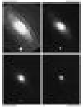
Machine Learning

Michele Sebag — Alexandre Allauzen
TAO, CNRS — INRIA — LRI — Université Paris-Sud

Orsay — Oct. 2017

Where we are

Ast. series 

Pierre de Rosette 

World

Natural phenomenons

Human–related phenomenons

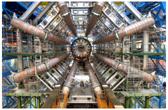
Data / Principles

Maths. Modelling

Common Sense

You are here

Where we are

Sc. data 

Google

World

Natural phenomenons

Human–related phenomenons

Data / Principles

Maths. Modelling

Common Sense

You are here

Types of application

Domain | **But : Modelling**

Physical phenomenons — **analysis & control**

manufacturing, experimental sciences, numerical engineering

Vision, speech, robotics..

Social phenomenons — **+ privacy**

Health, Insurance, Banks ...

Individual phenomenons — **+ dynamics**

Consumer Relationship Management, User Modelling

Social networks, games...

RoadMap

Decision trees

Types of Machine Learning problems

WORLD – DATA – USER

Observations	**+ Target**	+ Rewards
Understand **Code**	**Predict** **Classification/Regression**	Decide Policy
Unsupervised **LEARNING**	**Supervised** **LEARNING**	Reinforcement LEARNING

Data

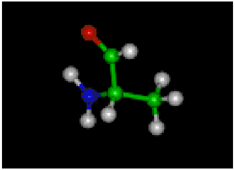
Example

- row : example/ case
- column : feature/ variable/ attribute
- attribute : class/ label

Instance space $\mathcal{X}$

- Propositionnal : $\mathcal{X} \equiv \mathbb{R}^d$
- Structured : sequential, spatio-temporal, relational.

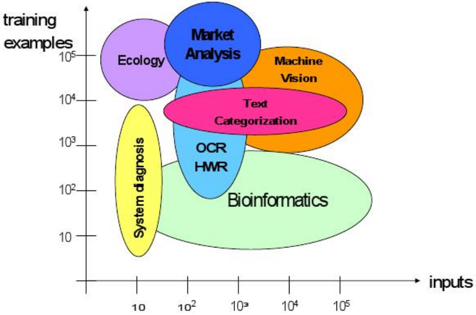
age	employme	education	edun	marital	...	job	relation	race	gender	hour	country	wealth
					...							
39	State_gov	Bachelors	13	Never_mar	...	Adm_cleri	Not_in_fam	White	Male	40	United_Sta	poor
51	Self_emp_	Bachelors	13	Married	...	Exec_man	Husband	White	Male	13	United_Sta	poor
39	Private	HS_grad	9	Divorced	...	Handlers_	Not_in_fam	White	Male	40	United_Sta	poor
54	Private	11th	7	Married	...	Handlers_	Husband	Black	Male	40	United_Sta	poor
28	Private	Bachelors	13	Married	...	Prof_speci	Wife	Black	Female	40	Cuba	poor
38	Private	Masters	14	Married	...	Exec_man	Wife	White	Female	40	United_Sta	poor
50	Private	9th	5	Married_sp	...	Other_serv	Not_in_fam	Black	Female	16	Jamaica	poor
52	Self_emp_	HS_grad	9	Married	...	Exec_man	Husband	White	Male	45	United_Sta	rich
31	Private	Masters	14	Never_mar	...	Prof_speci	Not_in_fam	White	Female	50	United_Sta	rich
42	Private	Bachelors	13	Married	...	Exec_man	Husband	White	Male	40	United_Sta	rich
37	Private	Some_coll	10	Married	...	Exec_man	Husband	Black	Male	80	United_Sta	rich
30	State_gov	Bachelors	13	Married	...	Prof_speci	Husband	Asian	Male	40	India	rich
24	Private	Bachelors	13	Never_mar	...	Adm_cleri	Own_child	White	Female	30	United_Sta	poor
33	Private	Assoc_acc	12	Never_mar	...	Sales	Not_in_fam	Black	Male	50	United_Sta	poor
41	Private	Assoc_voc	11	Married	...	Craft_repai	Husband	Asian	Male	40	*MissingV	rich
34	Private	7th_8th	4	Married	...	Transport_	Husband	Amer_India	Male	45	Mexico	poor
26	Self_emp_	HS_grad	9	Never_mar	...	Farming_fi	Own_child	White	Male	35	United_Sta	poor
33	Private	HS_grad	9	Never_mar	...	Machine_c	Unmarried	White	Male	40	United_Sta	poor
38	Private	11th	7	Married	...	Sales	Husband	White	Male	50	United_Sta	poor
44	Self_emp_	Masters	14	Divorced	...	Exec_man	Unmarried	White	Female	45	United_Sta	rich
41	Private	Doctorate	16	Married	...	Prof_speci	Husband	White	Male	60	United_Sta	rich
:	:	:	:	:	:	:	:	:	:	:	:	:



