Sequential/Adaptive Benchmarking

Manuel J. A. Eugster

Institut für Statistik
Ludwig-Maximiliams-Universität München

Validation in Statistics and Machine Learning, 2010

Benchmark experiments

Data generating process:

Given a data generating process DGP, we draw B independent and identically distributed learning samples:

$$\mathfrak{L}^1 = \{z_1^1, \ldots, z_n^1\} \sim DGP$$

$$\vdots$$

$$\mathfrak{L}^B = \{z_1^B, \ldots, z_n^B\} \sim DGP$$

(*) Following Hothorn, Leisch, Zeileis, and Hornik (2005).

Candidate algorithms:

There are $K > 1$ algorithms a_k $(k = 1, \ldots, K)$ with the function $a_k(\cdot \mid \mathfrak{L}^b)$ the fitted model on the learning sample $\mathfrak{L}^b$.

Performance measure:

The performance of algorithm a_k when provided with the learning sample $\mathfrak{L}^b$ is measured by a scalar function p:

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

Inference:
Given the K different random samples $\{p_{k1}, \ldots, p_{kB}\}$ with B iid samples drawn from the distributions $\mathcal{P}_k(DGP)$ the null hypothesis of interest for most problems is:

$$H_0 \ : \ \mathcal{P}_1 = \cdots = \mathcal{P}_K$$

Test procedure:
An algorithm a_k is better than an algorithm $a_{k'}$ with respect to a performance measure p and a functional ϕ iff $\phi(\mathcal{P}_k) < \phi(\mathcal{P}_{k'})$ $(k, k' \in \{1, \ldots, K\})$.

$$T \begin{cases} H_0: & \phi(\mathcal{P}_1) = \cdots = \phi(\mathcal{P}_K) \\ H_1: & \exists k, k' : \phi(\mathcal{P}_k) \neq \phi(\mathcal{P}_{k'}) \end{cases}$$

For $b = 1, \ldots, B$

1. Draw learning sample $\mathfrak{L}^b$.

2. Measure performance p_{kb} of the $k = 1, \ldots, K$ candidate algorithms.

Execute test procedure T on the K performance estimations $\{p_{1k}, \ldots, p_{Bk}\}$ and make a decision for a given α.

For $b = 1, \dots, B$

1. Draw learning sample $\mathfrak{L}^b$.

2. Measure performance p_{kb} of the $k = 1, \dots, K$ candidate algorithms.

Execute test procedure T on the K performance estimations $\{p_{1k}, \dots, p_{Bk}\}$ and make a decision for a given α.

- Benchmark experiments are considered as **fixed-sample** experiments; hypotheses of interests are tested using a test T at the end.
- In most benchmark experiments B is a **freely chosen** number (often specified depending on the algorithms' running time).
- The nature of benchmark experiments is **sequential**.

Do

1. Draw learning sample $\mathfrak{L}^b$.

2. Measure performance p_{bk} of the $k = 1, \ldots, K$ candidate algorithms.

3. Execute test procedure T on the K performance estimations $\{p_{1k}, \ldots, p_{bk}\}$.

While no decision for a given α (and $b \leq B$).

Do

1. Draw learning sample $\mathfrak{L}^b$.

2. Measure performance p_{bk} of the $k = 1, \ldots, K$ candidate algorithms.

3. Execute test procedure T on the K performance estimations $\{p_{1k}, \ldots, p_{bk}\}$.

While no decision for a given α (and $b \leq B$).

- **Sequential/Adaptive benchmarking:** execute test T successively on the accumulating data.
- This enables
  - **(1)** to monitor the benchmark experiment, and
  - **(2)** to make a decision – to *stop* or to *go on*.

Exemplar benchmark experiments

(1) $\mathfrak{L}$ is the Pima Indians Diabetes data set; **(2)** $\mathfrak{L}^b$ by bootstrapping; **(3)** linear discriminant analysis (`lda`), support vector machine with $C = 1.00$ (`svm1`), support vector machine with $C = 1.01$ (`svm2`), random forest (`rf`); **(4)** misclassification on the out-of-bag samples; **(5)** $B = 100$.

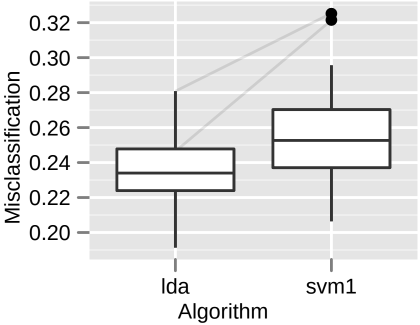
$\Rightarrow$ compare two algorithms at a time, i.e., test if algorithm a_1 is better than algorithm a_2.

(6) Wilcoxon Signed Rank test, $\alpha = 0.05$.

