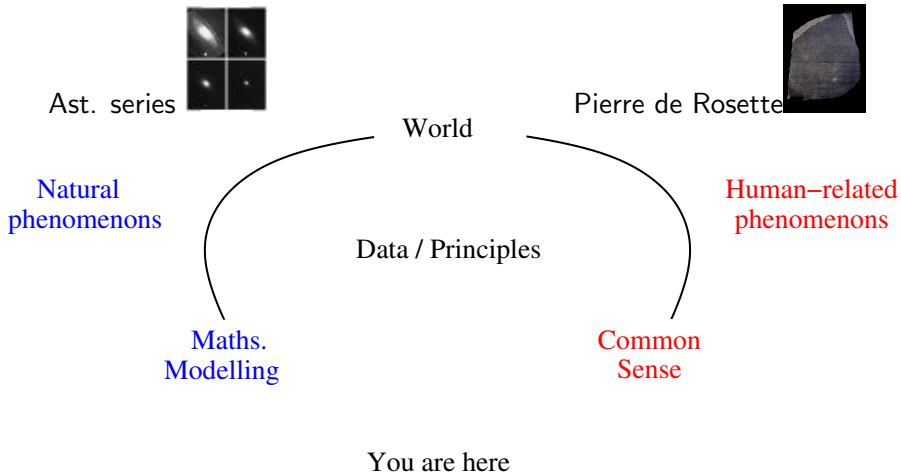


Master Recherche IAC
TC2: Apprentissage Statistique & Optimisation

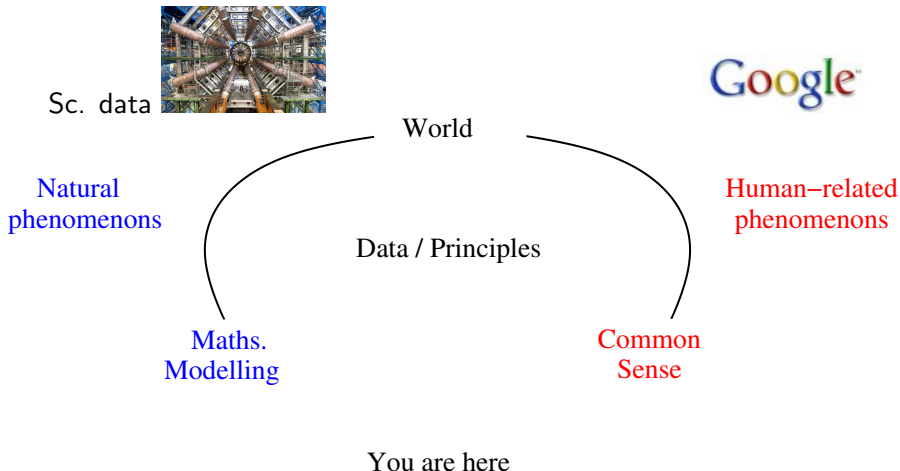
Alexandre Allauzen — Anne Auger — Michèle Sebag
LIMSI — LRI

Sept. 16th, 2013

Where we are



Where we are



Harnessing Big Data



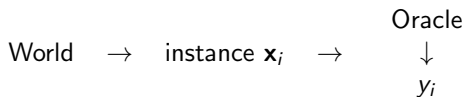
Watson (IBM) defeats human champions at the quiz game Jeopardy (Feb. 11)

<i>i</i>	1	2	3	4	5	6	7	8	
1000'	kilo	mega	giga	tera	peta	exa	zetta	yotta	bytes

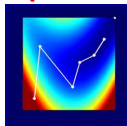
- ▶ Google: 24 petabytes/day
- ▶ Facebook: 10 terabytes/day; Twitter: 7 terabytes/day
- ▶ Large Hadron Collider: 40 terabytes/seconds

Machine Learning and Optimization

Machine Learning



Optimization



ML and Optimization

- ▶ ML is an optimization problem: find the best model
- ▶ Smart optimization requires learning about the optimization landscape

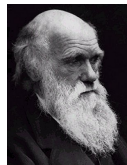
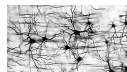
◀ ◻ ▶ ◀ ◻ ▶ ◀ ≡ ▶ ◀ ≡ ▶ ≡ ≡ ≡ ↺ 🔍 ↻

The module

1. Introduction. Decision trees. Validation.
2. Neural Nets
3. Statistics
4. Learning from sequences
5. Unsupervised learning
6. Representation changes
7. Bayesian learning
8. Optimisation



Rev. T. BAYES



- ▶ Slides of this module:
<http://tao.lri.fr/tiki-index.php?page=Courses>
<http://www.limsi.fr/Individu/allauzen/wiki/index.php/>
- ▶ Andrew Ng courses
<http://ai.stanford.edu/~ang/courses.html>
- ▶ PASCAL videos
<http://videlectures.net/pascal/>
- ▶ Tutorials NIPS Neuro Information Processing Systems
<http://nips.cc/Conferences/2006/Media/>
- ▶ About ML/DM
<http://hunch.net/>

