Aplikace matematiky

Zdeněk Režný

Algorithm. 45. Bayes factor test. An algorithm giving tables for a non-parametric two-sample test of the Bayesian discrimination power of an observed factor

Aplikace matematiky, Vol. 25 (1980), No. 5, 390–394

Persistent URL: http://dml.cz/dmlcz/103874

ALGORITMUS

45. BAYES FACTOR TEST

AN ALGORITHM GIVING TABLES FOR A NON-PARAMETRIC TWO-SAMPLE TEST OF THE BAYESIAN DISCRIMINATION POWER OF AN OBSERVED FACTOR

Ing. ZDENĚK REŽNÝ, CSc.,

Ústav biofyziky a nukleární medicíny Fakulty všeobecného lékařství Karlovy university, Salmovská 3, 120 00 Praha 2

The algorithm suggested in this paper concerns the Bayesian statistical decision problem (see [1], § 8.8) with the parameter space $\Omega = \{w_1, w_2\}$, decision space $D = \{d_1, d_2\}$, some loss function $L(w_i, d_j)$ and decision functions defined by

$$\delta(x; i, c) = d_i \quad \text{for} \quad x \leqq c\,, \qquad (1)$$
$$= d_{3-i} \quad \text{for} \quad x > c \quad (i = 1, 2;\ c \text{ real})\,.$$

The only assumed known property of the two conditional distribution functions of the observed variable X is their *continuity*, and therefore they are replaced by the *sample* distribution functions from random samples of sizes $n[1]$ and $n[2]$ made under the conditions $W = w_1$ and w_2 respectively; thus, the resulting Bayes risk is a random variable (which we shall call *sample B. risk*), and its distribution function, under the hypothesis H_0 that both conditional distributions are identical, can then be looked on as the *significance level* in testing the contrast between the two samples, or, in other words, the discrimination power of the variable X.

The apriori probabilities and loss values are postulated to be fractions with integer numerators $nprob[i] \geqq 0$ and $nloss[i, j]$ respectively so that

$$\xi_1 : \xi_2 = nprob[1] : nprob[2] \quad (\xi_1 + \xi_2 = 1)\,, \qquad (2)$$
$$L(w_i, d_j) = nloss[i, j]/dloss \qquad (i, j = 1, 2)\,,$$

and $dloss > 0$. The algorithm, without exercing any side effect on the input parameters, calculates first the Bayes risk $apr\,\varrho$ based exclusively on the apriori probabilities, i.e. without considering the observation X, and the number $imax$ of values $\varrho[i]$ which have positive probabilities of being taken on by the sample B. risk. In

the trivial case ($\xi_i = 0$ or $n[i] = 0$ for some i or one of the decisions d_1, d_2 is inadmissible or both are equivalent) *imax* turns out to be zero and the algorithm stops working. Otherwise, there are calculated the $\varrho[i]$ ($\varrho[1] < \varrho[2] < \ldots < \varrho[imax]$) and the probabilities $\alpha[i]$ that, under the above hypothesis H_0, the sample B. risk will not be greater than $\varrho[i]$ ($i = 1, \ldots, imax$; $\alpha[imax] = 1$). The α's are fractions with integer numerators $n\alpha[i]$ and integer common denominator $d\alpha$, which are also given. Finally, integer quantities $r[j_1, j_2]$ ($1 \leqq r[j_1, j_2] = r[n[1] - j_1, n[2] - j_2] \leqq \leqq imax + 1$ for $j_k = 0, \ldots, n[k]$; $k = 1, 2$) are calculated for the purposes of processing a single case, i.e. for determining the subscript i of the sample B. risk $\varrho[i]$ and its significance level $\alpha[i]$ or, more generally, the bounds within which it lies. Namely, if $\{x_k\colon 1 \leqq k \leqq l\}$ is the ordered set $(x_1 < x_2 < \ldots < x_l)$ of mutually different measurements from the combined sample (of $n[1] + n[2]$ values of X) and m_{jk} the number of such values from the j-th sample which are not greater than x_k, $m_{j0} = 0$ ($j = 1, 2$; $1 \leqq k \leqq l$), the relations

