DISKRIMINANTANALÜÜS (*DA*)

X(N,M), i=1,...,N; j=1,...,M.

DA - mitmemõõtmelise analüüsi meetod, mis klassifitseerib objektid teatavatesse etteantud klassidesse.

***Eesmärk*: etteantud teadaoleva klassikuuluvusega objektidele tuginedes**

- **leida etteantud P tunnuse (diskriminanttunnuse) alusel (S=1,...,P, P<M) eeskiri (diskrimineeriv funktsioon), mis võimalikult hästi eristaks objektide klasse,**
- **selle eeskirja alusel otsustada tundmatute (uute) objektide klassikuuluvus.**

DA *jaguneb*:

a) meetodid, mis võimaldavad interpreteerida klasside erinevusi, s.t.

- kas tunnuste X_1,..., X_P abil on võimalik eristada eri klasse,

- kui hästi nad mõõdavad erinevust klasside vahel,

- millised tunnused on informatiivsemad;

b) meetodid, mis klassifitseerivad objektid etteantud klassidesse, s.t. saada 1 või mitu diskriminantfunktsiooni.

NB! Iga objekt võib kuuluda ainult ühte klassi.

DA *rakendamise eeldused*:

1) klasside arv $G \geq 2$,

2) igas klassis K, K=1,...,G, objektide arv $N_K \geq 2$,

3) diskriminanttunnuste arv $1 \leq P < N-2$,

4) diskriminanttunnused lineaarselt sõltumatud,

5) klasside kovaratsioonimaatriksid võrdsed,

6)$\forall$ klassis diskriminanttunnustel normaaljaotus

Diskriminantfunktsioon (DF)

$f = U_0 + U_1X_1 + U_2X_2 + ... + U_PX_P$, kus

U_S - koefitsient, S=1,...,P.

$f_{KR} = U_0 + U_1X_{1KR} + U_2X_{2KR} + ... + U_PX_{PKR}$, kus

f_{KR} - DF väärtus K-nda klassi R-nda objekti korral, K=1,...,G; R=1,...,N_K;

X_{SKR} - S-nda tunnuse väärtus K-nda klassi R-ndal objektil.

Arvutatavate diskriminantfunktsioonide arv D=min(G-1, P), seejuures

1) esimese DF (f_1) korral U_S väärtused sellised, et DF keskväärtused eri klasside korral võimalikult palju erineksid,
2) teise DF (f_2) korral sama, mis 1), lisaks f2 väärtused ei korreleeruks f1 väärtustega,
3) ...

DF interpreteerimine

X1 - vanus, X2 - perekonnaseis, X3 - sissetulek, X4 - teadustööde arv,
klassifitseeriv tunnus - pereliikmete arv (kuni 5).

Standardiseeritud DF koefitsiendid

	Tunnus	DF_1	DF_2	DF_3	DF_4
U_1	vanus	-0,05	-0,06	0,2	0,04
U_2	perek.seis	2,39	1,17	0,3	-0,44
U_3	sissetulek	-0,03	0,01	0,0	0,0
U_4	publik. arv	-0,03	0,05	-0,09	-0,09
U_0		1,83	-0,71	-7,65	-3,08

$f_1 = 1,83 - 0,05X_1 + 2,39X_2 - 0,03X_3 - 0,03X_4$

f_2 = ...

Klasside tsentroidid
(keskmed koordinaatteljestikus)

Klass	DF_1	DF_2	DF_3	DF_4
1	-0,2	-1,02	-0,04	0,01
2	-2,04	0,37	-0,18	-0,01
3	0,83	0,18	-0,43	-0,01
4	0,31	0,14	0,54	-0,01
5	0,13	1,03	0,02	0,06

Kuidas hinnata tunnuste panuseid

1) standartiseeritud DF koefitsiendid U_S: mida suurem, seda mõjusam

Näide.

	DF_1	DF_2	DF_3
1	0,61	-0,39	1,22
2	0,71	-0,99	-0,30
3	-2,20	-0,53	-0,54
4	-0,48	-0,90	0,78
5	-0,81	0,80	0,27
6	1,02	0,73	0,14

2) Struktuursed koefitsiendid

- korrelatsioon diskriminanttunnuste ja DF vahel

Näide.