Today

1. Part 1. Generalities
2. Part 2. Decision trees
3. Part 3. Validation

Examples

- ▶ Vision
- ▶ Control
- ▶ Netflix
- ▶ Spam
- ▶ Playing Go
- ▶ Google



<http://ai.stanford.edu/~ang/courses.html>

Reading cheques

LeCun et al. 1990

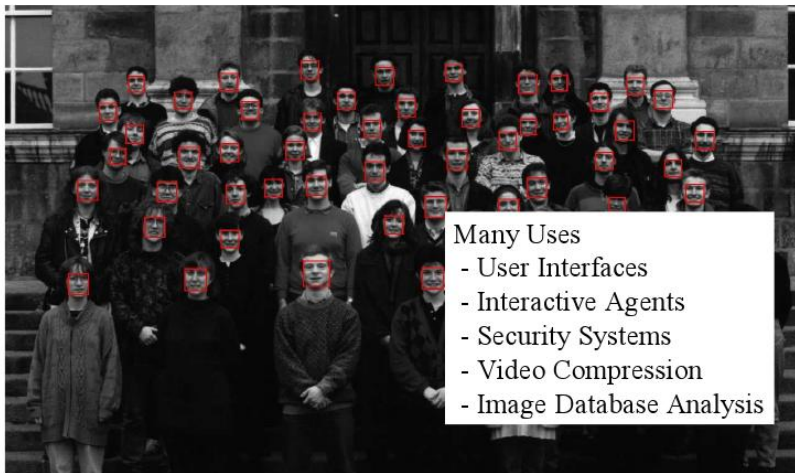


MNIST: The drosophila of ML



Fig. 4. Size-normalized examples from the MNIST database.

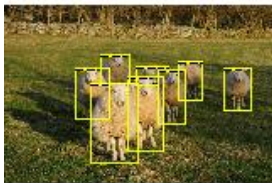
Detecting faces



Viola and Jones, Robust object detection using a boosted cascade of simple features, CVPR 2001

The 2005-2012 Visual Object Challenge

A. Zisserman, C. Williams, M. Everingham, L. v.d. Gool



The supervised learning setting

Input: set of (\mathbf{x}, y)

- ▶ An instance \mathbf{x} e.g. set of pixels, $\mathbf{x} \in \mathbb{R}^D$
- ▶ A label y in $\{1, -1\}$ or $\{1, \dots, K\}$ or \mathbb{R}

The supervised learning setting

Input: set of (\mathbf{x}, y)

- ▶ An instance \mathbf{x} e.g. set of pixels, $\mathbf{x} \in \mathbb{R}^D$
- ▶ A label y in $\{1, -1\}$ or $\{1, \dots, K\}$ or \mathbb{R}

Pattern recognition

- ▶ Classification concept ?

Does the image contain the target

$$h : \{ \text{Images} \} \mapsto \{1, -1\}$$

- ▶ Detection concept?

Does the pixel belong to the img of target

$$h : \{ \text{Pixels in an image} \} \mapsto \{1, -1\}$$

- Segmentation

Find contours of all instances of target concept in image

The 2005 Darpa Challenge

Thrun, Burgard and Fox 2005



Autonomous vehicle Stanley – Terrains

The Darpa challenge and the AI ag



What remains to be done

- ▶ Reasoning
- ▶ Dialogue
- ▶ Perception

Thrun 2005

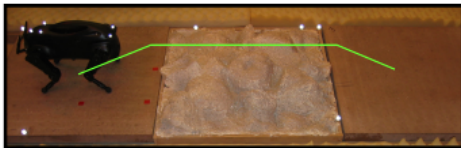
10%

60%

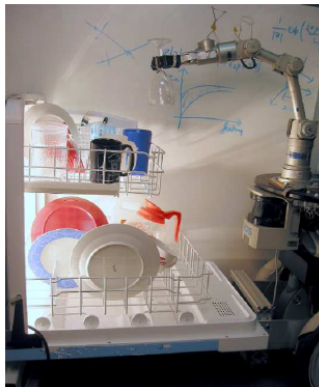
90%

Robots

Ng, Russell, Veloso, Abbeel, Peters, Schaal, ...



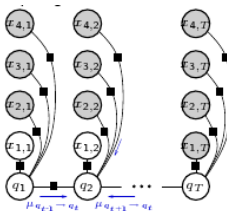
Reinforcement learning



Classification

Robots, 2

Toussaint et al. 2010



(a) Factor graph modelling the variable interactions

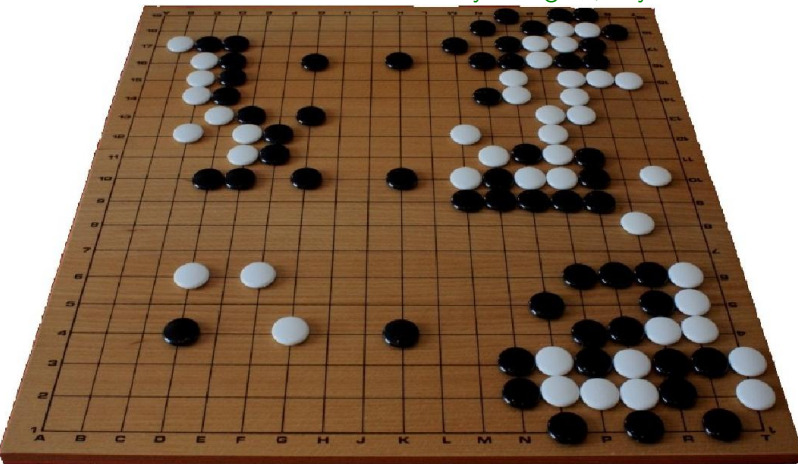


(b) Behaviour of the 39-DOF Humanoid:
Reaching goal under Balance and Collision constraints