aminoacid

Data / Applications

- Propositionnal data — 80% des applis.
- Spatio-temporal data — alarms, mines, accidents
- Relationnal data — chemistry, biology
- Semi-structured data — text, Web
- Multi-media — images, music, movies,..



Difficulty factors

Quality of data / of representation

- − Noise; missing data
- + Relevant attributes **Feature extraction**
- − Structured data: spatio-temporal, relational, text, videos,..

Data distribution

- + Independants, identically distributed examples
- − Other: robotics; data streams; heterogeneous data

Prior knowledge

- + Goals, interestingness criteria
- + Constraints on target hypotheses

Difficulty factors, 2

Learning criterion

- $+$ Convex optimization problem
- $\searrow$ Complexity : n, $nlogn$, n^2 **Scalability**
- $-$ Combinatorial optimization

Learning criteria, 2

The user's criteria

- Relevance, causality,
- INTELLIGIBILITY
- Simplicity
- Stability
- Interactive processing, visualisation
- ... Preference learning

Difficulty factors, 3

Crossing the chasm

- No *killer algorithm*
- Little expertise about algorithm selection

How to assess an algorithm

- Consistency

When number n of examples goes to infinity
and target concept h^* is in $\mathcal{H}$
h^* is found:

$$lim_{n \rightarrow \infty} h_n = h^*$$

- Speed of convergence

$$||h^* - h_n|| = \mathcal{O}(1/n), \mathcal{O}(1/\sqrt{n}), \mathcal{O}(1/\ln n)$$

Context

Disciplines et critres

- Data bases, Data Mining

  Scalability

- Statistics, data analysis

  Predefined models

- Machine learning

  Prior knowledge; complex data/hypotheses

- Optimisation

  well / ill posed problems

- Computer Human Interaction

  No final solution: a process

- High performance computing

  Distributed processing; safety

Supervised Machine Learning

Context

$$\text{World} \rightarrow \text{instance } \mathbf{x}_i \rightarrow \begin{array}{c}\text{Oracle}\\ \downarrow \\ y_i\end{array}$$

Input

Training set $\mathcal{E} = \{(\mathbf{x}_i, y_i), i = 1 \ldots n, \mathbf{x}_i \in \mathcal{X}, y_i \in \mathcal{Y}\}$

Milestones

- Select hypothesis space $\mathcal{H}$
- Assess hypothesis $h \in \mathcal{H}$ $\qquad score(h)$
- Find best hypothesis h^*

iid

iid: Independent identically distributed.

Independent

$$(x_i, y_i) \text{ does not depend on } (x_j, y_j)$$

Counter-example:

- x_i is the vector of sensor values of the robot at time i

Identically distributed

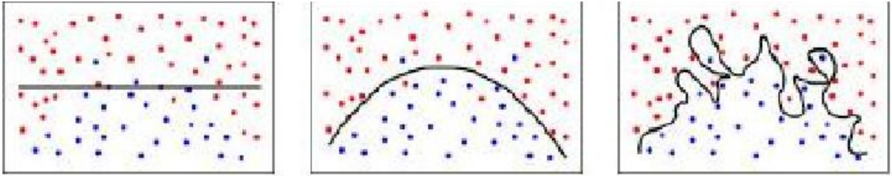
$$x_i \quad \text{are drawn after the same distribution}$$

Counter-example:

- x_i is the length travelled for fixed actuator values; the distribution changes as the robot goes on different types of ground.

