L3 Apprentissage

Michèle Sebag – Benjamin Monmège
LRI – LSV

13 février 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

Which indicator, which estimate: depends.

Settings

- Large/few data

Data distribution

- Dependent/independent examples
- balanced/imbalanced classes

Overview

Performance indicators

Estimating an indicator

Testing

Hyper-parameter tuning

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:

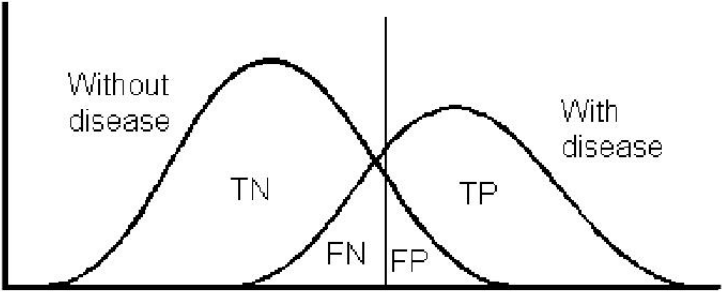
$$+++-+-++++---+---+-------------$$

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

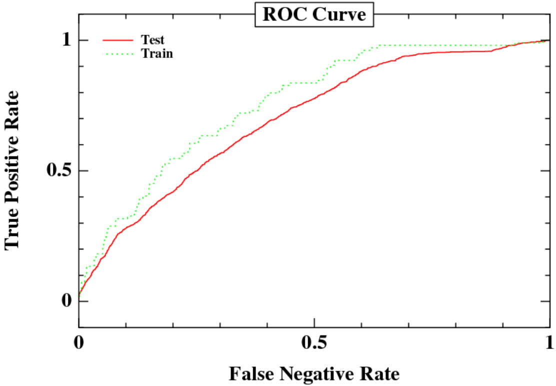
$$+++-+-++++ \mid ---+---+-------------$$

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

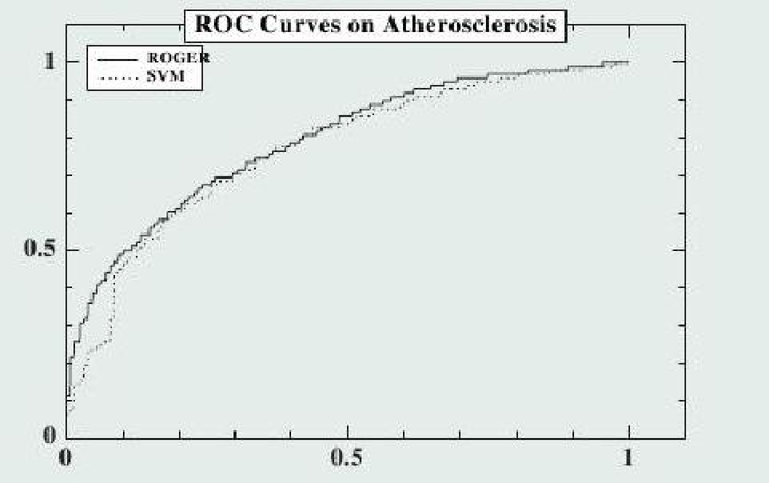
Used to compare learners

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

$$AUC = Pr(h(x) < h(x')|y = 1,\ y' = -1)$$

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

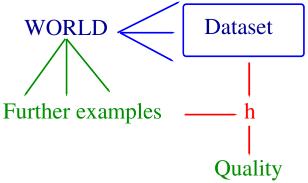
Estimating an indicator

Testing

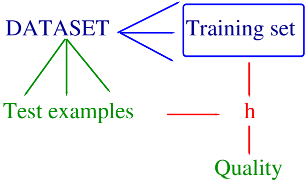
Hyper-parameter tuning

Validation, principle

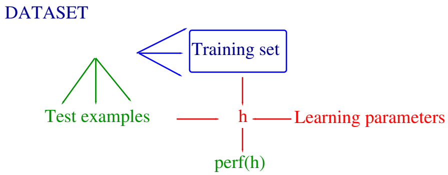
Desired: performance on further instances



