Diskriminační analýza pozorování kvalitativního typu

Ladislav Tomášek

Předpokládejme, že diskriminace mezi populacemi $H_1, \ldots, H_m$ je založena na vektoru dichotomických nezávislých veličin

$$x = (1, x_1, \ldots, x_p)'.$$

Uvažujme vyjádření posteriorních pravděpodobností ve tvaru

$$P(H_j/x) = \exp(x'a_j) \Big/ \sum_k \exp(x'a_k) ,$$

kde $a_j = (a_{j0}, a_{j1}, \ldots, a_{jp})'$, $j=1,\ldots,m$

$a_m = 0$,

$a = (a_1', a_2', \ldots, a_{m-1}')'.$

Označme dále

n_{jx} počet pozorování v populaci H_j při hodnotě x,
n_x počet pozorování ve všech populacích při hodnotě x,
n_j počet pozorování v populaci H_j a
n počet všech pozorování.

Uvážíme-li, že pro podmíněné pravděpodobnosti platí

$$P(x/H_j) = P(H_j/x)\, P(x) \,/\, P(H_j) ,$$

lze při označení $P(H_j)=p_j$, $P(x)=p_x$, $P(H_j/x)=p_{jx}$ logaritmus věrohodnostní funkce upravit na tvar:

$$\ln L = \text{konst.} + \sum_j \sum_x n_{jx} \ln p_{jx}(a) + \sum_x n_x \ln p_x .$$

Uvedená aditivní konstanta obsahuje členy nezávislé na parametrech a a p_x. Pravděpodobnosti p_{jx} a p_x jsou přitom vázány podmínkami:

$$\sum_x p_x = 1 , \tag{1}$$

$$\sum_x p_{jx} p_x = p_j , \quad j=1,\ldots,m-1 . \tag{2}$$

Snadno se lze přesvědčit, že parciální derivace posteriorních pravděpodobností splňují vztah:

$$\frac{\partial p_{ix}}{\partial a_{jt}} = p_{ix}(\delta_{ij} - p_{jx})\, x_t , \quad \begin{array}{l} i,j=1,\ldots,m \\ t=0,\ldots,p \end{array}$$

Na základě toho můžeme upravit parciální derivace funkce ln L

$$\frac{\partial \ln L}{\partial a_{jt}} = \sum_x \sum_i n_{ix} \frac{1}{p_{ix}} \frac{\partial p_{ix}}{\partial a_{jt}} = \sum_x (n_{jx} - n_x p_{jx}) x_t .$$

Dále platí:

$$\frac{\partial \ln L}{\partial p_x} = \frac{n_x}{p_x} .$$

Užitím Lagrangeových multiplikátorů dospějeme k rovnicím:

$$\frac{\partial}{\partial a_{jt}} \left(\ln L + u \sum_x p_x + \sum_{k=1}^{m-1} v_k \sum_x p_{kx}\, p_x \right) = 0 , \qquad (3)$$

pro $j=1,\dots,m-1$, $t=0,\dots,p$,

$$\frac{\partial}{\partial p_x} \left(\ln L + u \sum_x p_x + \sum_{k=1}^{m-1} v_k \sum_x p_{kx}\, p_x \right) = 0 \qquad (4)$$

pro všechna x .

Z (4) dostáváme pro všechna x:

$$n_x + u\, p_x = - \sum_{k=1}^{m-1} v_k\, p_{kx}\, p_x . \qquad (5)$$

Postupnou úpravou rovnic (3) a dosazením předchozího:

$$\sum_x (n_{jx} - n_x p_{jx}) x_t + \sum_{k=1}^{m-1} v_k p_{kx} (\delta_{jk} - p_{jx}) p_x x_t = 0$$

$$\sum_x (n_{jx} - n_x p_{jx} + v_j p_{jx} p_x - p_{jx} \sum_{k=1}^{m-1} v_k p_{kx} p_x) x_t = 0$$

$$\sum_x (n_{jx} + (u+v_j)\, p_{jx} p_x)\, x_t = 0$$

Speciálně pro t=0 je:

$$\sum_x (n_{jx} + (u+v_j)\, p_{jx} p_x) = 0 ,$$

tj. pro $j=1,\dots,m-1$ platí:

