeling, are available

❑ **Task:** group the data into a number of sensible clusters (groups)

Introduction

• Features and Feature vectors

➢ Features

- measurable quantities obtained from the patterns
- the classification task is based on their respective values
- features = random variables = attributes

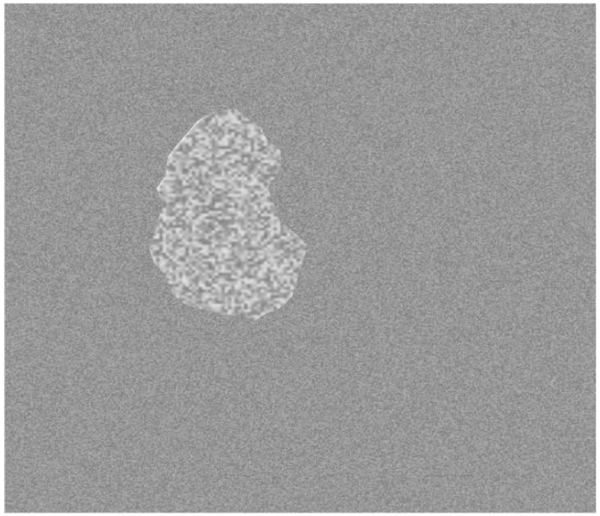
➢ Feature vector

- l features x_i, $i = 1, 2, \ldots, l$ are used
  and they form the feature vector $x = [x_1, x_2, \ldots, x_l]^T$
  where T denotes the transposition

Introduction

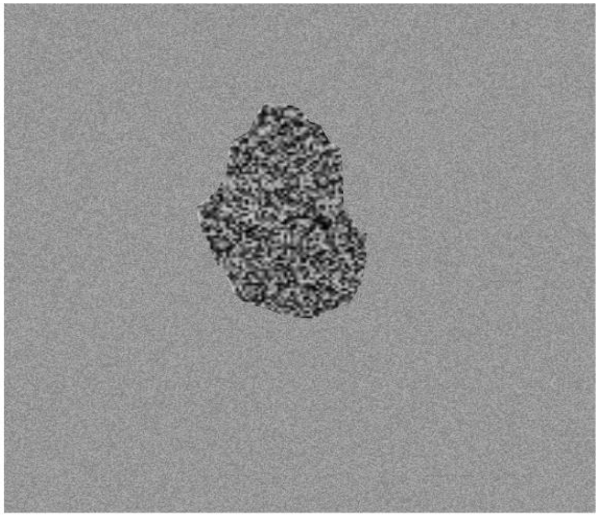
• Classification: an example

➢ A medical image classification task



(a)

benign lesion



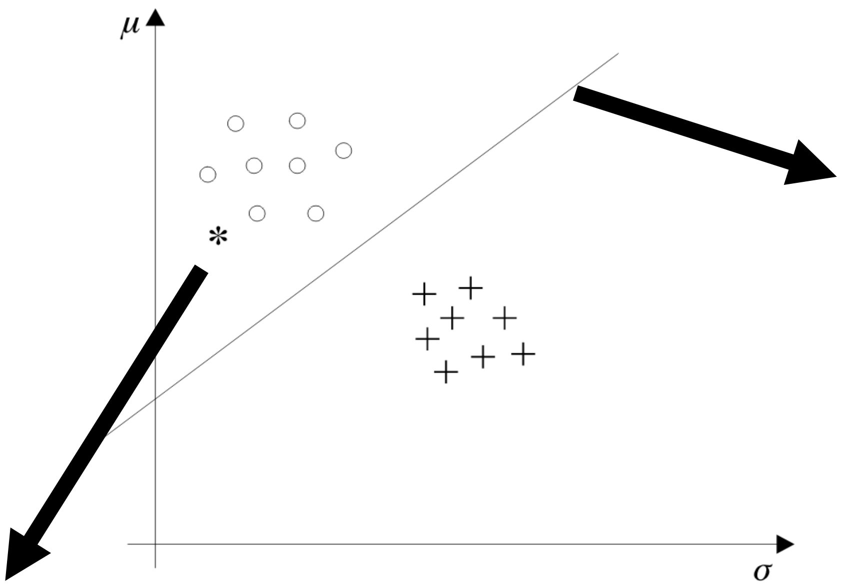
(b)

malignant lesion (cancer)

Introduction

• Classification: an example

Plot of the mean value *vs.* the standard deviation for a number of different images originating from class A (O) and class B (+)



a straight line separates the two classes
= decision line

the unknown pattern shown by the asterisk (∗) is more likely to belong to class A than class B

