M1 – Apprentissage

Michèle Sebag – Benoit Barbot
LRI – LSV

30 septembre 2013

Overview

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

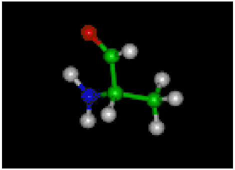
Example

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

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
:	:	:	:	:	:	:	:	:	:	:	:	:

Instance space $\mathcal{X}$

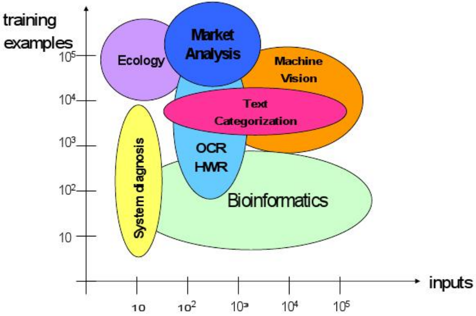
- Propositionnal : $\mathcal{X} \equiv \mathbb{R}^d$
- Structured : sequential, spatio-temporal, relational.



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, $n \log n$, n^2 **Scalability**
- $-$ Combinatorial optimization

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

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

Context

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

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

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

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

HYPOTHESIS SPACE

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

LOSS FUNCTION

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

OUTPUT

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

Classification and criteria

Supervised learning

- $\mathcal{Y}$ = True/False — classification
- $\mathcal{Y} = \{1, \ldots k\}$ — 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

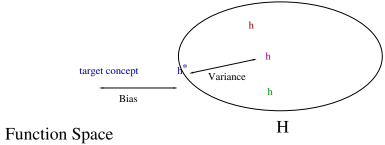
$$Err(h) < Err_e(h) + \mathcal{F}(n, d(\mathcal{H}))$$

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

The Bias-Variance Trade-off

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

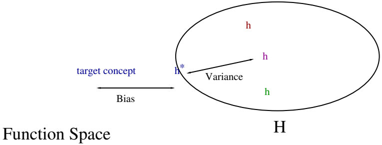
Variance Variance of h_n as a function of $\mathcal{E}$



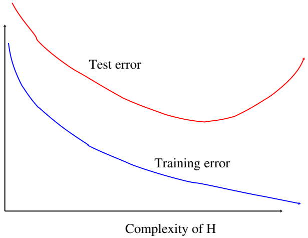
The Bias-Variance Trade-off

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 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

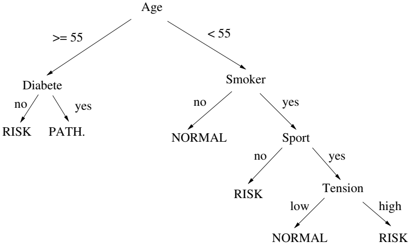
	$\mathcal{H}$	navigation operators
Version Space	Logical	spec / gen
Decision Trees	Logical	specialisation
Neural Networks	Numerical	gradient
Support Vector Machines	Numerical	quadratic opt.
Ensemble Methods	—	adaptation $\mathcal{E}$

Overview

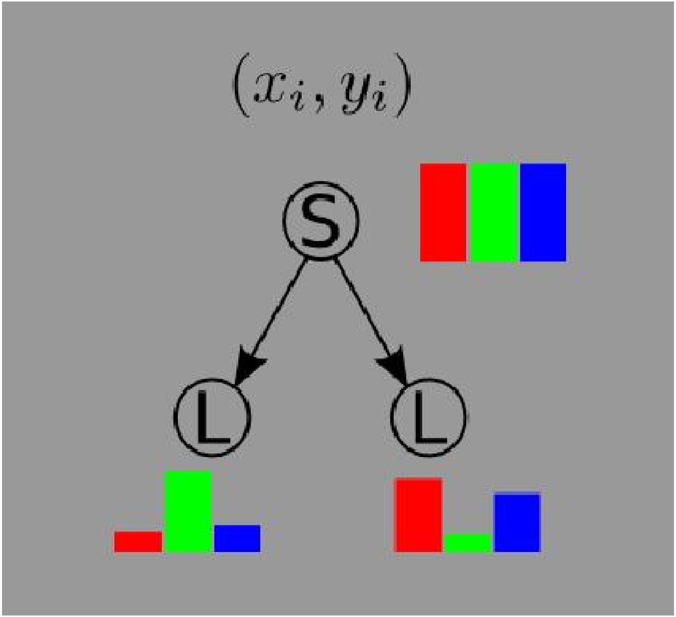
Decision Trees

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($\mathcal{E}$)