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

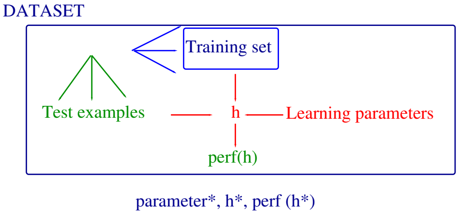


Validation, 2



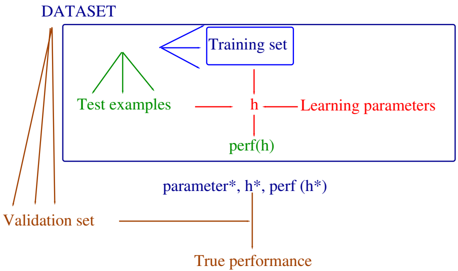
Unbiased Assessment of Learning Algorithms
T. Scheffer and R. Herbrich, 97

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, $n > 30$

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

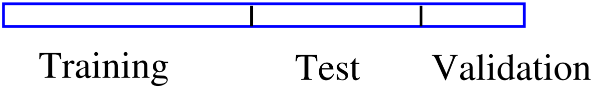
Table

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

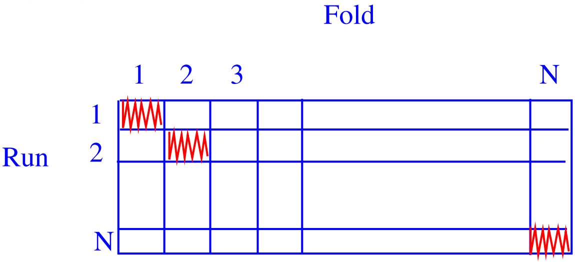
Empirical estimates

When data abound

(MNIST)



Cross validation



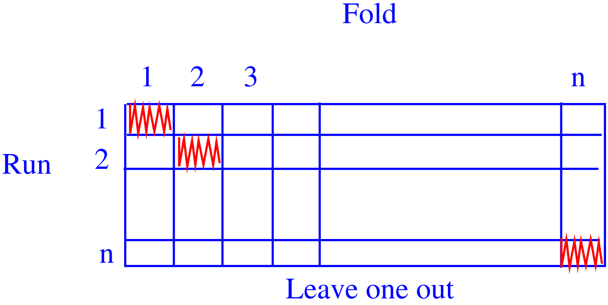
N–fold Cross Validation

Error = Average (error on of h

learned from)

Empirical estimates, foll'd

Cross validation → Leave one out



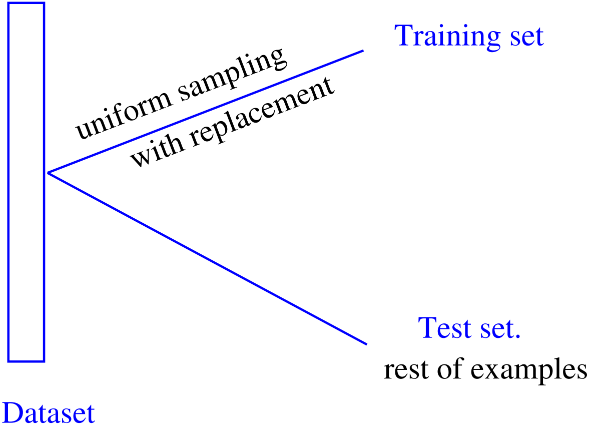
Same as N-fold CV, with N = number of examples.

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

- Video and slides: ICML 2012, Videolectures, Tutorial Japkowicz & Shah http://www.mohakshah.com/tutorials/icml2012/

Overview

Performance indicators

Estimating an indicator

Testing

Hyper-parameter tuning

Is $\hat{h}$ better than random ?