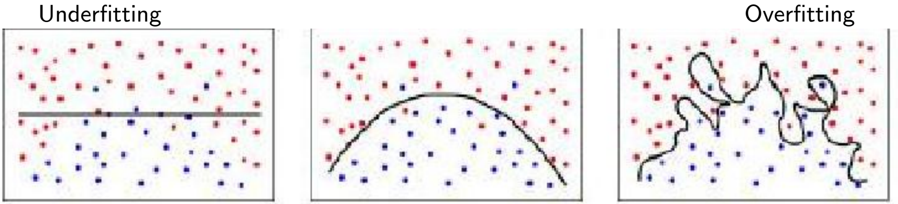
What is the goal ?

Underfitting Overfitting



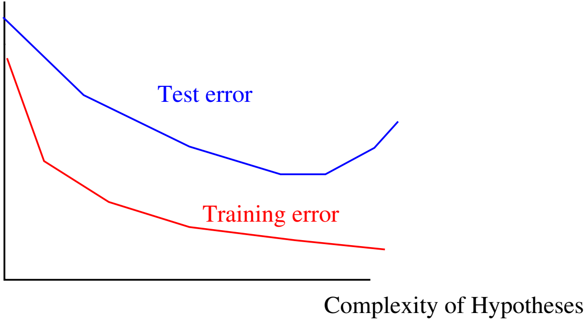
The goal is not to be perfect on the training set

What is the goal ?



The goal is not to be perfect on the training set

The villain: overfitting



What is the goal ?

Prediction good on future instances

Necessary condition:

Future instances must be similar to training instances

"identically distributed"

Minimize (cost of) errors $\ell(y, h(x)) \geq 0$

not all mistakes are equal.

Error: theoretical approach

Minimize expectation of error cost Generalization error

$$\text{Minimize } E[\ell(y, h(x))] = \int_{X \times Y} \ell(y, h(x)) p(x, y) dx\, dy$$

Error: theoretical approach

Minimize expectation of error cost Generalization error

$$\text{Minimize } E[\ell(y,h(x))] = \int_{X \times Y} \ell(y,h(x))p(x,y)dx\,dy$$

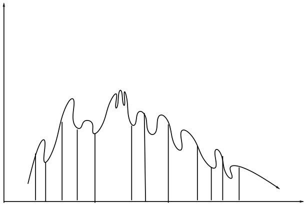
Define Empirical Error

$$Err_e(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, h(x_i))$$

Principle

Si function F "is well-behaved" on space $\mathcal{X}$ and $\mathcal{E}$ is a "sufficient" sample of $\mathcal{X}$, then integral of F on $\mathcal{X}$ is close to its empirical average on $\mathcal{E}$.

$$E[F] \leq \frac{\sum_{i=1}^{n} F(x_i)}{n} + c(F,n)$$



Classification, criteria

Generalisation error

$$Err(h) = E[\ell(y, h(x))] = \int \ell(y, h(x))dP(x, y)$$

Empirical error

$$Err_e(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, h(x_i))$$

Bound

risk minimization

$$Err(h) < Err_e(h) + \mathcal{F}(n, d(\mathcal{H}))$$

$d(\mathcal{H})$ = VC-dimension of $\mathcal{H}$

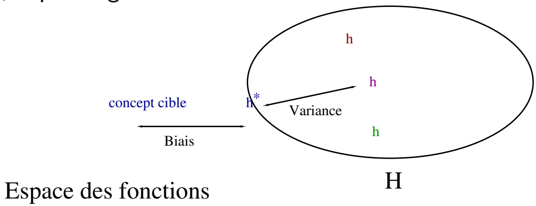
Classification: Ingredients of error

Bias

Bias ($\mathcal{H}$): error of the best hypothesis h^* in $\mathcal{H}$

Variance

Variance of h_n depending on $\mathcal{E}$

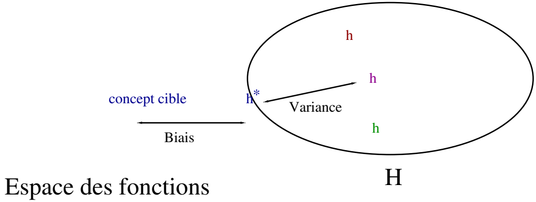


The Bias-Variance trade-off

As hypothesis space increases, bias decreases; but variance increases.