Bayesian Inference for Motion Control and Planning

Go as AI Challenge

Gelly Wang 07; Teytaud et al. 2008-2011



Reinforcement Learning, Monte-Carlo Tree Search

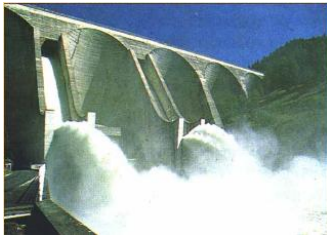
Energy policy

Claim

Many problems can be phrased as optimization in front of the uncertainty.

Adversarial setting	2 two-player game
uniform setting	a single player game

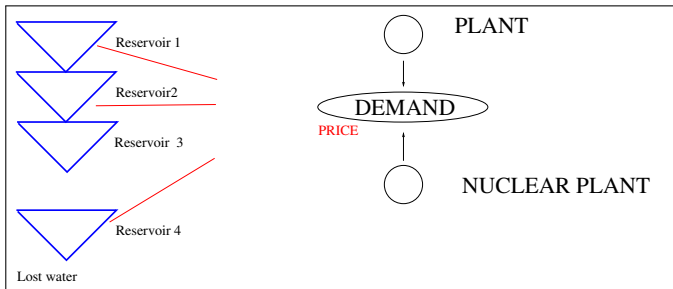
Management of energy stocks under uncertainty



States and Decisions

States

- ▶ Amount of stock (60 nuclear, 20 hydro.)
- ▶ Varying: price, weather alea or archive
- ▶ Decision: release water from one reservoir to another
- ▶ Assessment: meet the demand, otherwise buy energy



Netflix Challenge 2007-2008



NETFLIX

The **best** way to rent movies.

Plans start at
only \$9⁹⁹ a month!

A couple is sitting on a couch, smiling and watching a movie. In front of them is a coffee table with a bowl of popcorn, a glass of red wine, and several Netflix DVD cases. The scene is dimly lit, suggesting a cozy movie night.

Collaborative Filtering

Collaborative filtering

Input

- ▶ A set of users n_u , ca 500,000
- ▶ A set of movies n_m , ca 18,000
- ▶ A $n_m \times n_u$ matrix: person, movie, rating
Very sparse matrix: less than 1% filled...

Output

- ▶ Filling the matrix !

Collaborative filtering

Input

- ▶ A set of users n_u , ca 500,000
- ▶ A set of movies n_m , ca 18,000
- ▶ A $n_m \times n_u$ matrix: person, movie, rating
Very sparse matrix: less than 1% filled...

Output

- ▶ Filling the matrix !

Criterion

- ▶ (relative) mean square error
- ▶ ranking error

Spam – Phishing – Scam



Classification, Outlier detection



The power of big data

- ▶ Now-casting
- ▶ Public relations >> Advertizing

outbreak of flu

Mc Luhan and Google

We shape our tools and afterwards our tools shape us

Marshall McLuhan, 1964

First time ever a tool is observed to modify human cognition that fast.

Sparrow et al., Science 2011

Types of application

Domain

But : Modelling

Physical phenomena

analysis & control

manufacturing, experimental sciences, numerical engineering

Vision, speech, robotics..

Social phenomena

+ privacy

Health, Insurance, Banks ...

Individual phenomena

+ dynamics

Consumer Relationship Management, User Modelling

Social networks, games...

PASCAL : <http://pascallin2.ecs.soton.ac.uk/>

Banks, Telecom, CRN

Ex: KDD 2009 – Orange

1. Churn
2. Appetency
3. Up-selling

Objectives

1. Ads. efficiency
2. Less fraud



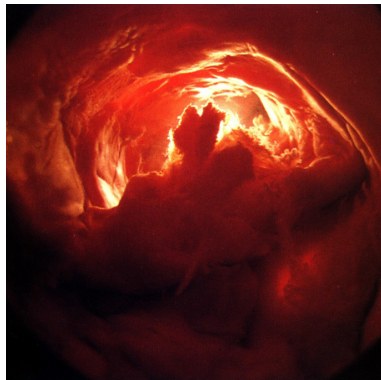
Health, bio-informatics

Ex: Risk factors

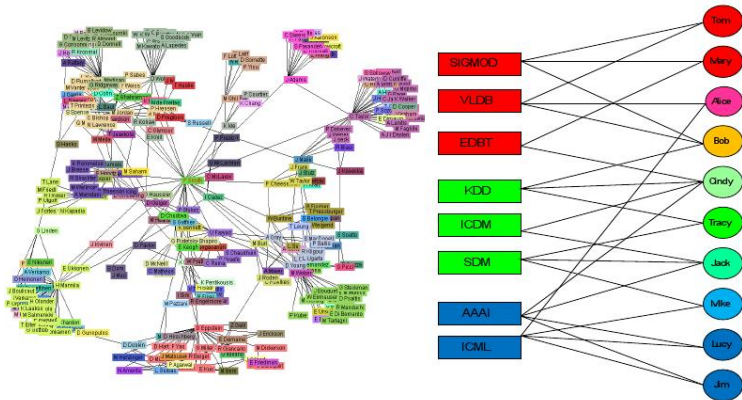
1. Cardio-vascular diseases
2. Carcinogenic Molecules
3. Obesity genes ...