The McNemar test

McNemar 47

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

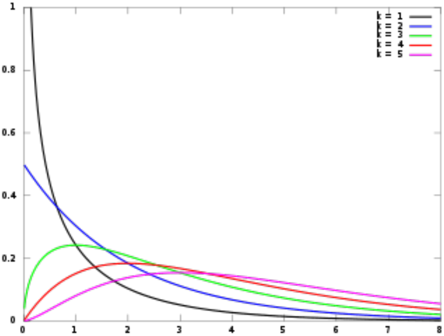
Property

$\frac{|b-c|-1}{b+c}$ follows a χ^2 law with degre of freedom 1

Chi-squared distribution with k degrees of freedom

What

Sum of k squared independent Gaussian normal variables.



Types of test error

Type I error
The hypothesis is not significant, and the test thinks it's significant

Type II error
The hypothesis is valid, and the test discards it.

Comparing algorithms A and B

	A	B	A-B
run 1	30	28	2
run 2	17	25	-8
	28	25	3
	17	28	-11
	30	26	4

Assumption

A and B have normal distribution

Simplest case

two samples with same size, (quasi) same variance.

Define

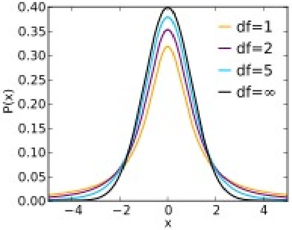
$$t = \frac{\bar{A} - \bar{B}}{S_{A,B} \cdot \sqrt{\frac{2}{n}}}$$

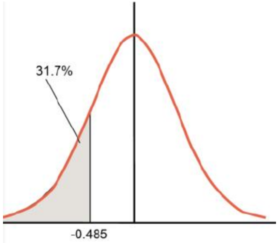
with $S_{A,B} = \sqrt{\frac{1}{2}(S_A^2 + S_B^2)}$ and $S_A^2 = \frac{1}{n}\sum(A_i - \bar{A})^2$

Comparing algorithms A and B

t follows a Student law with (2n-2)-dof

- Compute t
- See confidence of t





Comparing algorithms A and B

Recommended: Use paired t-test

- Apply A and B with same (training, test) sets
- Variance is lower:

$$Var(A - B) = Var(A) + Var(B) - 2coVar(A, B)$$

- Thus easier to make significant differences

What if variances are different ?

See Welch' test:

$$\frac{\bar{A} - \bar{B}}{\sqrt{\frac{S_A^2}{N_A} + \frac{S_B^2}{N_B}}}$$

Summary: single dataset (if we had enough data...)

The 5 x 2CV

Dietterich 98

- 5 times
- split the data into 2 halves
- gives 10 estimates of error indicator

+ More independent

− Each training set is 1/2 data.

With a single dataset

- 5x2 CV
- paired t-test
- McNemar test on a validation set

Multiple datasets

If A and B results don't follow a normal distribution

$$Z_i = A_i - B_i$$

A	B	|Z|	rank	sign
19	23	4	6th	$-$
22	21	1	1st	$+$
21	19	2	2nd	$+$
25	28	3	4th	$-$
24	22	2	2nd	$+$
23	20	3	4th	$+$

Wilcoxon signed rank test

1. Rank the $|Z_i|$
2. W_+ = sum of ranks when $Z_i > 0$
3. W_- = sum of ranks when $Z_i < 0$
4. $W_{min} = min(W_+, W_-)$

$$z = \frac{1/4n(n+1) - W_{min} - 1/2}{\sqrt{1/24n(n+1)(2n+1)}}$$

5. $z \sim \mathcal{N}(0,1)$ $\qquad n > 20$

Multiple hypothesis testing

Beware

- If you test many hypotheses on the same dataset
- one of them will appear confidently true...
  increase in type I error

Corrections Over n tests, the global significance level α_{global} is related to the elementary significance level α_{unit}:

$$\alpha_{global} = 1 - (1 - \alpha_{unit})^n$$

- Bonferroni correction pessimistic

$$\alpha_{unit} = \frac{\alpha_{global}}{n}$$

- Sidak correction

$$\alpha_{unit} = 1 - (1 - \alpha_{global})^{\frac{1}{n}}$$

Overview

How to set up my system ?

