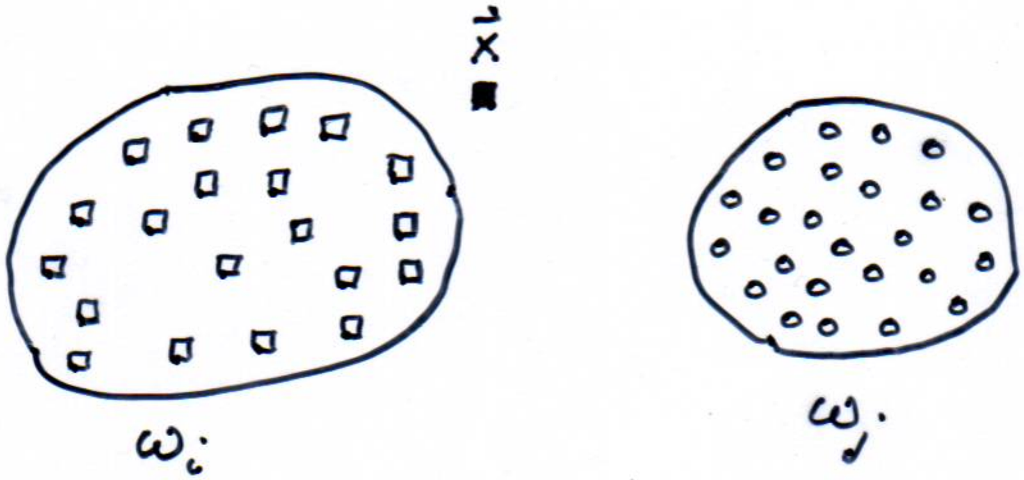
KLASIFIKACIJA UZORAKA POMOĆU FUNKCIJA UDALJENOSTI

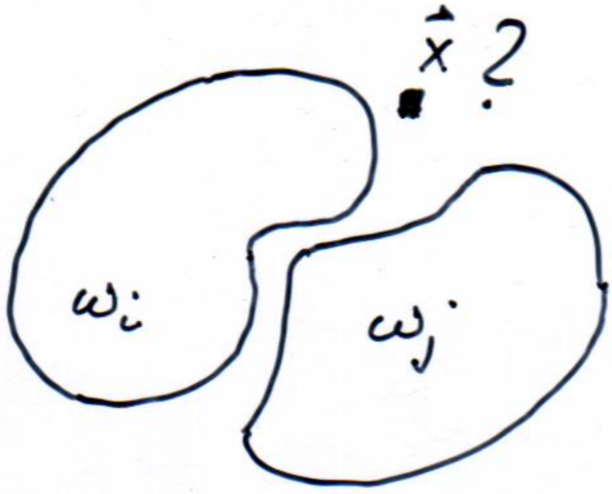
- jedna od najjednostavnijih metoda
- zasnovana na intuiciji

• mjera sličnosti između vektora uzoraka

vektor – točka u Euklidskom prostoru



- djelotvorna metoda kada razredi pokazuju svojstvo grupiranja



KLASIFIKACIJA POMOĆU NAJMANJE UDALJENOSTI

(MINIMUM - DISTANCE PATTERN CLASSIFICATION)

s. Ribarić

- mjera udaljenosti : bilo koja funkcija koja zadovoljava uvjete:

a) $D(\vec{x}_k, \vec{x}_l) = 0$ za $\vec{x}_k = \vec{x}_l$

$D(\vec{x}_k, \vec{x}_l) \neq 0$; $D(\vec{x}_k, \vec{x}_l) > 0$ za $\vec{x}_k \neq \vec{x}_l$

b) $D(\vec{x}_k, \vec{x}_l) = D(\vec{x}_l, \vec{x}_k)$

c) $D(\vec{x}_k, \vec{x}_l) \leq D(\vec{x}_k, \vec{x}_j) + D(\vec{x}_j, \vec{x}_l)$

primjer:

Euklidska udaljenost:

$$D = \| \vec{x}_k - \vec{x}_l \| = \sqrt{\sum_{i=1}^{n} (x_{ki} - x_{li})^2}$$

n - dimenzija vektora

Euklidska udaljenost - invarijantna na translaciju i rotaciju

Udaljenost Minkovskog:

$$D(\vec{x}_k, \vec{x}_l) = \left(\sum_{j=1}^{n} | x_{kj} - x_{lj} |^s \right)^{\frac{1}{s}}$$

- $s = 2 \rightarrow$ Euklidska udaljenost

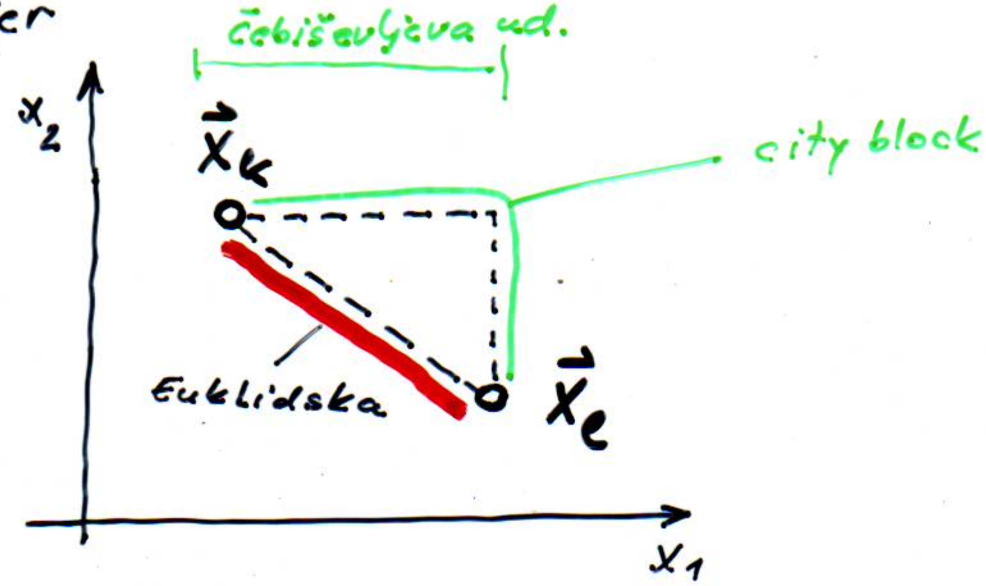
- $s = 1 \rightarrow$ Manhattan ili "city block" udaljenost