Objectives

1. Diagnostic
2. Personalized care
3. Identification



Scientific Social Network

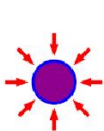
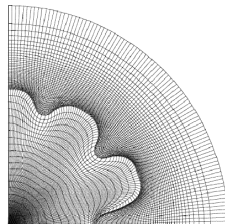


Questions

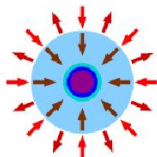
1. Who does what ?
2. Good conferences ?
3. Hot/emerging topics ?
4. Is Mr Q. Lee same as Mr Quoc N. Lee ?

Numerical Engineering

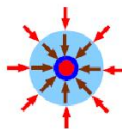
- ▶ Codes
- ▶ Computationally heavy
- ▶ Expertise demanding



Laser heating



DT compression



Hot spot ignition



Thermonuclear burn

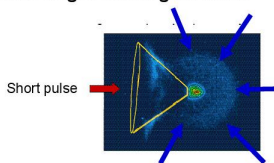
Fusion based on inertial confinement, ICF

Objectives

- ▶ Approximate answer
- ▶ .. in tenth of seconds
- ▶ Speed up the design cycle
- ▶ Optimal design

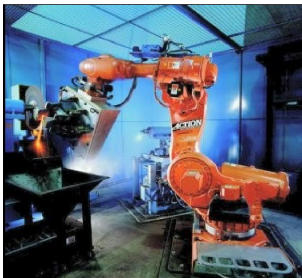
More is Different

Alternative scheme : spherical target with a gold cone*



* Kodama et al. Nature **412** 798 (2001); **418** 933 (2002);

Autonomous robotics

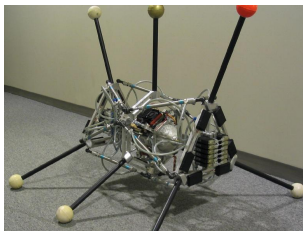


Complexe, monde fermé

Design



simple, random



Autonomous robotics, 2

Reality Gap

- ▶ Design in silico (simulator)
- ▶ Run the controller on the robot (in vivo)

Autonomous robotics, 2

Reality Gap

- ▶ Design in silico (simulator)
- ▶ Run the controller on the robot (in vivo)
- ▶ Does not work !

Closing the reality Gap

1. Simulator-based design
2. On-board trials
3. Log the data, update the simulator
4. Goto 1

safe environnement

Active learning

Co-evolution
[tr. Hod Lipson, 2010]

Overview

Examples

Introduction to Supervised Machine Learning

Decision trees

Empirical validation

- Performance indicators

- Estimating an indicator

Types of Machine Learning problem

WORLD — DATA — USER

Observations

+ **Target**

+ Rewards

Understand Code

Predict Classification/Regression

Decide Policy

Unsupervised LEARNING

Supervised LEARNING

Reinforcement LEARNING

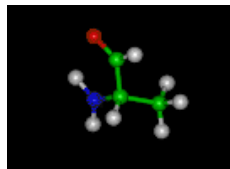
Example

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

age	employe	education	edur	marital	...	job	relation	race	gender	hour	country	wealth
39	State_gov	Bachelors	13	Never_mar	...	Adm_clerk	Not_in_fan	White	Male	40	United_States	poor
51	Self_employed	Bachelors	13	Married	...	Exec_manager	Husband	White	Male	13	United_States	poor
39	Private	HS_grad	9	Divorced	...	Handlers_cleaner	Not_in_fan	White	Male	40	United_States	poor
54	Private	11th	7	Married	...	Handlers_cleaner	Husband	Black	Male	40	United_States	poor
28	Private	Bachelors	13	Married	...	Prof_spec	Wife	Black	Female	40	Cuba	poor
38	Private	Masters	14	Married	...	Exec_manager	Wife	White	Female	40	United_States	poor
50	Private	9th	5	Married_spl	...	Other_serv	Not_in_fan	Black	Female	16	Jamaica	poor
52	Self_employed	HS_grad	9	Married	...	Exec_manager	Husband	White	Male	45	United_States	rich
31	Private	Masters	14	Never_mar	...	Prof_spec	Not_in_fan	White	Female	50	United_States	rich
42	Private	Bachelors	13	Married	...	Exec_manager	Husband	White	Male	40	United_States	rich
37	Private	Some_col	10	Married	...	Exec_manager	Husband	Black	Male	80	United_States	rich
30	State_gov	Bachelors	13	Married	...	Prof_spec	Husband	Asian	Male	40	India	rich
24	Private	Bachelors	13	Never_mar	...	Adm_clerk	Own_child	White	Female	30	United_States	poor
33	Private	Assoc_acc	12	Never_mar	...	Sales	Not_in_fan	Black	Male	50	United_States	poor
41	Private	Assoc_voc	11	Married	...	Craft_repair	Husband	Asian	Male	40	MissingV	rich
34	Private	7th_8th	4	Married	...	Transport	Own_child	Amer_Indi	Male	45	Mexico	poor
26	Self_employed	HS_grad	9	Never_mar	...	Farming_fish	Own_child	White	Male	35	United_States	poor
33	Private	HS_grad	9	Never_mar	...	Machine_op	Unmarried	White	Male	40	United_States	poor
38	Private	11th	7	Married	...	Sales	Husband	White	Male	50	United_States	poor
44	Self_employed	Masters	14	Divorced	...	Exec_manager	Unmarried	White	Female	45	United_States	rich
41	Private	Doctorate	16	Married	...	Prof_spec	Husband	White	Male	60	United_States	rich
:	:	:	:	:	:	:	:	:	:	:	:	:

