Exploratory and Inferential Analysis of Benchmark Experiments

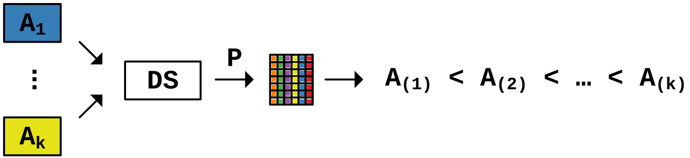
Manuel J. A. Eugster and Friedrich Leisch

Department of Statistics
Ludwig-Maximiliams-Universität München

International Conference on Computational Statistics, 2008

Benchmark experiments

Most popular scenario:



Benchmark experiments

Data set:

Given a data set $\mathfrak{L} = \{z_1, \ldots, z_n\}$, we draw B learning samples $(i = 1, \ldots, B)$:

$$\mathfrak{L}^i = \{z_1^i, \ldots, z_n^i\}$$

Candidate algorithms:

There are $K > 1$ algorithms a_k $(k = 1, \ldots, K)$;
$a_k(\cdot \mid \mathfrak{L}^b)$ is the fitted model based on the sample $\mathfrak{L}^b$ with the distribution $\mathcal{A}_k$:

$$a_k(\cdot \mid \mathfrak{L}^b) \sim \mathcal{A}_k(\mathfrak{L})$$

Benchmark experiments

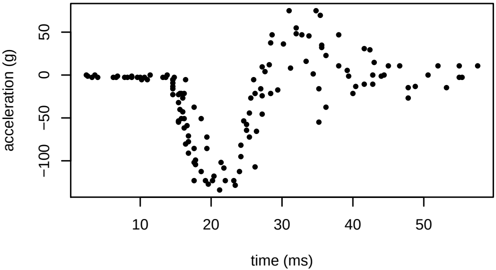
Performance measure:
Analytically, performance is measured by the scalar function:

$$p_{kb} = p(a_k, \mathfrak{L}^b) \sim \mathcal{P}_k = \mathcal{P}_k(\mathfrak{L})$$

The empirical analogue $\hat{p}_{kb}$ based on a test sample $\mathfrak{T}$; a common choice is $\mathfrak{T} = \mathfrak{L} \setminus \mathfrak{L}^b$.

Exemplar benchmark experiment

Experiment:



(1) regression problem `motorcycle`; **(2)** algorithms $\{\texttt{lm}, \texttt{nls}, \texttt{nnet}, \texttt{rpart}, \texttt{gam}, \texttt{loess}, \texttt{gamboost}\}$; **(3)** mean squared error; **(4)** bootstrap 250 samples; **(5)** out-of-bootstrap samples;

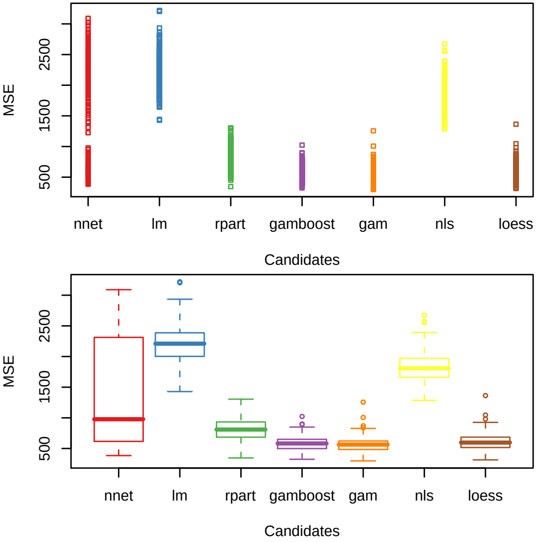
Exemplar benchmark experiment

Result:

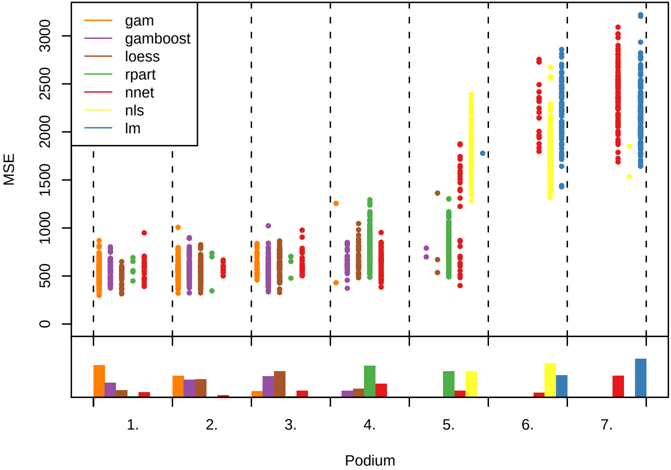
```
        nnet     lm     rpart gamboost    gam    nls  loess
  [1,]  669.2 2255.7     847.2    559.7  511.3 1933.3  548.8
  [2,]  722.3 2194.9     957.1    626.9  582.1 1737.5  613.9
   ...
[249,] 1967.4 2095.9     659.2    417.4  489.6 1561.9  579.3
[250,] 1508.2 1962.3     926.6    509.1  440.6 1674.9  614.3
```

Exploratory analysis

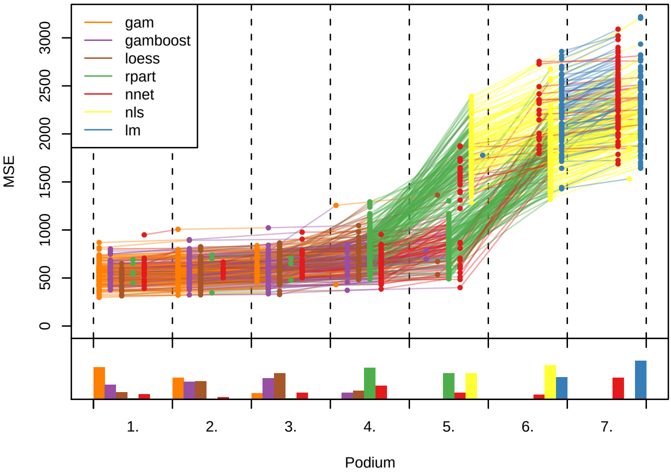
Basic plots


MSE
500
1500
2500
nnet
lm
rpart
gamboost
gam
nls
loess
Candidates

Benchmark experiment plot



Benchmark experiment plot



Summary statistics and simple rankings

	Mean	SD	Median	Max	95% CI (Mean)
nnet	1438.1	868.4	977.2	3090.2	[1329.4, 1546.8]
lm	2209.2	294.1	2209.8	3219.4	[2172.3, 2246.0]
rpart	812.4	181.2	809.2	1304.6	[789.7, 835.1]
gamboost	583.7	116.5	582.1	1022.9	[569.1, 598.3]
gam	565.2	122.6	563.6	1256.0	[549.9, 580.6]
nls	1818.1	242.3	1808.5	2674.7	[1787.8, 1848.4]
loess	604.3	134.6	596.6	1363.1	[587.4, 621.1]

Mean:

gam $<$ gamboost $<$ loess $<$ rpart $<$ nnet $<$ nls $<$ lm

Minimax:

gamboost $<$ gam $<$ rpart $<$ loess $<$ nls $<$ nnet $<$ lm

Mean – 95% CI:

gamboost $\approx$ gam $\approx$ loess $<$ rpart $<$ nnet $<$ nls $<$ lm

Inferential analysis

Inferential analysis

Random block design:

$$p_{ij} = \kappa_0 + \kappa_j + b_i + \epsilon_{ij},$$
$$i = 1, \dots, B, j = 1, \dots (K-1),$$

with different assumptions on κ_j, b_i and ϵ_{ij}.

Test problem:

$$H_0 : \ \kappa_1 = \cdots = \kappa_{K-1} = 0,$$
$$H_A : \ \exists j : \ \kappa_j \neq 0,$$

using parametric and non-parametric methods.