Introduction

- Classifier and Classification System

➢ The classifier consists of a set of functions, whose values, computed at *x*, determine the class to which the corresponding pattern belongs

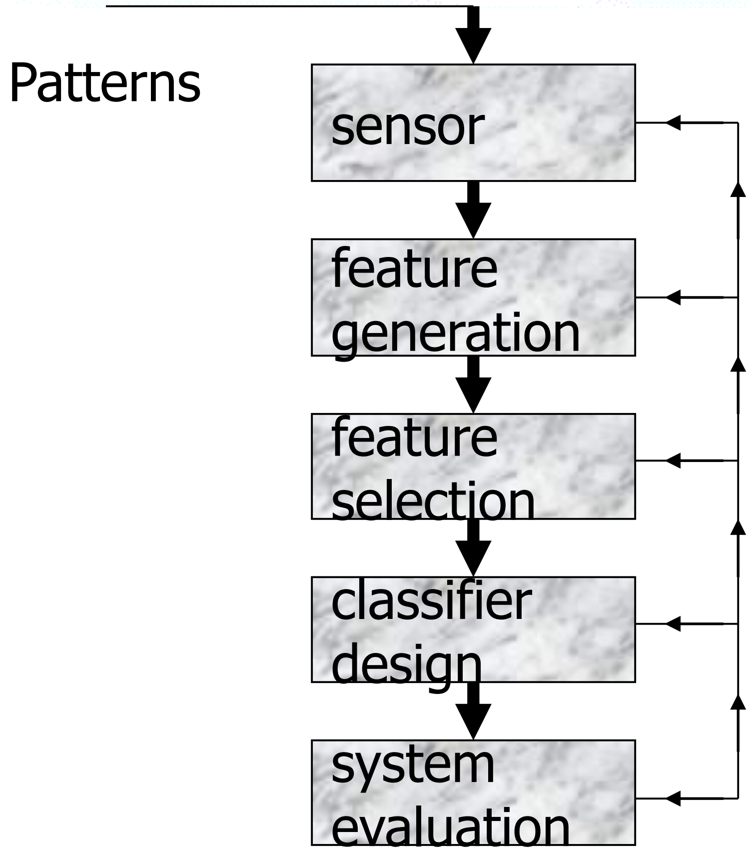
➢ The straight line in the previous example is known as the *decision line*, and it constitutes the *classifier* whose role is to divide the feature space into regions that correspond to either class A or class B

➢ When a decision made by the classifier is not correct, a *misclassification* has occurred

➢ The patterns whose true class is known and which are used for design of the classifier are known as *training patterns*

Introduction

- Classification system overview



how are the features generated?

what is the best number *l* of features to use?

how does one design the classifier?

how can one assess the performance of the designed classifier?

Introduction

- Scheme of the supervised learning (1/2)