Parameter tuning

- Setting the parameters for feature extraction
- Select the best learning algorithm
- Setting the learning parameters (e.g. type of kernel, the parameters in SVMs)
- Setting the validation parameters

Goal: find the best setting a pervasive concern

- Algorithm selection in Operational Research
- Parameter tuning in Stochastic Optimization
- Meta-Learning in Machine Learning

From Design of Experiments to ...

Main approaches

1. Design of experiments (Latin square)
2. Anova (Analysis of variance)-like methods:
   - Racing
   - Sequential parameter optimization

Parameter Tuning: A Meta-Optimization problem

Learner Dataset

Validation

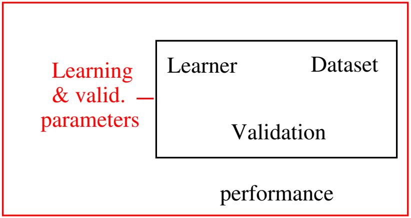
performance

Optimization: the Black-Box Scenario

- Need to perform several runs to compute performance

  Cross-Validation
- Need to specify the # runs and tune it optimally
- Overall cost is the total number of evaluations
- And don't forget to tune the parameters of the meta-optimizer!

Parameter Tuning: A Meta-Optimization problem

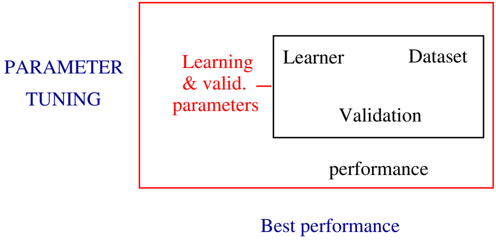


Optimization: the Black-Box Scenario

- Need to perform several runs to compute performance

  Cross-Validation
- Need to specify the # runs — and tune it optimally
- Overall cost is the total number of evaluations
- And don't forget to tune the parameters of the meta-optimizer!

Parameter Tuning: A Meta-Optimization problem



Optimization: the Black-Box Scenario

- Need to perform several runs to compute performance

  Cross-Validation
- Need to specify the # runs — and tune it optimally
- Overall cost is the total number of evaluations
- And don't forget to tune the parameters of the meta-optimizer!

Ingredients

Design Of Experiments (DOE)

- A long-known method from statistics
- Choose a finite number of parameter sets
- Compute their performance
- Return the *statistically significantly* best sets

Analysis of Variance (ANOVA)

- Assumes normally distributed data
- Tests if means are significantly different

  for a given confidence level; generalizes T-Test
- Perform pairwise tests if ANOVA reports some difference

  T-Test, rank-based tests, ...

DOE: Issues

Choice of sample parameter sets

- Full Factorial Design
  - Discretize all parameters if continuous
  - Choose all possible combinations
- Latin Hypercube Sampling: to generate k sets,
  - Discretize all parameters in k values
  - Repeat k times:
    for each parameter, (uniformly) choose one value out of k
  - For each parameter, each value is taken once

fine if no correlation

Cost

- For each parameter set, the full cost of learning validation
- Combinatorial explosion with number of parameters and precision

Racing algorithms

Birattari & al. 02, Yuan & Gallagher 04

Rationale

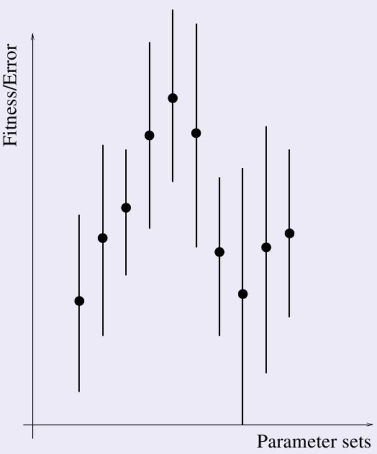
- All parameter settings are run the same number of times
  whereas very bad settings could be detected earlier

Implementation

- Repeat
  - Perform only a few runs per parameter set
  - Statistically check all sets against the best one
    at given confidence level
  - Discard the bad ones
- Until only survivor, or maximum number of runs per setting reached

Racing algorithms

How?

