M1 – Apprentissage

Michèle Sebag – Benoit Barbot
LRI – LSV

14 octobre 2013

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

Overview

Performance indicators

Measuring a performance indicator

Scalable validation: Bags of little bootstrap

Which indicator, which estimate: depends.

Settings

- Large/few data

Data distribution

- Dependent/independent examples
- balanced/imbalanced classes

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

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

h leads to order the examples:

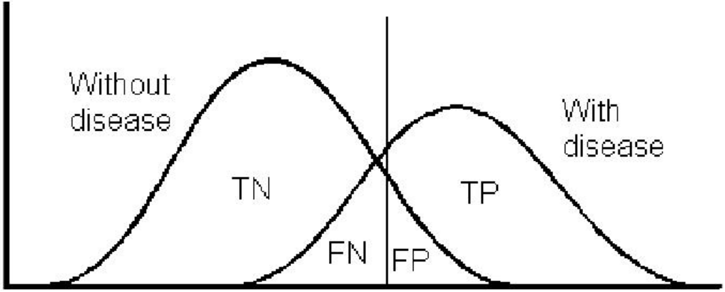
$$+++-+-++++---+---+-------------$$

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

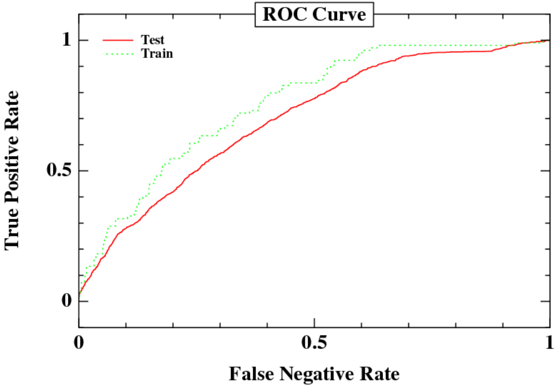
$$+++-+-++++ \mid ---+---+-------------$$

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

ROC



The ROC curve



Ideal classifier: (0 False negative,1 True positive)
Diagonal (True Positive = False negative) ≡ 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

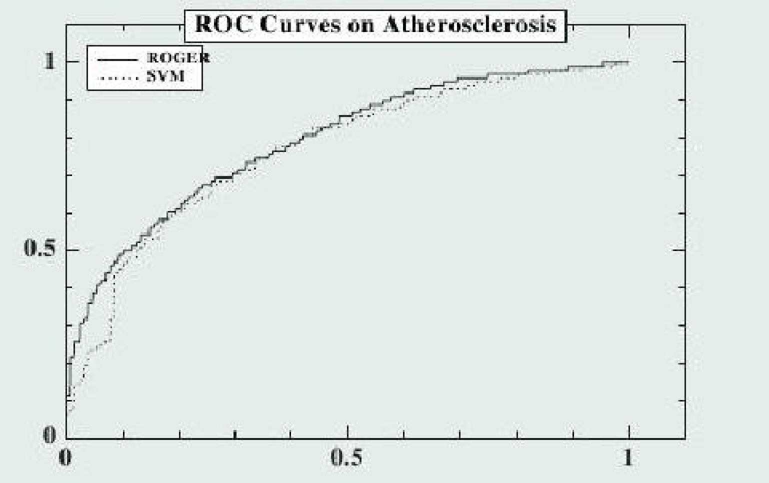
Used to compare learners

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

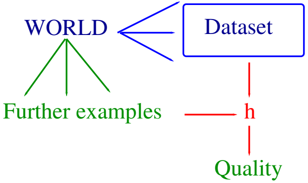
Performance indicators

Measuring a performance indicator

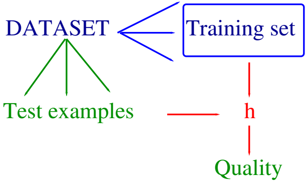
Scalable validation: Bags of little bootstrap