$$(3) \qquad i_1 = \operatorname*{Min}_{1 \leqq k \leqq l} \operatorname{Min} \{r[m_{1k}, m_{2,k-1}], r[m_{1,k-1}, m_{2k}]\} \leqq$$

$$\leqq i \leqq i_2 = \operatorname*{Min}_{1 \leqq k \leqq l} r[m_{1k}, m_{2k}],$$

$$i \leqq imax, \quad i_1 \geqq 1, \quad i_2 \leqq imax + 1$$

hold. The case $i_1 < i_2$ occurs if and only if there are some mixed ties in measured values of the combined sample, or, equivalently, if $m_{j,k-1} < m_{jk}$ for some k and both $j = 1, 2$. (With probability one, some higher precision of measurements of the same values of X would then lead to the equality $i_1 = i_2$.) If $i_2 = imax + 1$, then the Bayesian decision rule constructed on the basis of the actual measurements would be worse than that in the case of the most unfavorable arrangement of the combined sample values without ties.

The algorithm is based on the familiar Bayesian decision calculus (see [1], loc. cit.) as well as on some elementary number theory; moreover, the problem of evaluating the probabilities $\alpha[i]$ could be solved as a slight generalization of some discrete random walk models treated in [2], Ch. III.

procedure *BAYES FACTOR TEST*(*nprob*, *nloss*, *dloss*, *n*)
results: (*apr rho*, *imax*, *rho*, *nalpha*, *dalpha*, *alpha*, *r*);

integer *imax*, *dalpha*; **real** *dloss*, *apr rho*; **array** *rho*, *alpha*;

integer array *nprob*, *nloss*, *n*, *nalpha*, *r*;

begin

integer *i*, *j*, *k*, *m*, *p*, *q*; **real** *a*, *b*; **integer array** *s*, *v*, *w*, *x*[1 : 2], *d*[−1 : *n*[1], −1 : *n*[2]], *h*[1 : 2, 0 : **if** *n*[1] > 2 **then** *n*[1] **else** 2];

procedure *E*(*j*, *k*); **integer** *j*, *k*; **for** *i* := 1, 2 **do** *j* := *k*;

procedure *ADD*(*i*0, *i*1, *j*0);

```
    value i0, i1; integer i0, i1, j0;
    begin integer j;
    for i := i0 step 1 until i1 do
    for j := j0 step 1 until h[2, i] do d[i, j] := d[i - 1, j] + d[i, j - 1]
    end ADD;
imax := 0; b := (nprob[1] + nprob[2]) × dloss;
for j := 1, 2 do
    begin E(h[i, j], nprob[i] × nloss[i, j]); x[j] := h[1, j] + h[2, j] end;
apr rho := x[if x[1] < x[2] then 1 else 2]/b;
E(w[i], n[3 - i] × (h[i, 3 - i] - h[i, i]));
if w[1] × w[2] > 0 then
begin comment non-trivial case;
j := w[1]; k := w[2];
for i := j while k ≠ 0 do begin j := k; k := i - i ÷ j × j end;
E(w[i], w[i] ÷ j);
a := abs(j)/(b × n[1] × n[2]);
b := (if j > 0 then h[1, 1] + h[2, 2] else h[1, 2] + h[2, 1])/b;
k := -1;
for i := 1 step 1 until w[2] do
    begin k := k + w[1]; j := k ÷ w[2]; if k = j × w[2] then go to A1 end;
A1: x[1] := 0;
x[2] := n[2];
s[1] := i;
s[2] := j;
E(v[i], w[3 - i] - s[i]);
for j := 0 step 1 until n[1] do begin d[j, -1] := 0; E(h[i, j], x[i]) end;
for j := 1 step 1 until n[2] do d[-1, j] := 0;
d[-1, 0] := 1;
ADD(0, n[1], 0);
dalpha := d[n[1], n[2]];
for k := 0, k + 1 while m > 0 do
    begin
    if x[2] > n[2] then go to A2; imax := imax + 1; j := x[2];
    for i := x[1], i + w[2] while i ≦ n[1] ∧ j ≦ n[2] do
```