Classification: Ingredients of error

Bias
Bias ($\mathcal{H}$): error of the best hypothesis h^* in $\mathcal{H}$
Variance Variance of h_n depending on $\mathcal{E}$



Optimization
negligible in small scale
takes over in large scale (Google)

Classification, Problem posed

INPUT $\sim P(x, y)$

$$\mathcal{E} = \{(x_i, y_i), x_i \in \mathcal{X}, y_i \in \{0, 1\}, i = 1 \dots n\}$$

HYPOTHESIS SPACE SEARCH SPACE

$$\mathcal{H} \qquad h : \mathcal{X} \mapsto \{0, 1\}$$

LOSS FUNCTION

$$\ell : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$$

OUTPUT

$$h^* = arg \ max\{score(h), h \in \mathcal{H}\}$$

Key notions

- The main issue regarding supervised learning is overfitting.
- How to tackle overfitting:
  - Before learning: use a sound criterion — regularization
  - After learning: cross-validation — **Case studies**

Summary

- Learning is a search problem
- What is the space ? What are the navigation operators ?

Hypothesis Spaces

Logical Spaces

$$\text{Concept} \leftarrow \bigvee \bigwedge \text{Literal,Condition}$$

- Conditions $=$ [color $=$ blue]; [age $<$ 18]
- Condition $f : X \mapsto \{True, False\}$
- Find: disjunction of conjunctions of conditions

- Ex: (unions of) rectangles of the 2D-plane$X$.

Hypothesis Spaces

Numerical Spaces

$$\text{Concept} \; = (h() > 0)$$

- $h(x) =$ polynomial, neural network, ...
- $h : X \mapsto \mathbb{R}$
- Find: (structure and) parameters of h

Hypothesis Space $\mathcal{H}$

Logical Space

- h covers one example x iff $h(x) = True$.
- $\mathcal{H}$ is structured by a partial order relation

$$h \prec h' \text{ iff } \forall x, h(x) \rightarrow h'(x)$$

Numerical Space $\mathcal{H}$

- $h(x)$ is a real value (more or less far from 0)
- we can define $\ell(h(x), y)$
- $\mathcal{H}$ is structured by a partial order relation

$$h \prec h' \text{ iff } E[\ell(h(x), y)] < E[\ell(h'(x), y)]$$

Hypothesis Space $\mathcal{H}$ / Navigation

	$\mathcal{H}$	navigation operators
Version Space	Logical	spec / gen
Decision Trees	Logical	specialisation
Neural Networks	Numerical	gradient
Support Vector Machines	Numerical	quadratic opt.
Ensemble Methods	–	adaptation $\mathcal{E}$

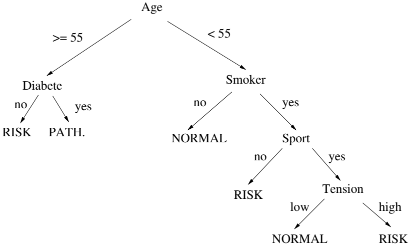
RoadMap

Decision trees

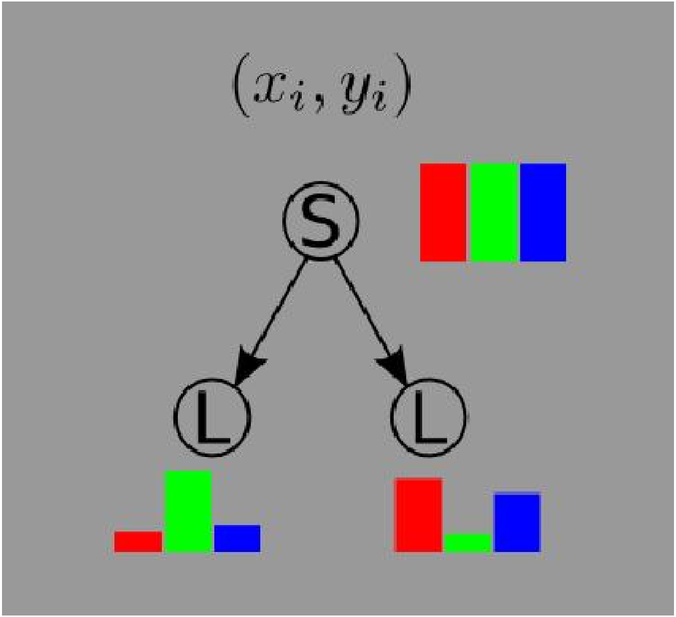
Decision Trees

C4.5 (Quinlan 86)