	DF_1	DF_2	DF_3
1	-0,56	0,35	0,33
2	0,34	0,27	-0,43
3	-0,86	0,24	-0,16
4	0,27	-0,67	0,25
5	-0,30	0,78	0,08
6	-0,14	0,75	-0,27

NB! Korreleeritud diskriminanttunnuste korral tunnuste panused väga erinevad (eespoololija panus suurim), samal ajal korrelatsioonid DFga aga ei erine palju.

Mitme DF-ga arvestada?

Otsustamise mõõdud

1. DF omaväärtus.

Mida suurem, seda mõjusam:

- **osa nullilähedased,**
- **osa statistiliselt ebaolulised:**

H_0: omaväärtus ei erine oluliselt nullist.

Kontroll: Wilks'i Λ (lambda):

$\Lambda = {}_L\Pi\ 1/(1+\Lambda_L)$, L=K+1,..., G, kus

K - juba hinnatud DF-de arv,

O_L - L-nda DF (f_L) omaväärtus.

***Otsustamine*. Arvutuslik olulisus > 0,05 =>H_0.**

$\leq$ 0,05 =>H_1.

2. Kanooniline korrelatsioon

$r_L = \sqrt{\Lambda_L/(\Lambda_L+1)}$, kus

L- DF number (f_L).

H_0: $r_L = 0$

H_1: $r_L \neq 0$

***Otsustamine*. Arvutuslik olulisus > 0,05 =>H_0.**

$\leq$ 0,05 =>H_1.

3. Omaväärtuse suhteline panus.

$\%(f_L) = [\Lambda_L/(\Lambda_L / {}_L\Sigma\Lambda_L)]$ *100%

Näide.

f_L	Oma väärt	%	r_L	Wilks 'i Λ	Hii-ruut χ^2	df	Olulisus$_{ARV}$
1	1,01	63,4	0,71	0,30	26,9	16	0,04
2	0,41	25,7	0,54	0,61	11,3	9	0,26
3	0,17	10,9	0,38	0,85	3,6	4	0,46
4	0,0	0,03	0,02	0,99	0,01	1	0,91

Klassifitseerimine

Millisesse klassi tundmatu objekt DF alusel klassifitseerida?

Klassifikatsioonifunktsioon igale klassile:

$$I_{iK} = {}_S\sum a_{SK} X_{iS} + a_{0K} \text{ , } S = 1,...,P, \text{ kus}$$

I_{Ki} - i-nda objekti kaal K-nda klassi korral,

X_{iS} - S-nda tunnuse väärtus i-ndal objektil,

a_{0K} - klassi K klassif.funktsiooni konstant.

Klassifikatsioonikoefitsiendid

Tunnus	Klass 1	Klass 2	Klass 3	Klass 4	Klass 5
vanus	1,81	1,79	1,61		
perek.seis	-4,32	-7,12	-0,54		
sissetulek	0,06	0,07			
publik. arv	-0,94	-0,8			
konstant	35,86	-40,63			

$$I_1 = 35,86 + 1,81X_{i1} - 4,32X_{i2} + 0,06\ X_{i3} - 0,94X_{i4}$$

Näide.

Vanus - 40; perek.seis - 1 (abielus);
sissetulek - 800 krooni; publik.arv - 40.

$I_1 = 43,3$; $I_2 = 47,5$; $I_3 = 44,0$;
$I_4 = 43,3$; $I_5 = 47,1$;

Kuidas hinnata klassifitseerimise headust?

Millisesse klassi <u>valimi X(N,M) iga objekt</u> DF alusel klassifitseerida?
Klassifitseerimistabel

<u>Näide.</u>
N = 28

Klassifitseerimine DF alusel

		KL. 1	KL. 2	KL. 3	KL. 4	KL. 5
LÄH	KL. 1	6	0	0	0	0
TE	KL. 2	0	7	0	0	1
KLAS	KL. 3	0	0	5	0	0
SID	KL. 4	0	0	0	7	0
	KL. 5	0	0	0	0	2

Klassifitseerimise *täpsuse* (headuse) *mõõdud*:

T = (1-Y/N)*100% , kus

Y - ümberpaiknenud objektide arv.

N = 28; T = (1-1/28)*100% = 96,4%.