

ANALIZA DATELOR

Metode și Tehnici de cunoaștere a formelor

MTFR → Recunoaștere Ne-supravidă (sau analiză CLUSTER)

Rec. Supervizată

Metode de cunoaștere a formelor:

metode de laud filii:

îerarhice

algoritmi de partizionare

k-means

ascendentă

descendentă

Mhd. descendentă DIANA

ROMA (dacă date binare)

Recunoaștere Supervizată

cls. Bayesian

cls. Fisher

cls. Mahalanobis

că nu este apropiat vecinilor

k-nn (nu este apropiat vecinilor)

SVM (Support Vector Machine)

RNA

Probleme legate de clustere

- Evaluarea similarității sau disimilarității dintre formele

observații

- formarea --- (verticuri)

- evaluarea variabilității între-clasă și între-discriminator

- evaluarea puterii de discriminare a variabililor

Natările

C → centroid

AS → agregare simplă

AM → agregare medie

AT? → agregare complexă / totală

$n \rightarrow$ conțin variabile (ex $P(B), P_1, \dots, P_S$)

$T \rightarrow$ obs \rightarrow forme

k clase (clustere) notă cu $w = w_1, w_2, \dots, w_k$.

1). Distanță euclidiană $d_e(x, y) = \left[\sum_{i=1}^n (x_i - y_i)^2 \right]^{1/2}$

\rightarrow distanță linie dreaptă

1'). Distanță euclidiană ponderată $(x, y) =$

$d_{ep} = \left[\sum_i w_i (x_i - y_i)^2 \right]^{1/2} \quad \sum_i w_i = 1.$

2). Distanță Manhattan $d_m(x, y) =$

$= \sum_{i=1}^n |x_i - y_i|$

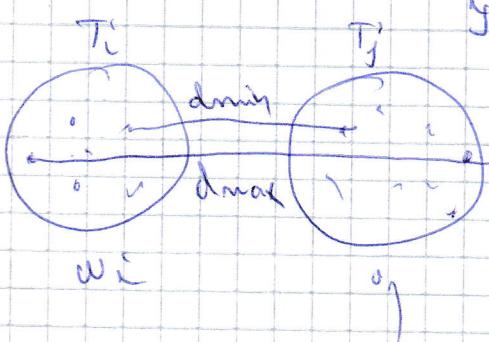
3). Distanță Mahalanobis:

$d_{MAH}(x, y) = (x - y)^T \cdot \Sigma^{-1} (x - y)$ $\begin{matrix} \text{vector} \\ \text{t - transpus.} \end{matrix}$

Evaluarea distanților dintre clustere:

1). Metoda celor mai apropiate vecini (agregare simplă).

$d(w_i, w_j) = \min_{\substack{x \in w_i \\ y \in w_j}} d(x, y)$



- distanță vecinilor cel mai apropiat

2). Metoda celor mai deosebite vecini (agregare complexă)

$d(w_i, w_j) = \max_{\substack{x \in w_i \\ y \in w_j}} d(x, y)$

Exemplu pur didactic.

Obs.	x_1	x_2
O_1	-0,62	-0,96
O_2	0,42	-0,88
O_3	1,42	0,59
O_4	0,34	1,4
O_5	-0,3	-0,92
O_6	0,12	-0,96

← matricea de proximitate

dE	O_1	O_2	O_3	O_4	O_5	O_6
O_1	0					
O_2	1,03	0				
O_3	2,55	1,77	0			
O_4	2,55	2,28	1,35	0		
O_5	0,52	0,72	2,29	2,41	0	
O_6	0,82	0,23	1,36	2,36	0,5	0

$$d(O_1, O_2) = 1,03 = \sqrt{(-0,62 - 0,42)^2 + (-0,96 - 0,88)^2} = \\ = 1,03$$

Aplicăm metoda centroidului

Criteriul general al clasificării: Clasile trebuie să fie cât mai aproape la interior și cât mai diferențiate la exterior. Variabilitatea întraclassă să fie cât mai mică și variabilitatea interclassă să

PAS 2

(Tab B)
Pe noua matrice obținută din nouă distanță
minimă, $\Rightarrow 0,32$

$$w_2 = \{0_3, 0_5\} = 0,32$$

$$\text{centroid } w_2(0_1, 0_5) = \left(\frac{-0,62 - 0,3}{2}, \frac{-0,96 - 0,92}{2} \right) \\ = (-0,46, -0,94).$$

Refacem matricea de proximitate luând în calcul
matricea formată.

de	w_1	w_2	0_3	0_4	
w_1	0				
w_2	0,77	0			
0_3	1,87	2,42	0		
0_4	2,31	2,47	1,35	0	