- Among the most widely used algorithms
- Easy
  - to understand
  - to implelement
  - to use
  - and cheap in CPU time
- J48, Weka, SciKit



Decision Trees



Decision Trees (2)

Procedure DecisionTree($\mathcal{E}$)

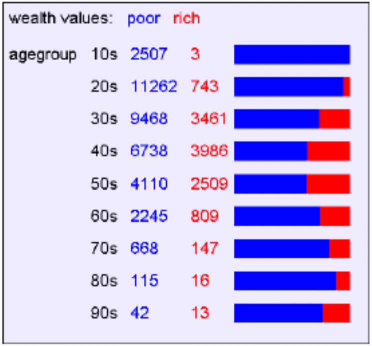
1. Assume $\mathcal{E} = \{(x_i, y_i)_{i=1}^n, \; x_i \in \mathbb{R}^D, \; y_i \in \{0,1\}\}$
   - If $\mathcal{E}$ single-class (i.e., $\forall i, j \in [1, n]; y_i = y_j$), return
   - If n too small (i.e., $<$ threshold), return
   - Else, find the most informative attribute *att*
2. Forall value *val* of *att*
   - Set $\mathcal{E}_{val} = \mathcal{E} \cap [att = val]$.
   - Call DecisionTree($\mathcal{E}_{val}$)

Criterion: information gain

$$
\begin{array}{rcl}
p & = & Pr(Class = 1 | att = val) \\
I([att = val]) & = & -p \, log \, p - (1-p) \, log \, (1-p) \\
I(att) & = & \sum_i Pr(att = val_i).I([att = val_i])
\end{array}
$$

Decision Trees (3)

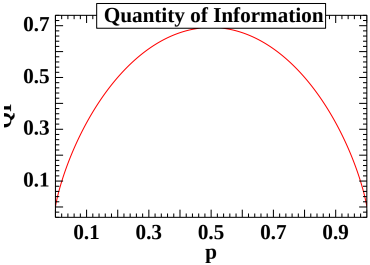
Contingency Table



wealth values: poor rich

		poor	rich
agegroup	10s	2507	3
	20s	11262	743
	30s	9468	3461
	40s	6738	3986
	50s	4110	2509
	60s	2245	809
	70s	668	147
	80s	115	16
	90s	42	13

Quantity of Information (QI)



Computation

value	p(value)	p(poor | value)	QI (value)	p(value) * QI (value)
[0,10[	0.051	0.999	0.00924	0.000474
[10,20[	0.25	0.938	0.232	0.0570323
[20,30[	0.26	0.732	0.581	0.153715

Decision Trees (4)

Limitations

- XOR-like attributes
- Attributes with many values
- Numerical attributes
- Overfitting

Limitations

Numerical Attributes

- Order the values $val_1 < \ldots < val_t$
- Compute $\mathrm{QI}([att < val_i])$
- $\mathrm{QI}(\mathrm{att}) = \max_i \mathrm{QI}([att < val_i])$

The XOR case

Bias the distribution of the examples

Complexity

Quantity of information of an attribute

$$n \ln n$$

Adding a node

$$D \times n \ln n$$

Tackling Overfitting

Penalize the selection of an already used variable

- Limits the tree depth.

Do not split subsets below a given minimal size

- Limits the tree depth.

Pruning

- Each leaf, one conjunction;
- Generalization by pruning litterals;
- Greedy optimization, QI criterion.

Decision Trees, Summary

Still around after all these years

- Robust against noise and irrelevant attributes
- Good results, both in quality and complexity

Random Forests Breiman 00