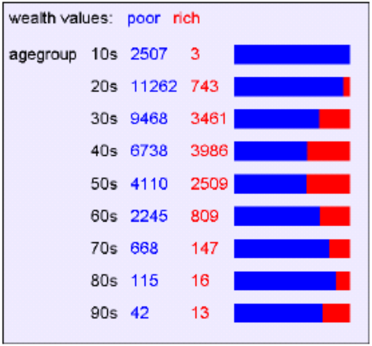
1. Assume $\mathcal{E} = \{(x_i, y_i)_{i=1}^n, \ x_i \in \mathbb{R}^D, \ y_i \in \{0,1\}\}$
   - If $\mathcal{E}$ single-class (i.e., $\forall i, j \in [1, n]; y_i = y_j$), return
   - If n too small (i.e., $<$ threshold), return
   - Else, find the most informative attribute *att*
2. Forall value *val*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

Criterion: information gain

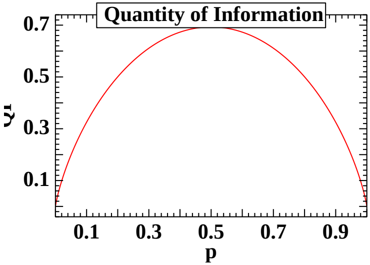
$$\begin{aligned}
p &= Pr(Class = 1 | att = val) \\
I([att = val]) &= -p \, log \, p - (1-p) \, log \, (1-p) \\
I(att) &= \textstyle\sum_i Pr(att = val_i).I([att = val_i])
\end{aligned}$$

Decision Trees (3)

Contingency Table



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

Overview

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: depends.

Settings

- Large/few data

Data distribution

- Dependent/independent examples
- balanced/imbalanced classes

Overview

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.

Performance indicators, 3

The Area under the ROC curve

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

Principle

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

h leads to order the examples:

$$+ + + - + - + + + + - - - + - - - + - - - - - - - - - - - - -$$

Performance indicators, 3

The Area under the ROC curve

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

Principle

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

h leads to order the examples:

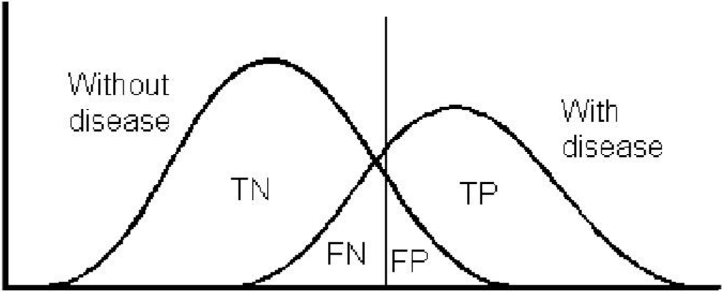
$$+++-+-++++---+---+-------------$$

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

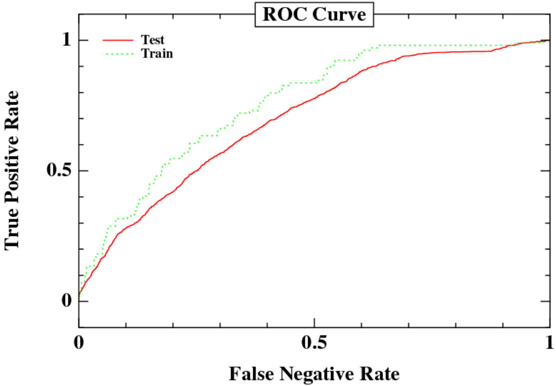
$$+++-+-++++ \;|\; ---+---+-------------$$

Here, TP (θ)= .8; FN (θ) = .1

ROC



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.

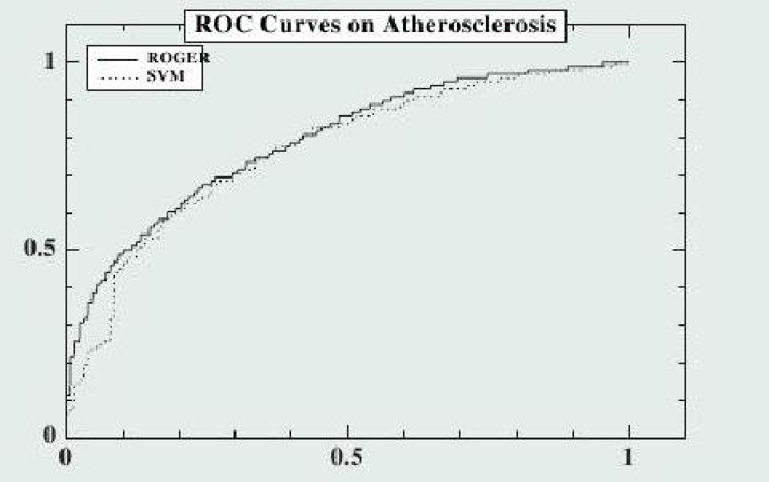
Best solution: intersection of Pareto front with $\Delta(-c, -1)$