Instance space \mathcal{X}

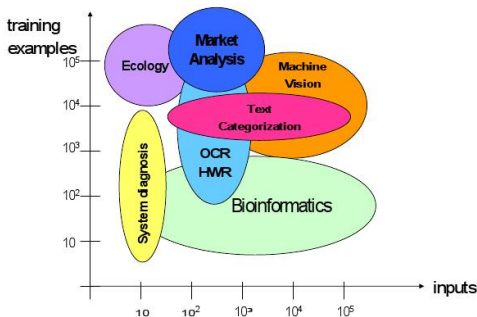
- ▶ Propositionnal :
 $\mathcal{X} \equiv \mathbb{R}^d$
- ▶ Structured :
sequential,
spatio-temporal,
relational.



aminoacid

Data / Applications

- ▶ Propositional data 80% des applis.
- ▶ Spatio-temporal data alarms, mines, accidents
- ▶ Relational 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
- Structured data: spatio-temporal, relational, text, videos,..

Feature extraction

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
- ↘ Complexity : n , $n \log n$, n^2
- Combinatorial optimization

Scalability

H. Simon, 1958:

In complex real-world situations, optimization becomes approximate optimization since the description of the real-world is radically simplified until reduced to a degree of complication that the decision maker can handle.

Satisficing seeks simplification in a somewhat different direction, retaining more of the detail of the real-world situation, but settling for a satisfactory, rather than approximate-best, decision.

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

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

Disciplines et critères

- ▶ 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 Learning, notations

Context

World \rightarrow Instance $\mathbf{x}_i \rightarrow$ Oracle
 \downarrow
 y_i



INPUT

$\sim P(\mathbf{x}, y)$

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

HYPOTHESIS SPACE

$$\mathcal{H} \quad h : \mathcal{X} \mapsto \mathcal{Y}$$

LOSS FUNCTION

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

OUTPUT

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

Classification and criteria

Supervised learning

- ▶ $\mathcal{Y} = \text{True/False}$ classification
- ▶ $\mathcal{Y} = \{1, \dots, k\}$ multi-class discrimination
- ▶ $\mathcal{Y} = \mathbb{R}$ regression

Generalization Error

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

Empirical Error

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

Bound

structural risk

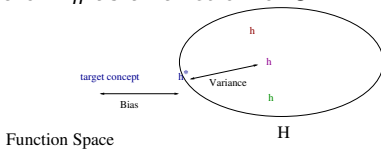
$$Err(h) < Err_e(h) + \mathcal{F}(n, d(\mathcal{H}))$$

$d(\mathcal{H}) = \text{Vapnik Cervonenkis dimension of } \mathcal{H}$, see later

The Bias-Variance Trade-off

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

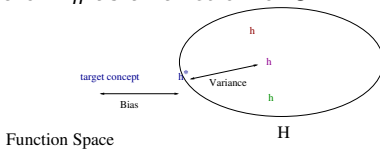
Variance Variance of h_n as a function of \mathcal{E}



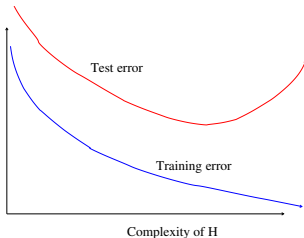
The Bias-Variance Trade-off

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

Variance Variance of h_n as a function of \mathcal{E}



Overfitting



Key notions

- ▶ The main issue regarding supervised learning is overfitting.
- ▶ How to tackle overfitting:
 - ▶ Before learning: use a sound criterion
 - ▶ After learning: cross-validation

regularization
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

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 $h(x) = \text{True}$.
- ▶ \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

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

A set of small navigation icons typically found in Beamer presentations, including symbols for back, forward, search, and other slide controls.