begin

$p := n[1] - i$; $q := n[2] - j$; $r[i, j] := r[p, q] := imax$;
$d[i, j] := d[p, q] := 0$; $h[2, i] := h[2, i] - 1$; $h[1, p] := h[1, p] + 1$;
$j := j + w[1]$
end i;

if $q > x[2]$ **then** $p := n[1] - (k - x[2] \times w[2]) \div w[1]$;
$ADD(x[1] + 1, p - 1, x[2])$;
$ADD(p, n[1], h[1, i])$;
$m := d[n[1], n[2]]$;
$rho[imax] := a \times k + b$;
$nalpha[imax] := dalpha - m$;
$alpha[imax] := 1 - m/dalpha$;

$A2$: **if** $x[1] < v[1]$ **then** $E(x[i], x[i] + s[i])$ **else** $E(x[i], x[i] - v[i])$
end k;

for $i := 0$ **step** 1 **until** $n[1]$ **do**

for $j := h[1, i]$ **step** 1 **until** $h[2, i]$ **do** $r[i, j] := imax + 1$

end non-trivial case

end BAYES FACTOR TEST

The application of the algorithm is connected with the single constraint that the binomial coefficient $\binom{n[1] + n[2]}{n[1]}$, which gives the value of $d\alpha$, will not exceed the maximum integer permitted by the individual compuitng device. Further, as to setting the upper subscript bound in declarations of the actual parameters corresponding in the main program to ϱ, α and $n\alpha$, we may use the inequality

$$imax \leqq (n[1] + 1)(n[2] + 1)/2 \tag{4}$$

or, if desirable, a stronger upper estimate, which is established in the following manner: Let $j_i = nprob[i]\, n[3 - i]\ (nloss[i, 3 - i] - nloss[i, i])$ for $i = 1, 2$. If $j_1 j_2 \leqq 0$ then $imax = 0$ (trivial case). Otherwise, let k_i denote positive integers with the greatest common divisor one and such that $k_1 : k_2 = j_1 : j_2$. Then, if $k_i n[i] \leqq k_{3-i} n[3 - i]$ and $h = \operatorname{Min}\{k_i,\ [k_i(n[i] + 1)/k_{3-i}] + 1\}$, it may be shown that

$$imax \leqq h(n[i] + 1) - \binom{h}{2}(k_{3-i} + 1)/k_i\,. \tag{4'}$$

Check example: Given values according to following table

i	$nprob[i]$	$nloss[i,j]$		$n[i]$
		$j=1$	$j=2$	
1	5	5	15	4
2	15	4	1	6

the statement *BAYES FACTOR TEST* ($nprob$, $nloss$, 10, n, $apr\ rho$, $imax$, rho, $nalpha$, $dalpha$, $alpha$, r) leads to the results

$$apr\ rho = .425\,, \quad imax = 12\,, \quad dalpha = 210\,,$$

i	1	2	3	4	$5 \leqq i \leqq 12$
$rho[i]$	.2000	.2375	.2625	.2750	$.0125i + .2375$

i	1	2	3	4	5	6	7	8	9	10	11	12
$nalpha[i]$	2	10	20	38	62	93	116	146	174	194	206	210

$$alpha[i] = nalpha[i]/210\,,$$

$r[i,j]$

	$j=0$	1	2	3	4	5	6
$i=0$	13	12	9	6	4	2	1
1	12	13	13	11	8	5	3
2	7	10	13	13	13	10	7
3	3	5	8	11	13	13	12
4	1	2	4	6	9	12	13

The algorithm has been tested in its transcription from the presented version into FORTRAN IV and implemented in the Institute of Biophysics and Nuclear Medicine, Faculty of General Medicine, Charles University for the computer ICL-4/72.

[1] *M. H. DeGroot:* Optimal Statistical Decisions, McGraw-Hill Co., New York 1970.
[2] *W. Feller:* An Introduction to Probability Theory and Its Applications, Vol. 1, 3rd ed., J. Wiley, New York 1967.