Validation, principle

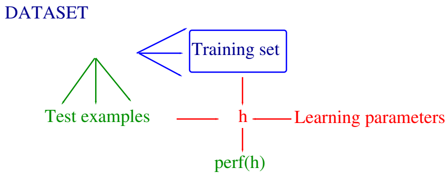
Desired: performance on further instances



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

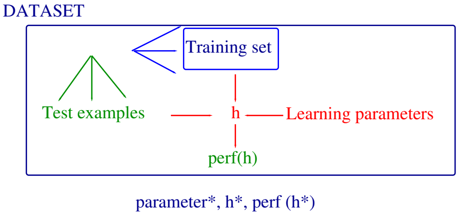


Validation, 2



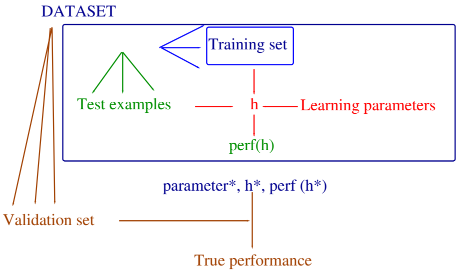
Unbiased Assessment of Learning Algorithms
T. Scheffer and R. Herbrich, 97

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

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, $n > 30$

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

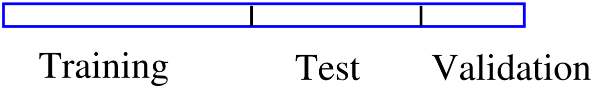
Table

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

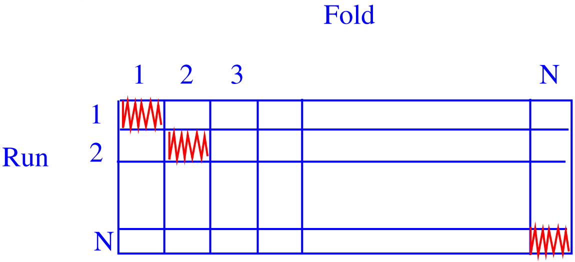
Empirical estimates

When data abound

(MNIST)



Cross validation



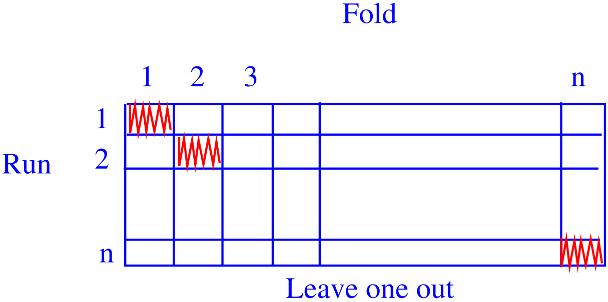
N–fold Cross Validation

Error = Average (error on of h

learned from)

Empirical estimates, foll'd

Cross validation → Leave one out



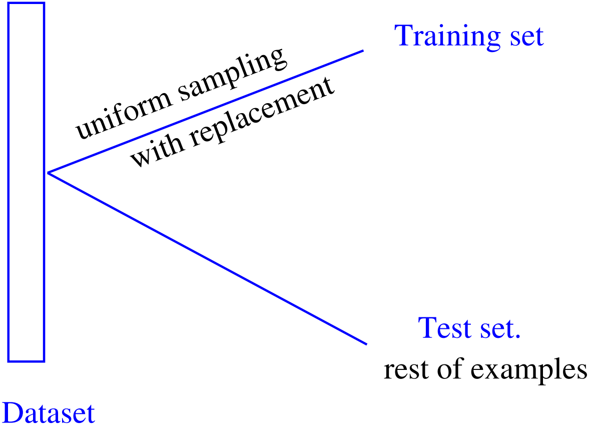
Same as N-fold CV, with N = number of examples.

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

Overview

Performance indicators

Measuring a performance indicator

Scalable validation: Bags of little bootstrap