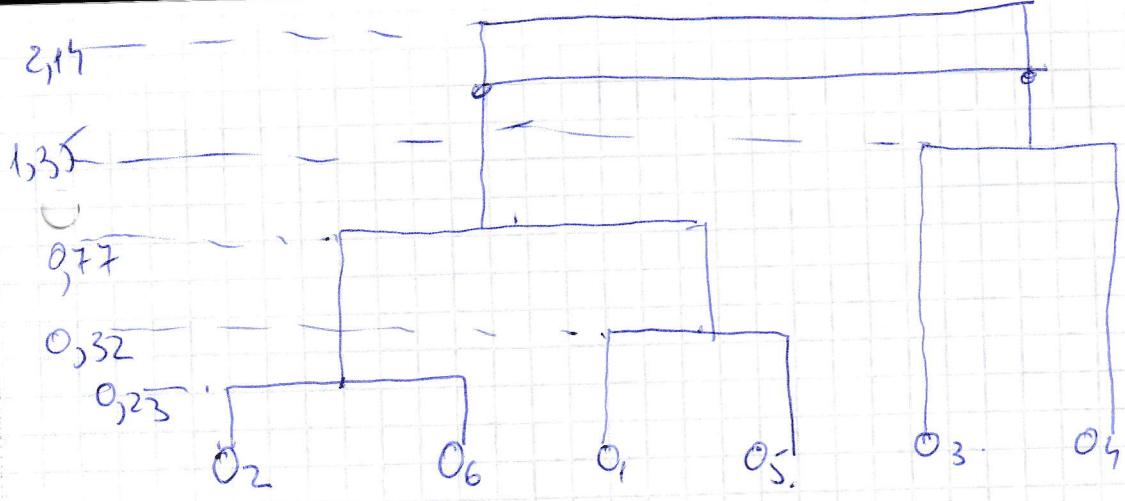
$$0,77 = d(w_1, w_2) = d(\text{centroid}(w_1), \text{centroid}(w_2)) =$$

$$= \sqrt{(0,21 + 0,46)^2 + (-0,92 + 0,94)^2} = 0,77.$$

PAS 3

w_1 și w_2 se unesc

$$w_1 = \{w_1^{\text{pas1}}, w_2\} = \{0_1, 0_2\} \\ = \{\overbrace{0_2, 0_6}, \overbrace{0_1, 0_5}\}$$



↗ DENDROGRAMA

Observăm că mai mare diferență între poziții 1,35 și 2,14, deci
însemnăm că limii \Rightarrow vom avea două clase

A treia problema: evaluarea variabilității interclasă și intraclassă.

a). caracteristica emică: \bar{x}

$$\begin{aligned} w_1, w_2, \dots, w_K &\quad \left. \right\} \Rightarrow SPA_T = \sum_{t=1}^T (x_t - \bar{x})^2 = \\ T_1, T_2, \dots, T_K &\quad \left. \right\} \\ &= \underbrace{\sum_{k=1}^K \sum_{t=1}^T (x_t^{(k)} - \bar{x}^{(k)})^2}_{SPA_W} + \underbrace{\sum_{k=1}^K T_k (\bar{x}^{(k)} - \bar{x})^2}_{SPA_B \text{ (between)}} \end{aligned}$$

SPA_W
(within
intra class).

SPA_W să fie cel mai mică

$SPA_B \rightarrow$ cel mai mare.

$$SPA_T = SPA_W + SPA_B \quad | : G_f = S_x^2 = S_w^2 + S_b^2$$

$G_f = \text{grade libertate}$

b). cas general m-variabile $(x_1, x_2 \dots x_n)$.

$$\begin{aligned}
 SPA_T &= \sum_{i=1}^n \sum_{t=1}^{T_i} (x_{ti} - \bar{x}_i)^2 = SPA_W \\
 &= \sum_{k=1}^K \sum_{i=1}^{n_k} \sum_{t=1}^{T_k} (x_{ti}^{(k)} - \bar{x}_i^{(k)})^2 + \\
 &\quad + \sum_{k=1}^K T_k \sum_{i=1}^{n_k} (\bar{x}_i^{(k)} - \bar{\bar{x}}_i)^2
 \end{aligned}$$

$$SPA_T = SPA_W + SPA_B \quad | : GL \Rightarrow$$

$$\Rightarrow \Sigma_T = \Sigma_W + \Sigma_B.$$

Revenim la problema:

$$\bullet \quad \omega_1 = \{o_1, o_2, o_3, o_4\}$$

$$\omega_2 = \{o_3, o_4\}$$

→ de făcut în Excel

ω_1	OBS	x_1	x_2	$(x_1 - \bar{x}_1)^2$	$(x_2 - \bar{x}_2)^2$
o_1					
o_2					
o_3					
o_4					
	medie	\bar{x}_1	medie	\bar{x}_2	medie
				0,66	0,94

$$\begin{aligned}
 SPA_K(\omega_1) &= 0,66 + 0,04 \\
 &= 0,664
 \end{aligned}$$

w_2	OBS	x_1	x_2	$(x_1 - \bar{x}_1)^2$	$(x_2 - \bar{x}_2)^2$
	O_3				
	O_4			0,584	0,329