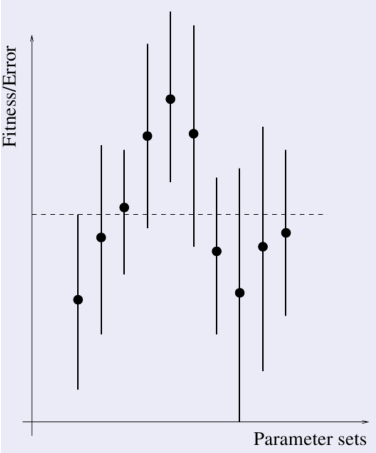


Example: Initialization

- $R = 0$
- while $R < R_{max}$ and more than 1 set
  - Compute empirical value of performance for all sets doing r additional runs
    average, median, ...
  - Compute X% confidence intervals
    Hoeffding bounds, Friedman tests, ...
  - Remove sets whose best possible value is worse than worse possible value of the best empirical set.
  - $R+ = r$

Racing algorithms

How?

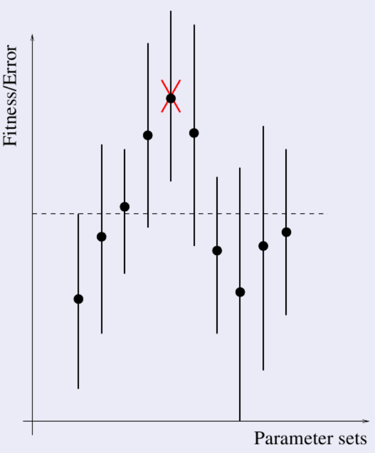


Example: Initialization

- $R = 0$
- while $R < R_{max}$ and more than 1 set
  - Compute empirical value of performance for all sets doing r additional runs
    average, median, ...
  - Compute X% confidence intervals
    Hoeffding bounds, Friedman tests, ...
  - Remove sets whose best possible value is worse than worse possible value of the best empirical set.
  - $R+ = r$

Racing algorithms

How?

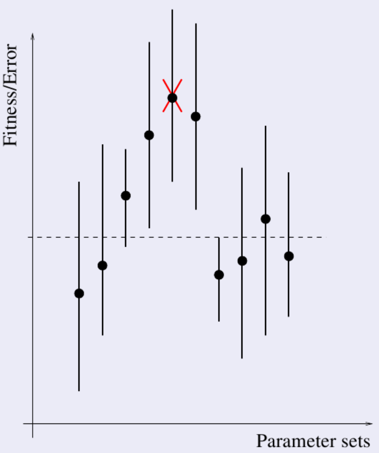


Example: Initialization

- $R = 0$
- while $R < R_{max}$ and more than 1 set
  - Compute empirical value of performance for all sets doing r additional runs
    average, median, ...
  - Compute X% confidence intervals
    Hoeffding bounds, Friedman tests, ...
  - Remove sets whose best possible value is worse than worse possible value of the best empirical set.
  - $R+ = r$

Racing algorithms

How?

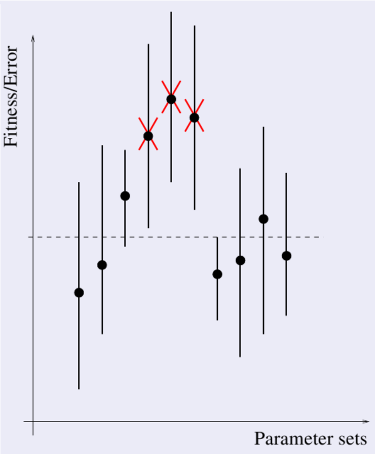


Example: Iteration 1

- $R = 0$
- while $R < R_{max}$ and more than 1 set
  - Compute empirical value of performance for all sets doing r additional runs
    average, median, ...
  - Compute X% confidence intervals
    Hoeffding bounds, Friedman tests, ...
  - Remove sets whose best possible value is worse than worse possible value of the best empirical set.
  - $R+ = r$

Racing algorithms

How?