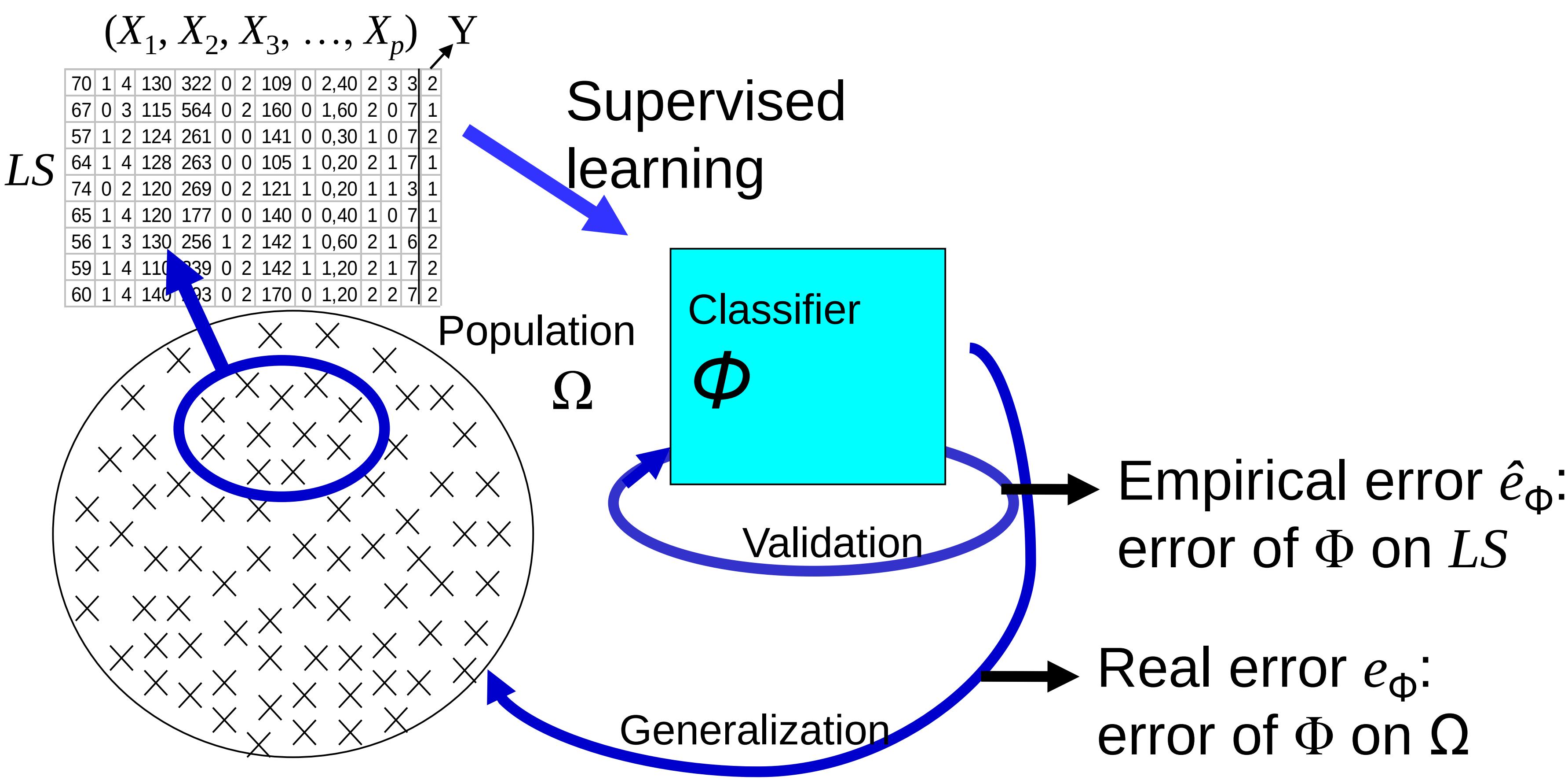
Input: ⟶ a set of training data is available (the learning set)

⟶ $LS = \{(X(\omega) = (X_1(\omega), X_2(\omega), \ldots, X_p(\omega)), Y(\omega))\}$

Output: ⟶ a classifier $\Phi(\omega) : X \rightarrow Y, \forall \omega \in \Omega$

Introduction

- Scheme of the supervised learning (2/2)



70	1	4	130	322	0	2	109	0	2,40	2	3	3	2
67	0	3	115	564	0	2	160	0	1,60	2	0	7	1
57	1	2	124	261	0	0	141	0	0,30	1	0	7	2
64	1	4	128	263	0	0	105	1	0,20	2	1	7	1
74	0	2	120	269	0	2	121	1	0,20	1	1	3	1
65	1	4	120	177	0	0	140	0	0,40	1	0	7	1
56	1	3	130	256	1	2	142	1	0,60	2	1	6	2
59	1	4	110	239	0	2	142	1	1,20	2	1	7	2
60	1	4	140	[illegible]93	0	2	170	0	1,20	2	2	7	2

Introduction

- Estimation of the generalization error

1) Estimation using the learning set LS (Resubstitution Method):

→ This method estimates the generalization (or real) error e of Φ directly from LS, using the empirical error $\hat{e}$ computed from LS

→ $\Phi: X \rightarrow Y$

$\omega \in LS \rightarrow \Phi(\omega)$

→ We can deduce an estimate empirical error $\hat{e}$ of the generalization error e:

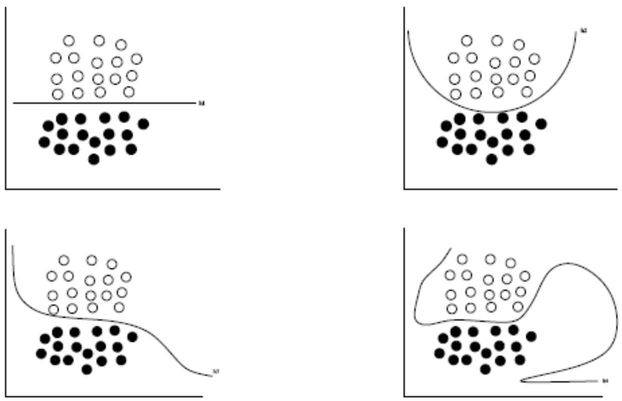
$$\hat{e} = \frac{1}{|LS|} \sum_{\omega \in LS} 1_{[\phi(\omega) \neq Y(\omega)]}$$

Drawback: this way of proceeding is too optimistic because it tends to overestimate the generalization ability of Φ, and does not allow us to detect overfitting situations (Breiman 84)

Introduction

- Estimation of the generalization error

What is "overfitting" in machine learning?