Monitoring

Observe and interpret the test result, mainly the p-value, on the accumulating performance measures.

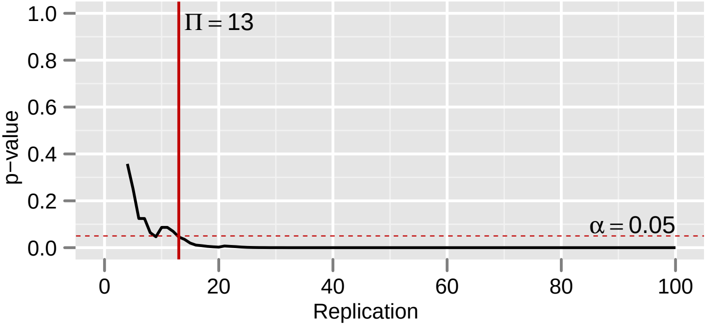
Scenario 1 – Different algorithm performances:



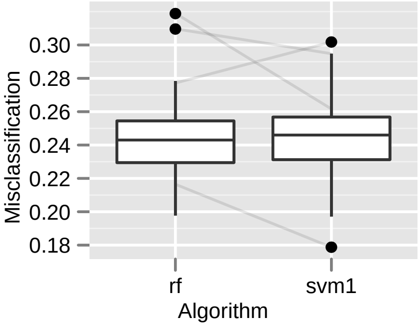
p-value $= 1.312e - 12$

$\Rightarrow$ `lda` $\prec$ `svm1`

Scenario 1 – Different algorithm performances:



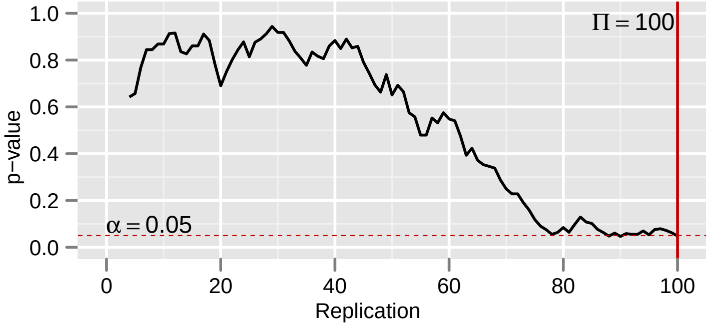
Scenario 2 – Similar algorithm performances:



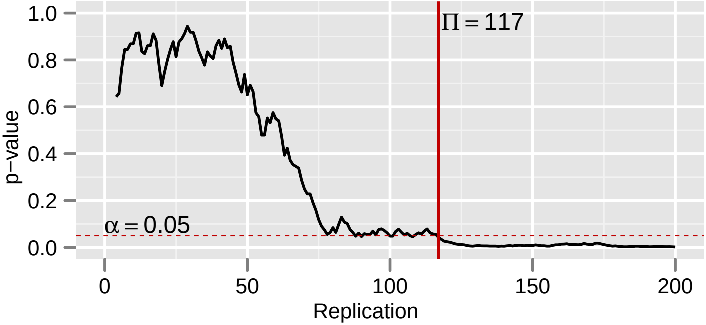
p-value $= 0.0484$

$\Rightarrow$ `rf` $\prec$ `svm1`

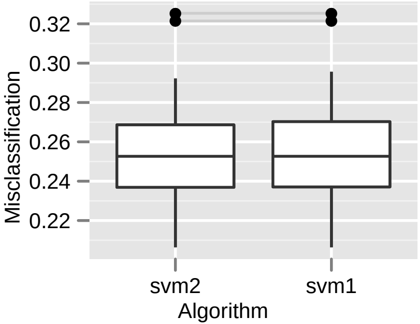
Scenario 2 – Similar algorithm performances:



Scenario 2 – Similar algorithm performances:



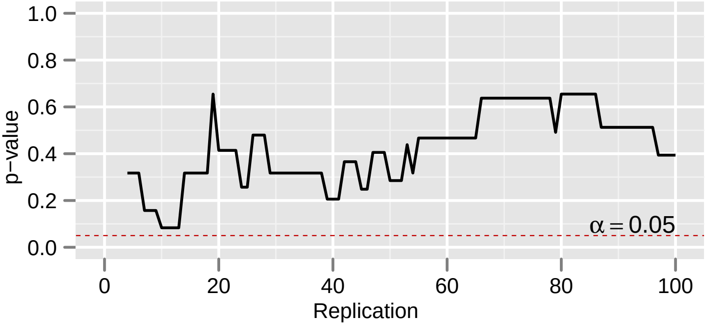
Scenario 3 – Equal algorithm performances:



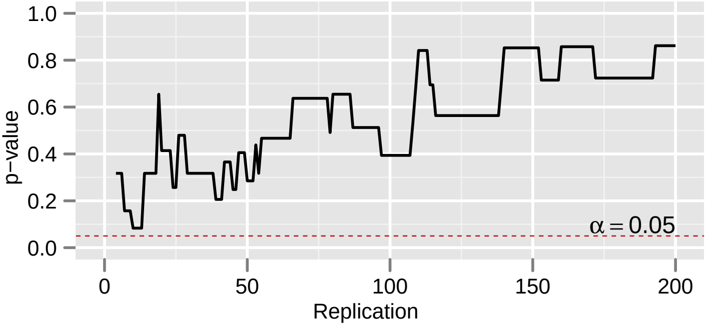
p-pvalue $= 0.9302$

$\Rightarrow$ `svm2` $\approx$ `svm1`

Scenario 3 – Equal algorithm performances:



Scenario 3 – Equal algorithm performances:



Interpretation

Point consecutively significance:

$$\Pi_{\text{Scenario 1}} = 13,\ \Pi_{\text{Scenario 2}} = 117,\ \Pi_{\text{Scenario 3}} = \infty$$

Measure of "how big the difference" is – indicator for relevance?

Decision making

Execute a benchmark experiment as long as needed – either until H_0 is rejected or H_0 is “accepted” (failed to reject).

Analyses on accumulating data

Sequential: Sample observations one by one; the test is executed after each new observation – the experiment can be stopped at any point.

Group sequential: Sample groups of observations; the test is executed after each group – the experiment can be stopped after each group.

Adaptive: Group sequential with more flexibility, e.g., to change hypothesis, group sample size, etc.

(*) Following Vandemeulebroecke (2008).

Sequential analysis of benchmark experiments:

General differences to the common field (e.g., clinical trials).

1. Compared to clinical trials it is easy and (relatively) cheap to make additional replications until a final decision, i.e., to reject or accept H_0; so, (theoretically) there is no undecidable situation.
2. Benchmark experiments are computer experiments often executed on remote servers, etc; so decisions made in the interim and planning phases need a sound automatization (or "interactive" interim and planning phases).

Recursive combination tests

Adaptive test procedure based on the recursive application of two-stage combination tests; and a *p*-value function to combine *p*-values from two stages.

(*) Defined by Brannath, Posch, and Bauer (2002).

Two-stage combination tests:

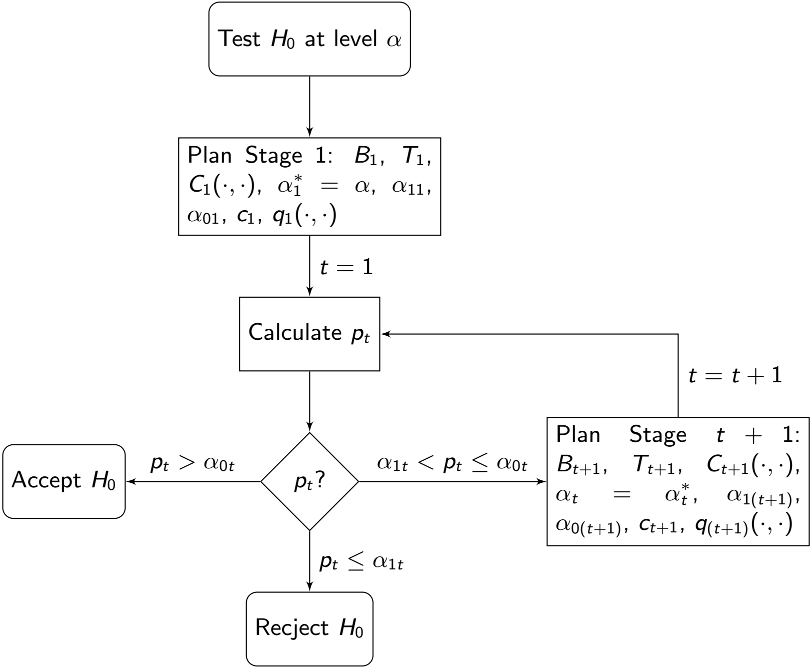
Test a one-sided null hypothesis H_0 at level α using two stages.