Overview

Examples

Introduction to Supervised Machine Learning

Decision trees

Empirical validation

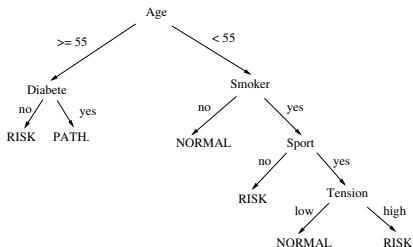
Performance indicators

Estimating an indicator

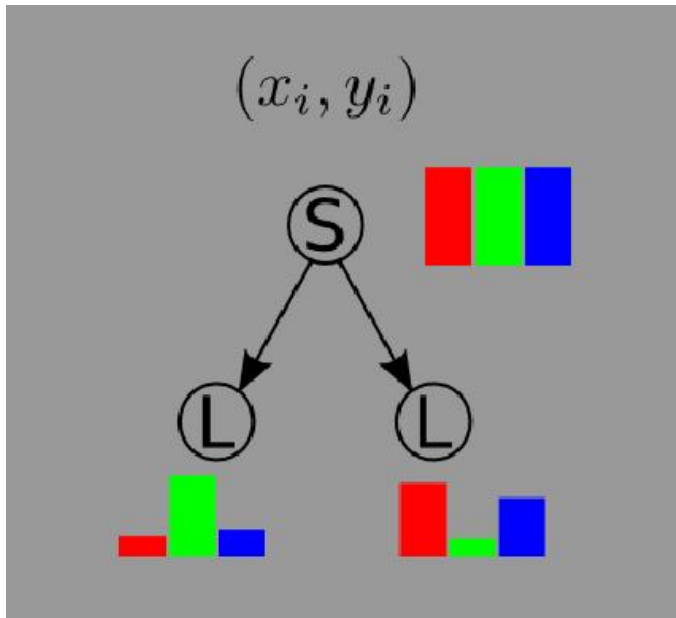
Decision Trees

C4.5 (Quinlan 86)

- ▶ Among the most widely used algorithms
- ▶ Easy
 - ▶ to understand
 - ▶ to implement
 - ▶ 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., $< \text{threshold}$), return
 - Else, find the most informative attribute att
2. For all value val of att
 - Set $\mathcal{E}_{val} = \mathcal{E} \cap [att = val]$.
 - Call DecisionTree(\mathcal{E}_{val})

Criterion: information gain

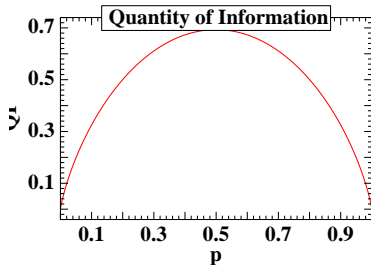
$$\begin{aligned}
 p &= Pr(Class = 1 | att = val) \\
 I([att = val]) &= -p \log p - (1 - p) \log (1 - p) \\
 I(att) &= \sum_i Pr(att = val_i) \cdot I([att = val_i])
 \end{aligned}$$

Decision Trees (3)

Contingency Table

wealth values:		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

Numerical Attributes

- ▶ Order the values $val_1 < \dots < val_t$
- ▶ Compute $QI([att < val_i])$
- ▶ $QI(att) = \max_i QI([att < val_i])$

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 literals;
- ▶ 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

Overview

Examples

Introduction to Supervised Machine Learning

Decision trees

Empirical validation

- Performance indicators

- Estimating an indicator

Validation issues

1. What is the result ?
2. My results look good. Are they ?
3. Does my system outperform yours ?
4. How to set up my system ?

Validation: Three questions

Define a good indicator of quality

- ▶ Misclassification cost
- ▶ Area under the ROC curve

Computing an estimate thereof

- ▶ Validation set
- ▶ Cross-Validation
- ▶ Leave one out
- ▶ Bootstrap

Compare estimates: Tests and confidence levels

Which indicator, which estimate: de

Settings

- ▶ Large/few data

Data distribution

- ▶ Dependent/independent examples
- ▶ balanced/imbalanced classes

Overview

Examples

Introduction to Supervised Machine Learning

Decision trees

Empirical validation

- Performance indicators

- Estimating an indicator

Performance indicators

Binary class

- ▶ h^* the truth
- ▶ \hat{h} the learned hypothesis

Confusion matrix

\hat{h} / h^*	1	0	
1	a	b	a+b
0	c	d	c+d
	a+c	b+d	a + b + c + d

Performance indicators, 2

\hat{h} / h^*	1	0	
1	a	b	a+b
0	c	d	c+d
	a+c	b+d	a + b + c + d

- ▶ Misclassification rate $\frac{b+c}{a+b+c+d}$
- ▶ Sensitivity (recall), True positive rate (TP) $\frac{a}{a+c}$
- ▶ Specificity, False negative rate (FN) $\frac{b}{b+d}$
- ▶ Precision $\frac{a}{a+b}$

Note: always compare to random guessing / baseline alg.