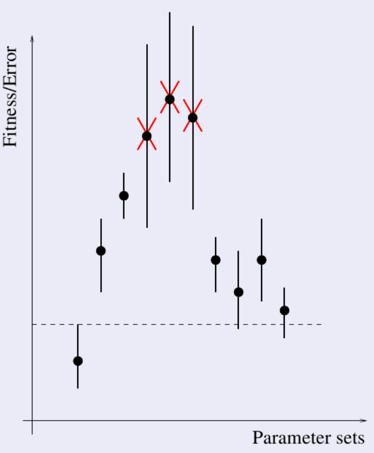


Example: Iteration 1

- $R = 0$
- while $R < R_{max}$ and more than 1 set
  - Compute empirical value of performance for all sets doing r additional runs
    average, median, ...
  - Compute X% confidence intervals
    Hoeffding bounds, Friedman tests, ...
  - Remove sets whose best possible value is worse than worse possible value of the best empirical set.
  - $R+ = r$

Racing algorithms

How?

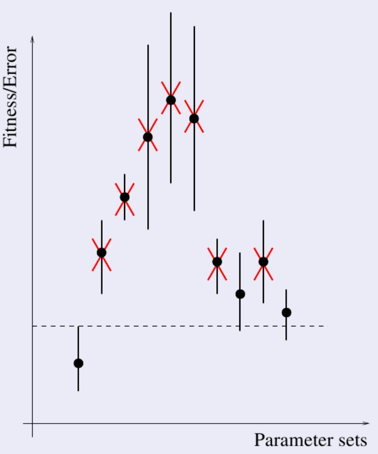


Example: Iteration N

- $R = 0$
- while $R < R_{max}$ and more than 1 set
  - Compute empirical value of performance for all sets doing r additional runs
    average, median, ...
  - Compute X% confidence intervals
    Hoeffding bounds, Friedman tests, ...
  - Remove sets whose best possible value is worse than worse possible value of the best empirical set.
  - $R+ = r$

Racing algorithms

How?

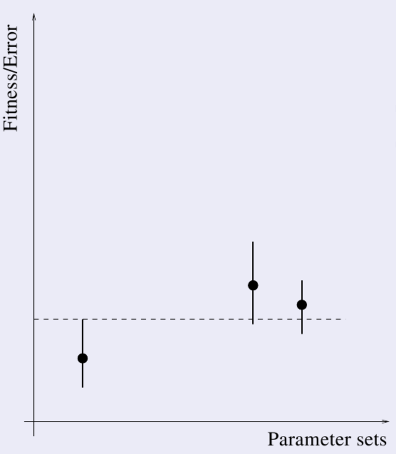


Example: Iteration N

- $R = 0$
- while $R < R_{max}$ and more than 1 set
  - Compute empirical value of performance for all sets doing r additional runs
    average, median, . . .
  - Compute X% confidence intervals
    Hoeffding bounds, Friedman tests, . . .
  - Remove sets whose best possible value is worse than worse possible value of the best empirical set.
  - $R+ = r$

Racing algorithms

How?



Example: Best parameter sets

- $R = 0$
- while $R < R_{max}$ and more than 1 set
  - Compute empirical value of performance for all sets doing r additional runs
    average, median, ...
  - Compute X% confidence intervals
    Hoeffding bounds, Friedman tests, ...
  - Remove sets whose best possible value is worse than worse possible value of the best empirical set.
  - $R+ = r$

Racing algorithms: Discussion

Results

- Published results claim saving between 50 and 90% of the runs

Useful for

- Multiple algorithms on single problem — for efficiency
- Single algorithm on multiple problems — to assess problem difficulties
- Multiple algorithms on multiple problems — for robustness

Issues

- Nevertheless costly
- Can only find the best one in initial sample

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