Linear mixed effects model

Assumptions:

κ_j fixed effect, b_i random effect,

$$b_i \sim N(0, \sigma_b^2), \epsilon_{ij} \sim N(0, \sigma^2).$$

Test problem:

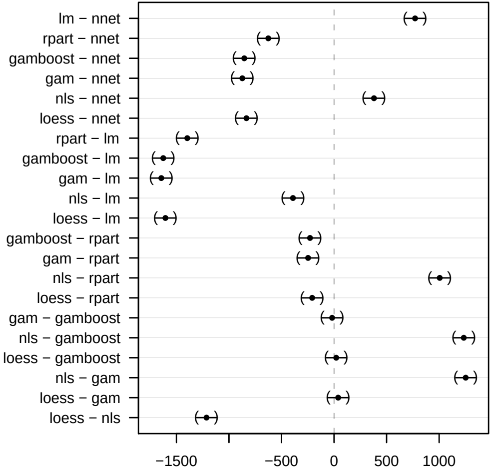
Pairwise comparisons with Tukey contrasts.

Pairwise comparisons

```
 General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts

Linear Hypotheses:
                         Estimate
lm - nnet == 0             771.05
rpart - nnet == 0         -625.70
gamboost - nnet == 0      -854.45
gam - nnet == 0           -872.86
nls - nnet == 0            379.98
loess - nnet == 0         -833.83
rpart - lm == 0          -1396.75
gamboost - lm == 0       -1625.50
gam - lm == 0            -1643.91
nls - lm == 0             -391.06
loess - lm == 0          -1604.88
gamboost - rpart == 0     -228.75
gam - rpart == 0          -247.16
nls - rpart == 0          1005.69
loess - rpart == 0        -208.13
gam - gamboost == 0        -18.41
nls - gamboost == 0       1234.43
loess - gamboost == 0       20.62
nls - gam == 0            1252.85
loess - gam == 0            39.03
loess - nls == 0         -1213.81
```



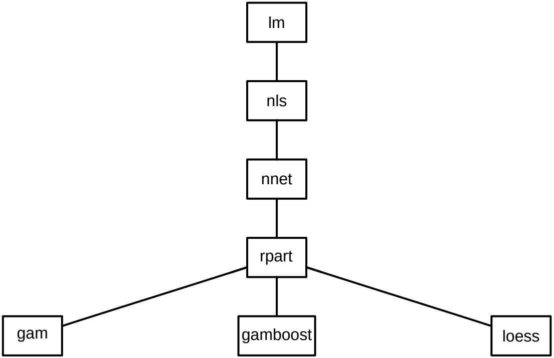
Order relation

In case of a significant difference between two algorithms we define a strict total order $<$, otherwise the algorithms are $\approx$-related.

Pairwise orders:
$\texttt{nnet} < \texttt{lm}, \texttt{rpart} < \texttt{nnet}, \texttt{gamboost} < \texttt{nnet}, \ldots, \texttt{gam} \approx \texttt{gamboost}, \texttt{gamboost} < \texttt{nls}, \texttt{gamboost} \approx \texttt{loess}, \ldots$

Binary relation:
Domain is $\{\mathcal{A}, \mathcal{A}\}$, where $\mathcal{A}$ is the set of candidate algorithms; the graph is the set $\{(\texttt{nnet}, \texttt{lm}), (\texttt{rpart}, \texttt{nnet}), \ldots\}$.

Order relation

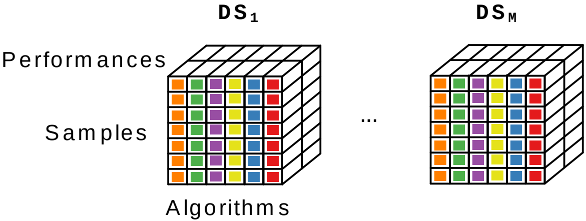


Topological sort:

$\texttt{gam} \approx \texttt{gamboost} \approx \texttt{loess} < \texttt{rpart} < \texttt{nnet} < \texttt{nls} < \texttt{lm}$

Further developments

More complex scenarios



Exploratory and inferential analysis tools, e.g.:

Consensus: overall order based on different data sets and different performance measures.

Inference: model the design with two experimental factors, their interactions and blocking factors at two levels.

Papers & Software

... at `http://statistik.lmu.de/~eugster/benchmark/`:

`R` **Package:**

`benchmark` version 0.01.

Reports:

Exploratory and Inferential Analysis of Benchmark Experiments.
Manuel J. A. Eugster, Torsten Hothorn and Friedrich Leisch. Technical Report 30, LMU Munich. **R supplement "The *uci621* benchmark experiment"**.