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

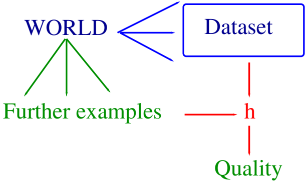
HOW

- SVM-Ranking Joachims 05; Usunier et al. 08, 09
- Stochastic optimization

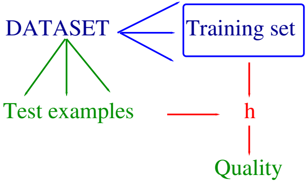
Overview

Validation, principle

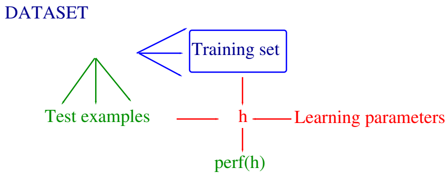
Desired: performance on further instances



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

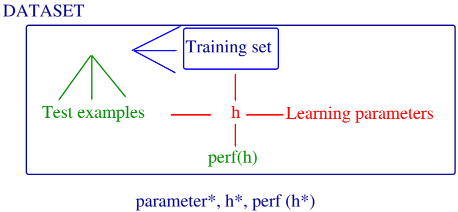


Validation, 2



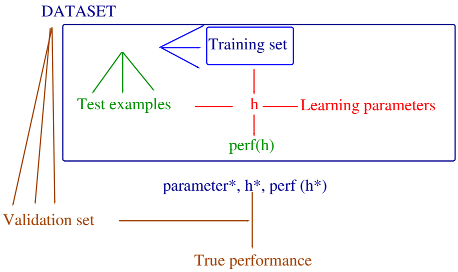
Unbiased Assessment of Learning Algorithms
T. Scheffer and R. Herbrich, 97

Validation, 2



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

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, $n > 30$

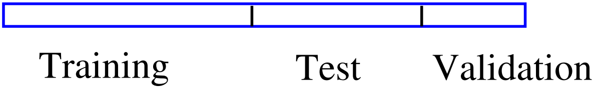
$$Pr(|\hat{x}_n - x^*| > \underset{z}{1.96} \sqrt{\frac{\hat{x}_n.(1-\hat{x}_n)}{n}}) < \underset{\varepsilon}{.05}$$

Table

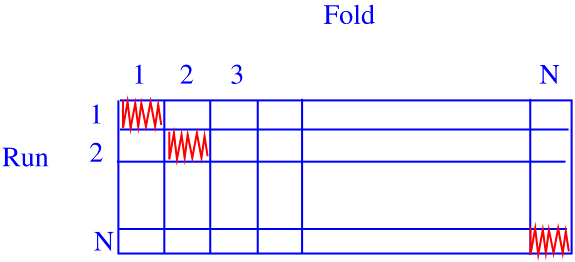
z	.67	1.	1.28	1.64	1.96	2.33	2.58
ε	50	32	20	10	5	2	1

Empirical estimates

When data abound (MNIST)



Cross validation



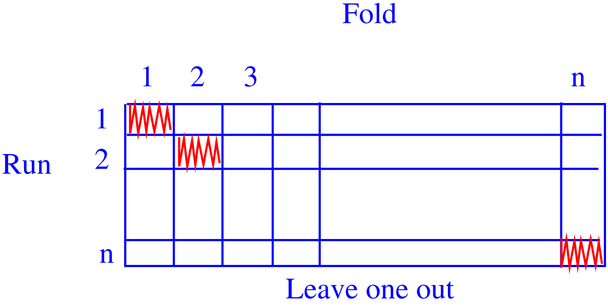
N−fold Cross Validation

Error = Average (error on of h

learned from)

Empirical estimates, foll'd

Cross validation → Leave one out



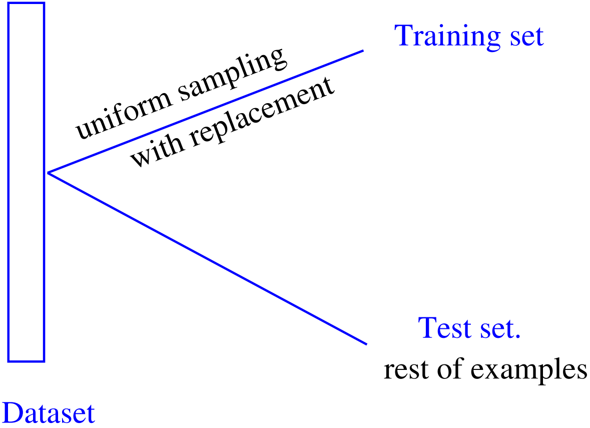
Same as N-fold CV, with N = number of examples.

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.

Beware

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