$$n_j + (u+v_j)\, p_j = 0 .$$

Jestliže zvolíme priorní pravděpodobnosti $p_j=n_j/n$, bude pro $j=1,\dots,m-1$:

$$n + u + v_j = 0 \; .$$

Z rovností (5) dostaneme sumací s využitím (1) a (2) :

$$n + u + \sum_{k=1}^{m-1} v_k p_k = 0 \; ,$$

z čehož nutně plyne pro $j=1,\dots,m-1$

$$v_j = 0$$

a

$$u = -n \; .$$

Rovnice (3) a (4) se tedy při volbě $p_j=n_j/n$ zjednoduší na tvar

$$\sum_x (n_{jx} - n_x p_{jx}) \, x_t = 0$$

pro $j=1,\dots,m-1$, $t=0,\dots,p$ a

$$p_x = n_x/n$$

pro všechna x.

Druhé parciální derivace funkce $\ln L(a)$ můžeme na základě předchozího vyjádřit ve tvaru:

$$\frac{\partial^2 \ln L}{\partial a_{is} \partial a_{jt}} = - \sum_x n_x \, (\delta_{ij} \, p_{ix} - p_{ix} p_{jx}) \, x_s x_t \; .$$

Označme dále:

$$d(a) = \left(\frac{\partial \ln L}{\partial a_{jt}} \right) \quad \text{pro } j=1,\dots,m-1 \text{ a } t=0,\dots,p \; ,$$

$$D(a) = \left(\frac{\partial^2 \ln L}{\partial a_{is} \partial a_{jt}} \right) \text{ pro } i,j=1,\dots,m-1 \text{ a } s,t=0,\dots,p \; .$$

Newton-Raphsonovou metodou lze najít s požadovanou přesností kořeny věrohodnostních rovnic při počátečním odhadu

$$a^{(0)} = 0 \; .$$

Tento iterační postup můžeme vyjádřit ve tvaru:

$$a^{(k+1)} = a^{(k)} - (D(a^{(k)}))^{-1} \, d(a^{(k)}) \; .$$

Odhady $\hat{a}$ získané uvedenou metodou mají asymptoticky normální rozložení se střední hodnotou a . Asymptotickou kovarianční matici můžeme odhadnout maticí

$$-D(\hat{a})^{-1} .$$

Vhodnost modelu lze posoudit na základě statistiky

$$2 \sum_x \sum_i n_{ix} \ln \frac{n_{ix}}{n_x p_{ix}(\hat{a})} ,$$

která má asymptoticky rozložení chí-kvadrát o $N-(m-1)(p+1)$ stupních volnosti, kde N je počet různých hodnot x v populaci.

Uvažujme nyní obecné priorní pravděpodobnosti q_j $j=1,\ldots,m$. Musí ovšem platit pro všechna x a $j=1,\ldots,m$

$$P(x/H_j) = p_{jx}\, p_x \,/\, p_j = q_{jx}\, q_x \,/\, q_j ,$$

kde označení indexů u pravděpodobností q je analogické předchozímu. Další úpravou postupně dostaneme:

$$\frac{p_{jx}}{p_{mx}} \frac{p_m}{p_j} = \frac{q_{jx}}{q_{mx}} \frac{q_m}{q_j} \qquad \text{pro } j=1,\ldots,m-1$$

a logaritmováním

$$x'(b_j-a_j) = \ln q_j p_m / q_m p_j ,$$

kde vektory b_j odpovídají pravděpodobnostem q.
Speciálně pro $x=(1,0,\ldots,0)'$ platí

$$b_{j0} = a_{j0} + \ln q_j p_m / q_m p_j ,$$

takže pro $t=1,\ldots,p$ dostáváme:

$$b_{jt} = a_{jt} .$$

Lze tedy uzavřít, že obecné priorní pravděpodobnosti ovlivní aditivně pouze člen svázaný s proměnnou x_0. Ostatní koeficienty jsou vůči volbě priorních pravděpodobností invariantní.

Literatura:

Anderson, J.A., Logistic Discrimination with Medical Applications In: Discriminant Analysis and Applications (ed. T.Cacoullos) Academic Press, N.York, 1973