The Area under the ROC curve

- ▶ ROC: Receiver Operating Characteristics
- ▶ Origin: Signal Processing, Medicine

Principle

$h : X \mapsto \mathbb{R}$ $h(x)$ measures the risk of patient x

h leads to order the examples:

+++ - + - + + + + - - - + - - - - - - - - - - - - - - - -

The Area under the ROC curve

- ▶ ROC: Receiver Operating Characteristics
- ▶ Origin: Signal Processing, Medicine

Principle

$h : X \mapsto \mathbb{R}$ $h(x)$ measures the risk of patient x

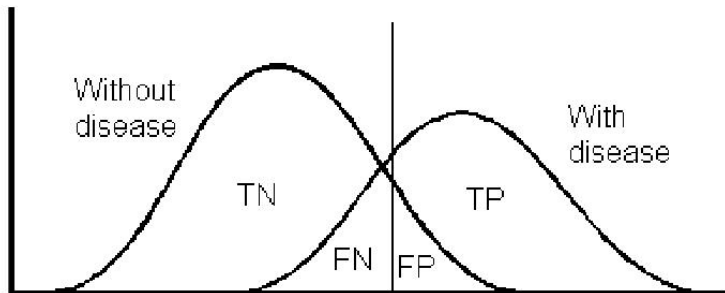
h leads to order the examples:

+++ - + - + + + + - - - + - - - - - - - - - - - - - - - -

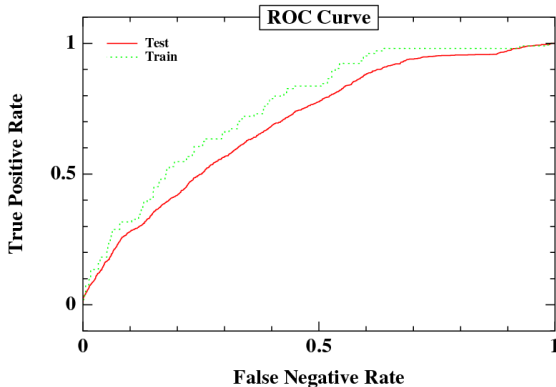
Given a threshold θ , h yields a classifier: Yes iff $h(x) > \theta$.

+++ - + - + + + + | - - - + - - - + - - - - - - - - - - - - - - - -

Here, TP $(\theta) = .8$; FN $(\theta) = .1$



The ROC curve



Ideal classifier: (0 False negative, 1 True positive)

Diagonal (True Positive = False negative) \equiv nothing learned.

ROC Curve, Properties

Properties

ROC depicts the trade-off True Positive / False Negative.

Standard: misclassification cost (Domingos, KDD 99)

$$\text{Error} = \# \text{ false positive} + c \times \# \text{ false negative}$$

In a multi-objective perspective, ROC = Pareto front.

Best solution: intersection of Pareto front with $\Delta(-c, -1)$

ROC Curve, Properties, foll'd

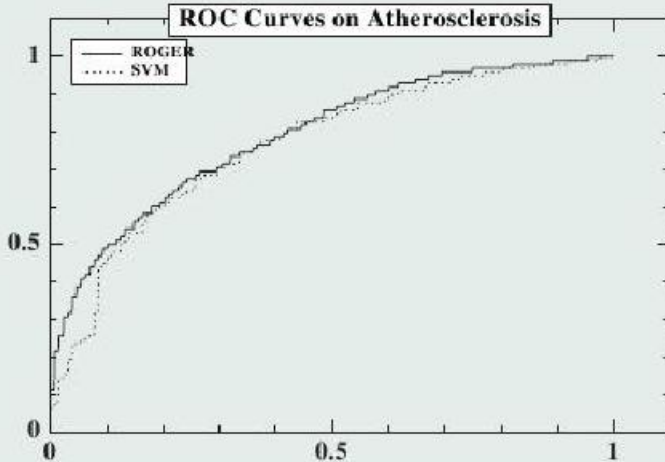
Used to compare learners

Bradley 97

multi-objective-like

insensitive to imbalanced distributions

shows sensitivity to error cost.



Area Under the ROC Curve

Often used to select a learner

Don't ever do this !

Hand, 09

Sometimes used as learning criterion

Mann Whitney

Wilcoxon

$$AUC = Pr(h(x) > h(x') | y > y')$$

WHY

Rosset, 04

- ▶ More stable $\mathcal{O}(n^2)$ vs $\mathcal{O}(n)$
- ▶ With a probabilistic interpretation

Clemençon et al. 08

HOW

- ▶ SVM-Ranking
- ▶ Stochastic optimization

Joachims 05; Usunier et al. 08, 09

Overview

Examples

Introduction to Supervised Machine Learning

Decision trees

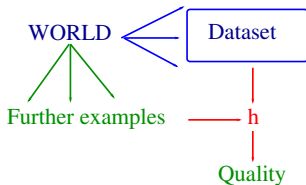
Empirical validation

- Performance indicators

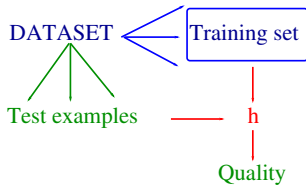
- Estimating an indicator

Validation, principle

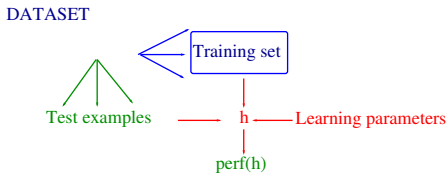
Desired: performance on further instances



Assumption: Dataset is to World, like Training set is to Dataset.