$SPAW(w_2) = 0,913.$

Insumam $SPAW_{w_1}$ si $SPAW_{w_2}$

$$SPAW = 0,664 + 0,913 = 1,577$$

$$SPAT = 7,7$$

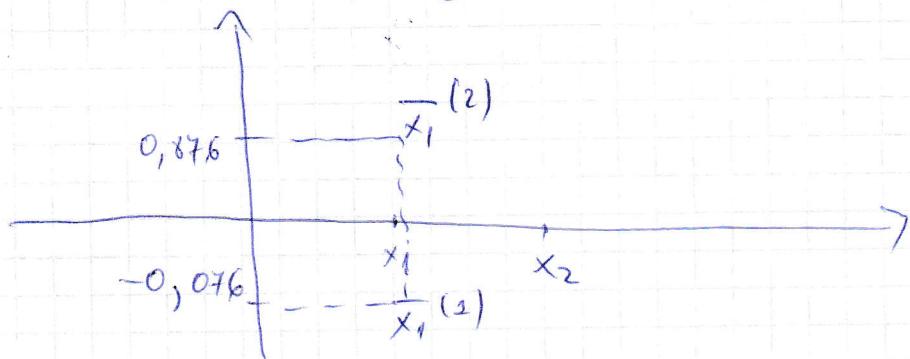
Facem un tabel ascendenta, dar punem și O_3, O_4 , tot cu diferențe.

Pentru diferență, vom afla $SPAb = 7,7 - 1,577 \approx 6,121$

A patra problemă: determinarea puterii de discriminare a variabilei

Două metode: se reprezintă grafic mediile variabilității fiecare dosar.

Variabilitatea între dosare este ceeață mai mare, cu cat mediile sunt mai diferențiate.



b). Pentru fiecare x_i ^{var} se calculează un raport,

$$x_i : \frac{SPA_b}{SPA_W} = R_{xi}$$

$R_{x_2} > R_{x_1}$ (în exemplul nostru).

DETALII EXAMEN

Proiect format din trei proiecte mai mici

alexandru.alixa@cncl.ase.ro

Pt. proiect date male reprezentate la $f_{\text{m}}, f_{\text{d}}, f_{\text{st}}$.

Observații : peste 30 ^{observații} (judecăți, sumă, etc.)

Variabile : o regulă nr. observații / 5 ≈ 6

judecăți

ACP mai poate fi folosită pentru eliminarea redundantei informaționale.

ANALIZA DISCRIMINANT

* Metode de recunoaștere a formelor de tip SUPERVIZAT

Construim clasele (numărul) și indivizi în că se încadrează (apartența obiectului).

Aveam o populație $\Omega \rightarrow$ set de antrenament (70%)
↓ set de testare (restul) 25%

Se presupune că formele sunt separate în K clase: w_1, w_2, \dots, w_K .

Multimea de clase acoperă în întregime populația, nu avem indivizi în afara categoriilor.

Clasele w au proprietăți $w_k \subseteq \Omega ; k \in \{1, 2, \dots, K\}$

$$w_1 \cup w_2 \cup w_3 \dots w_K = \Omega$$

Clasele sunt disjuncte dacă și doar dacă $w_i \cap w_j = \emptyset$
 $(\forall) i, j = 1, K, i \neq j$

În ceea ce constă modelului se urmărește obținerea unei clase predizate. Acestea sunt notate cu $\tilde{w}_1, \tilde{w}_2, \dots, \tilde{w}_K$.
Abilitatea predictivă
~~Cădutudinea~~ modelului poate fi evaluată prin intermediul matricei confundanții clasificării.

Matr. coecitudinii clasificării

Clasa reală

Clasa reală	Clasa predictată			
	\tilde{w}_1	\tilde{w}_2	...	\tilde{w}_k
w_1	T_{11}	T_{12}	...	T_{1k}
w_2	T_{21}	T_{22}	...	T_{2k}
:	:	:	...	:
w_k	T_{k1}	T_{k2}	...	T_{kk}
	$T_{\cdot 1}$	$T_{\cdot 2}$...	$T_{\cdot k}$

Elementele de pe diagonala principală T_{ii} reprezintă forme încastrate cînd în clasa predictată.

T_{ii} reprezintă numărul formelor existente în clasa reală.

$T_{\cdot j}$ numărul formelor ale căror clase predictate sunt \tilde{w}_j .