Het ALLOC pakket voor verdelingsvrije Discriminant Analyse

door
J.D.F. Habbema[1,2]
J. Hermans[1]
J. Remme[1]

Het uitgangspunt van discriminant analyse is als volgt:
- k verschillende populaties (groepen, categorieën) zijn gegeven, $A_1, \ldots, A_k$.
- aan elk steekproef element (object) zijn p kenmerken (variabelen) te meten: $x = (x_1, x_2, \ldots, x_p)$
- de kansverdeling van x zal in het algemeen per populatie verschillend zijn: $f(x|A_j)$ $j = 1, \ldots, k$.
- uit elke populatie A_j is een steekproef van omvang N_j van trainings-elementen (leerpopulatie, referentie waarnemingen) gegeven, waaruit informatie geput moet worden betreffende de verdeling van x.

Het doel is de verschillen tussen de populaties met betrekking tot x zo goed mogelijk te beschrijven. Liefst zó, dat een nieuw element van onbekende herkomst aan de hand van zijn waarneming x zo goed mogelijk te klassificeren is.
Voor zo'n element worden daartoe de kansen uitgerekend dat het afkomstig zou zijn uit (zou behoren tot) populatie A_j (j= 1,..., k), gegeven zijn waarneming x. Deze achteraf- of posterior-kansen $P(A_j|x)$ kwantificeren de (on)zekerheid waarmee gesteld kan worden dat het element tot elk der populaties A_j (j= 1,..., k) behoort, gezien zijn waarneming x en aannemende dat het tot één der A_j's moet behoren. Het uitrekenen van deze achteraf-kansen geschiedt met de regel van Bayes:

$$P(A_j|x) = \frac{P(A_j)\, f(x|A_j)}{\sum_{i=1}^{k} P(A_i) f(x|A_i)} \qquad j=1,\ldots,k$$

Naast de kansverdelingen $f(x|A_j)$ komen in deze formule ook de vooraf- of prior-kansen $P(A_j)$ voor. Deze kwantificeren de (on)zekerheid waarmee, vóórdat de waarneming verricht is, gesteld kan worden dat een element tot een populatie A_j behoort. De vooraf-kansen zijn in sommige toepassingen goed te schatten; in andere toepassingen niet zo betrouwbaar te geven.

Van veel belang is verder de behandeling van de kansverdelingen $f(x|A_j)$. De specifieke kenmerken van het ALLOC pakket hebben ook juist betrekking op dit punt. Momenteel is het meest gebruikelijke te veronderstellen dat x een p-dimensionale normale verdeling volgt met in elke populatie een gelijke covariantiestructuur:

$$f(x|A_j) = N^{(p)}(x|\mu_j,\Sigma)$$

Bij de toepassing worden dan de populatie parameters μ_j en Σ vervangen door de steekproef gemiddelden $\bar{x}_j$ en gepoolde steekproefcovariantie matrix S, berekend uit de trainingselementen. Dit leidt tot de zogenaamde lineaire discriminant analyse. De programma's in SPSS, BMD en BMDP zijn hierop gebaseerd. In het ALLOC pakket is getracht de discriminant analyse uit te voeren met minder stringente verdelingsveronderstellingen om zodoende bij toepassingen wat minder risico te lopen op het niet goed vervuld zijn van deze veronderstellingen. Hiertoe wordt gewerkt met zogenaamde kernel functies of potentiaal functies. Het basisidee hiervan is alsvolgt. Per populatie wordt rond elk trainingselement y_{ij} een 'kernel' $K(x,y_{ij})$ met kansmassa $1/N_j$ gelegd in de p-dimensionale ruimte. Optelling van deze 'kernels' over de N_j trainingselementen geeft een kansdichtheid, die als schatting van $f(x|A_j)$ wordt gebruikt:

$$\hat{f}(x|A_j) = \frac{1}{N_j} \sum_{i=1.}^{N_j} K(x,y_{ij})$$

Voor een nieuw element zijn deze schattingen van de k dichtheden ter plaatse van zijn waarneming x uit te rekenen en leiden zij via de Bayes formule tot schattingen van de k achteraf-kansen. De thans beschikbare programma's hebben allen betrekking op continue kenmerken x. Deze 'kernel' aanpak geeft een tamelijk flexibel systeem voor het beschrijven van de variabiliteit, die optreedt bij de trainingselementen en is aanzienlijk minder restrictief dan de veronderstelling van multinormaliteit van de data. Asymptotisch beschouwd is de zaak heel eenvoudig. De kernel schatting convergeert naar de ware onderliggende dichtheid; de schatting met de normale kansdichtheid doet dit (met kans 1) niet. Een ons inziens aan te bevelen werkwijze bij toepassingen is de data met twee verschillende typen discriminant analyses te analyseren. Bij veel en/of grote verschillen tussen deze analyseuitkomsten, dient dan een nadere inspectie van de data uitgevoerd te worden.

Een uitvoerig manual is beschikbaar. Hierin wordt een algemene inleiding in de discriminant analyse gegeven, wordt de kernel methode zoals deze is geimplementeerd in de ALLOC programma's besproken en worden tenslotte de invoer, de uitvoer en de restricties van de programma's vermeld.

Het pakket telt thans drie programma's.

ALLOC-1. Hiermee kan een stapsgewijze, voorwaartse selectie van variabelen worden uitgevoerd. Het selectie criterium is de fractie fout geklassificeerde trainingselementen (error rate).
Een vergelijking van dit programma met de BMD(P) en SPSS programma's voor selectie van variabelen in discriminant analyse, vindt plaats in een artikel van Habbema en Hermans in Technometrics, vol. 19 (1977), november nummer.

ALLOC-2. Dit is geschikt indien men enkel geinteresseerd is in de toewijzing of klassificatie van een aantal nieuwe elementen. Het gebruik van kernel functies brengt met zich mee, dat wel alle data van de trainingselementen beschikbaar moeten zijn; deze worden evenwel niet geanalyseerd.

ALLOC-3. Dit is bestemd voor het analyseren van de trainingsdata, gegeven een vaste set van p variabelen. Hiermee is een beeld te krijgen van de mate van overlap tussen de k populaties en het discriminerend vermogen met deze set variabelen.

Op het ogenblik is het pakket in Nederland beschikbaar op de universitaire rekencentra van Leiden, Utrecht, Amsterdam (Technisch Centrum), en Nijmegen en op het rekencentrum van Unilever.

[1]Afd. Medische Statistiek, Universiteit Leiden, Wassenaarseweg 80, Leiden
[2]Afd. Maatschappelijke Gezondheidszorg, Erasmus Universiteit, Rotterdam