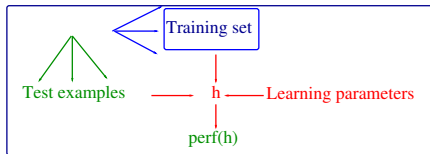
Validation, 2



Unbiased Assessment of Learning Algorithms
T. Scheffer and R. Herbrich, 97

Validation, 2

DATASET

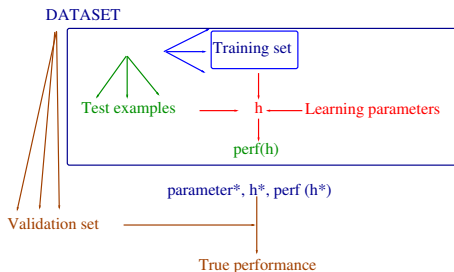


parameter*, h^* , $\text{perf}(h^*)$

Unbiased Assessment of Learning Algorithms

T. Scheffer and R. Herbrich, 97

Validation, 2



Unbiased Assessment of Learning Algorithms
T. Scheffer and R. Herbrich, 97

Overview

Examples

Introduction to Supervised Machine Learning

Decision trees

Empirical validation

- Performance indicators

- Estimating an indicator

Confidence intervals

Definition

Given a random variable X on \mathbb{R} , a $p\%$ -confidence interval is $I \subset \mathbb{R}$ such that

$$Pr(X \in I) > p$$

Binary variable with probability ϵ

Probability of r events out of n trials:

$$P_n(r) = \frac{n!}{r!(n-r)!} \epsilon^r (1-\epsilon)^{n-r}$$

- ▶ Mean: $n\epsilon$
- ▶ Variance: $\sigma^2 = n\epsilon(1-\epsilon)$

Gaussian approximation

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{-\frac{1}{2} \frac{x-\mu}{\sigma}^2}$$

Confidence intervals

Bounds on (true value, empirical value) for n trials, $n > 30$

$$Pr(|\hat{x}_n - x^*| > \underset{z}{1.96} \sqrt{\frac{\hat{x}_n \cdot (1 - \hat{x}_n)}{n}}) < \underset{\epsilon}{.05}$$

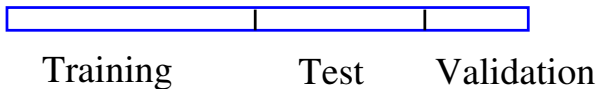
Table

z	.67	1.	1.28	1.64	1.96	2.33	2.58
ϵ	50	32	20	10	5	2	1

Empirical estimates

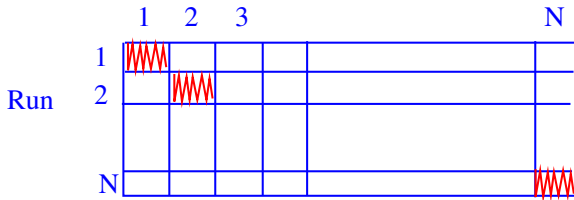
When data abound

(MNIST)





Cross validation

Fold



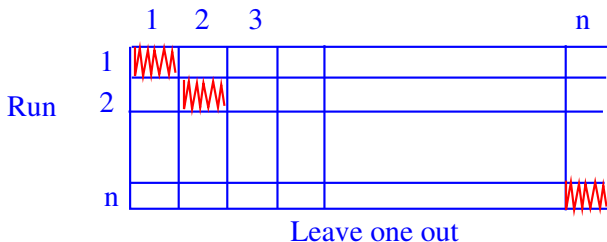
N-fold Cross Validation

Error = Average (error on  of h
learned from )

Empirical estimates, foll'd

Cross validation → Leave one out

Fold



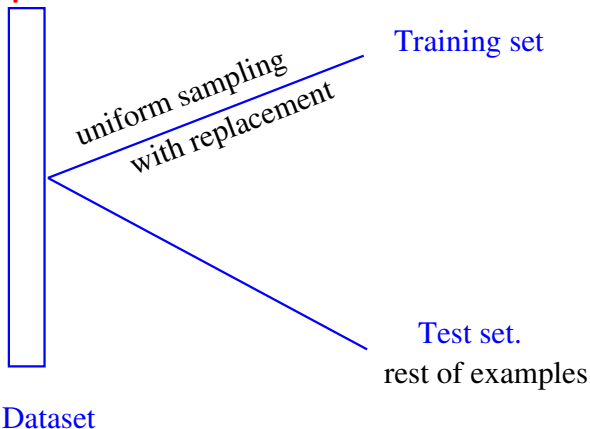
Same as N-fold CV, with $N = \text{number of examples}$.

Properties

Low bias; high variance; underestimate error if data not independent

Empirical estimates, foll'd

Bootstrap



Average indicator over all (Training set, Test set) samplings.

Multiple hypothesis testing

- ▶ If you test many hypotheses on the same dataset
- ▶ one of them will appear confidently true...

More

- ▶ Tutorial slides:
http://www.lri.fr/~sebag/Slides/Validation_Tutorial_11.pdf
- ▶ Video and slides (soon): ICML 2012, Videolectures, Tutorial Japkowicz & Shah
<http://www.mohakshah.com/tutorials/icml2012/>

Validation, summary

What is the performance criterion

- ▶ Cost function
- ▶ Account for class imbalance
- ▶ Account for data correlations

Assessing a result

- ▶ Compute confidence intervals
- ▶ Consider baselines
- ▶ Use a validation set

If the result looks too good, don't believe it