Stage 1: B_1 learning samples, decision boundaries α_{01} and α_{11} $(0 \leq \alpha_{11} < \alpha < \alpha_{01} \leq 1)$.

$$\text{Decision} = \begin{cases} \text{reject } H_0, & p_1 \leq \alpha_{11} \\ \text{accept } H_0, & p_1 > \alpha_{01} \\ \text{perform Stage 2}, & \alpha_{11} < p_1 \leq \alpha_{01} \end{cases}$$

Stage 2: B_2 learning samples.

$$\text{Decision} = \begin{cases} \text{reject } H_0, & C(p_1, p_2) \leq c \\ \text{undecidable}, & \text{otherwise} \end{cases}$$



Test H_0 at level α
Plan Stage 1: B_1, T_1, $C_1(\cdot,\cdot)$, $\alpha_1^* = \alpha$, α_{11}, α_{01}, c_1, $q_1(\cdot,\cdot)$
$t = 1$
Calculate p_t
$t = t + 1$
$p_t?$
$p_t > \alpha_{0t}$
Accept H_0
$\alpha_{1t} < p_t \leq \alpha_{0t}$
Plan Stage $t+1$: B_{t+1}, T_{t+1}, $C_{t+1}(\cdot,\cdot)$, $\alpha_t = \alpha_t^*$, $\alpha_{1(t+1)}$, $\alpha_{0(t+1)}$, c_{t+1}, $q_{(t+1)}(\cdot,\cdot)$
$p_t \leq \alpha_{1t}$
Recject H_0

Fisher's combination test:

$$C(p_{t-1}, p_t) = p_{t-1} \cdot p_t$$

Critical value:

$$c_t = \frac{\alpha_t^* - \alpha_{1t}}{\ln \alpha_{0t} - \ln \alpha_{1t}}$$

Conditional significance level:

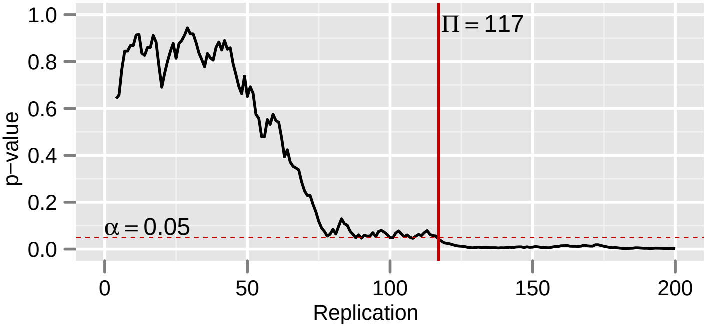
$$\alpha_1^* = \alpha,\ \alpha_t^* = \frac{c_{t-1}}{p_{t-1}}$$

Conditional decision boundaries:

$$\alpha_{1t} < \alpha_t^* \leq \alpha_{0t}$$

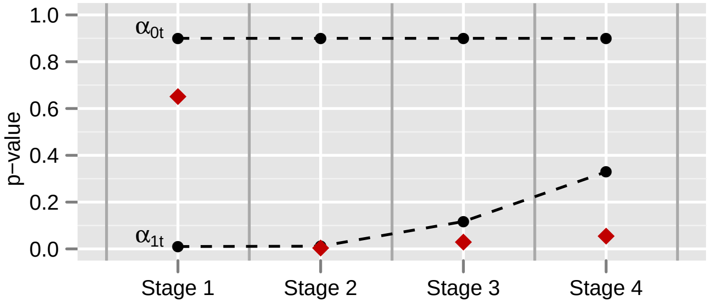
Scenario 2 – Similar algorithm performances:

Split the scenario into four stages, i.e., $B_t = \frac{B}{4} = 50$; define $\alpha = \alpha_1^* = 0.05$, $\alpha_{11} = 0.01$, $\alpha_{01} = 0.9$; and define the rule for the conditional decision boundaries as $\alpha_{0t} = \frac{\alpha_t^*}{1.2}$ and $\alpha_{1t} = \alpha_{1(t-1)}$.



Scenario 2 – Similar algorithm performances:

Split the scenario into four stages, i.e., $B_t = \frac{B}{4} = 50$; define $\alpha = \alpha_1^* = 0.05$, $\alpha_{11} = 0.01$, $\alpha_{01} = 0.9$; and define the rule for the conditional decision boundaries as $\alpha_{0t} = \frac{\alpha_t^*}{1.2}$ and $\alpha_{1t} = \alpha_{1(t-1)}$.



Summary

Monitoring:

- Point of consecutively significance.

Decision making:

- Statistical justification for the number of replications.
- Sound interim phase (or better an "interactive" one)?
- General advantage in real-world benchmark experiments?

`http://CRAN.R-project.org/package=benchmark`

Werner Brannath, Martin Posch, and Peter Bauer. Recursive combination tests. *Journal of the American Statistical Association*, 97(457):236–244, 2002.

Torsten Hothorn, Friedrich Leisch, Achim Zeileis, and Kurt Hornik. The design and analysis of benchmark experiments. *Journal of Computational and Graphical Statistics*, 14(3):675–699, 2005.

Marc Vandemeulebroecke. Group sequential and adaptive designs – a review of basic concepts and points of discussion. *Biometrical Journal*, 50(3), 2008.