→ From a same machine learning problem, several families of classifiers can be used leading to the same error rate

Occam's razor: "*No sunt multiplicanda entia praeter necessitatem*"

→ the best solution is often the one that calls on the smallest number of concepts

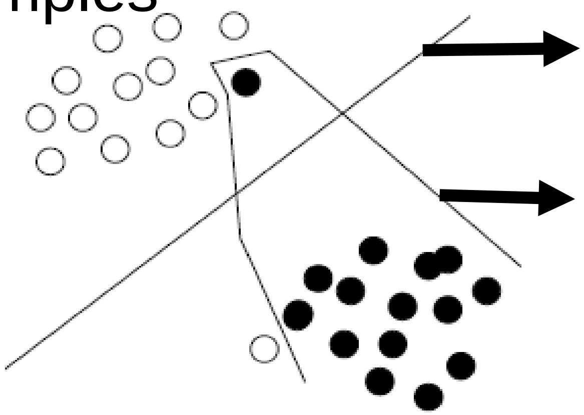
→ between two classifiers (with the same empirical error on LS), choose the simplest one

Introduction

- Estimation of the generalization error

What is "overfitting" in machine learning?

→ In some situations, we can even prefer to build a classifier making some errors on LS, rather than learning by heart the examples



→ We can bound the generalization error e of a classifier as follows: $e < \hat{e} + \Lambda(\frac{d_H}{|LS|})$

Introduction

- Estimation of the generalization error

2) Estimation using a test set *TS* (Holdout Method):

→ This method consists of splitting *LS* in two subsets such that $LS = LS^* \cup T$. LS^* is used to build Φ, while T is used to test Φ on examples that have not been used for its inference, but for which the label $Y(\omega)$ is known

→ $\Phi: X \rightarrow Y$

$\omega \in LS^* \rightarrow \Phi(\omega)$

→ We can deduce an estimate empirical error $\hat{e}'$ of the generalization error e: $\hat{e}' = \frac{1}{|T|} \sum_{\omega \in T} 1_{[\phi(\omega) \neq Y(\omega)]}$

Drawback: this solution reduces the number of examples available for learning Φ

Introduction

- Estimation of the generalization error

3) Estimation by Cross-Validation:

Input: A learning algorithm LA, a set of examples LS

Output: an estimate $\hat{e}'$

Divide randomly LS in k subsets $S_1, \ldots, S_k$;

for *i=1 to k* **do**

  Run the algorithm LA on the sample $S - S_i$ and generate the classifier ϕ_i;

Deduce the estimate of the error such that $\hat{e}' = \frac{1}{k}\sum_{i=1}^{k} \hat{e}'_i$ where $\hat{e}'_i$ is the error of ϕ_i on the subset S_i;

Introduction

- Estimation of the generalization error

4) Estimation by Bootstrap:

→ Algorithm

Input: A learning algorithm LA, a set of examples LS

Output: an estimate $\hat{e}'$

for i=*1 to* k **do**

Draw with replacement, a subset S_i of size $|LS|$;

Run the algorithm LA on S_i and generate the classifier ϕ_i;

Deduce the estimate of the error such that $\hat{e}' = \frac{1}{k}\sum_{i=1}^{k} \hat{e}'_i$ where $\hat{e}'_i$ is the error of ϕ_i on the subset S_i;

→ "bootstrap" = new data sets are artificially generated by *random* sampling with *replacement*

Bibliography

- Fayyad, U. M., G. Piatetsky-Shapiro, and P. Smyth (1996). "Knowledge discovery and data mining: Towards a unifying framework". In E. Simoudis, J. Han, and U. M. Fayyad (Eds.), *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pp. 82–88. AAAI Press. https://www.aaai.org/Papers/KDD/1996/KDD96-014.pdf
- Hand, D. J. (2000). Data mining: New challenges for statisticians. *Social Science Computer Review, 18* (4), pp. 442–449. https://journals.sagepub.com/doi/pdf/10.1177/089443930001800407
- Peck R., C. Olsen, and J. L. Devore (2016). *Introduction to Statistics and Data Analysis*, 5th edition, Boston: Cengage Learning