$$D(\vec{x}_k, \vec{x}_l) = \sum_{j=1}^{n} | x_{kj} - x_{lj} |$$

- $s \rightarrow \infty$ Čebiševljeva udaljenost

$$D(\vec{x}_k, \vec{x}_l) = \max \{ | x_{kj} - x_{lj} | \}$$

Primjer



S. Ribarić

- težinska udaljenost Minkovskog:

$$D(\vec{x}_k, \vec{x}_l) = \left(\sum_{j=1}^{n} w_j \cdot |x_{kj} - x_{lj}|^s \right)^{1/s}$$

w_j - težina pojedine značajke

$w_j \geq 1$

- Mahalanobisova udaljenost:

$$D(\vec{x}_k, \vec{x}_l) = (\vec{x}_k - \vec{x}_l)^T \mathbf{C}^{-1} (\vec{x}_k - \vec{x}_l)$$

$\mathbf{C}$ - kovarijacijska matrica dobivena iz S_N ; S_N - skup uzoraka za učenje

$\mathbf{C} = I$?

- Normirani korelacijski koeficijenti:

$$C(\vec{x}_k, \vec{x}_l) = \frac{\sum_{j=1}^{n} (x_{kj} - m)(x_{lj} - m)}{\sqrt{\sum_{j=1}^{n} (x_{kj} - m)^2 \sum_{j=1}^{n} (x_{lj} - m)^2}},$$

gdje je m srednja vrijednost značajke x_j.

S. Ribarić

"Hi kvadrat" udaljenost

$$D(\vec{X}_k, \vec{X}_l) = \sum_{j=1}^{n} \frac{1}{X_{\bullet j}} \left(\frac{X_{kj}}{X_{k\bullet}} - \frac{X_{lj}}{X_{l\bullet}} \right)^2$$

pri čemu su:

$$X_{\bullet j} = \sum_{k=1}^{N} X_{kj} \qquad X_{k\bullet} = \sum_{j=1}^{n} X_{kj}$$

i $X_{l\bullet} = \sum_{j=1}^{n} X_{lj}$; N je broj uzoraka u skupu za učenje S_N

- mjere sličnosti za binarne uzorke

$$\text{sličnost} = D^{-1}(\vec{x}_k, \vec{x}_l)$$

- $\vec{x}_k$ i $\vec{x}_l$ opisani binarnim značajkama

- udaljenost između binarnih uzoraka :

Hammingova udaljenost

$$D(\vec{x}_k, \vec{x}_l) = \sum_{j=1}^{n} d_j$$

$$d_j = \begin{cases} 1, & \text{ako je } x_{kj} \neq x_{lj} \\ 0, & \text{ako je } x_{kj} = x_{lj} \end{cases}$$

- težinska Hammingova udaljenost

$$D(\vec{x}_k, \vec{x}_l) = \sum_{j=1}^{n} w_j d_j$$

$w_j \geq 1$ - težina pojedinih značajki

$\vec{x}_k$ i $\vec{x}_l$ – opisani binarnim značajkama

		x_k = 1	x_k = 0
x_l	1	a	b
	0	c	d

Russelova i Raova sličnost:

$$s(\vec{x}_k, \vec{x}_l) = \frac{a}{a+b+c+d}$$

Jaccardova i Needhamova sličnost:

$$s(\vec{x}_k, \vec{x}_l) = \frac{a}{a+b+c}$$

Sokalova i Sneathova sličnost

$$s(\vec{x}_k, \vec{x}_l) = \frac{a}{a+2(b+c)}$$

Tanimotova sličnost:

$$s(\vec{x}_k, \vec{x}_l) = \frac{a+d}{a+d+2(b+c)}$$

Korelacija:

$$s(\vec{x}_k, \vec{x}_l) = \frac{ad+bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

S. Ribarić

Primjer:

$$\vec{x}_1 = (0, 1, 1, 0, 0, 1, 1, 1, 0, 1)$$

$$\vec{x}_2 = (0, 1, 1, 1, 0, 1, 0, 0, 1, 1)$$

		x_1 = 1	x_1 = 0
x_2	1	4 a	2 b
	0	2 c	2 d

$a = 4 \quad b = 2$

$c = 2 \quad d = 2$

Tanimotova sličnost:

$$S(\vec{x}_1, \vec{x}_2) = \frac{a+d}{a+d+2(b+c)} = \frac{6}{4+2+2(2+2)}$$

$$S(\vec{x}_1, \vec{x}_2) = \frac{6}{14} = 0.429$$

S. Ribarić

A) KLASIFIKACIJA NA TEMELJU JEDNOG PROTOTIPA - KARAKTERISTIČNOG PREDSTAVNIKA POJEDINOG RAZREDA

- uzorci teže grupiranju oko tipičnog ili reprezentativnog uzorka za pojedini razred
(npr. čitanje čekova)

M - razreda

prototipovi $\vec{z}_1, \vec{z}_2, \ldots, \vec{z}_M$!

Euklidova udaljenost

$\vec{x}$ i i-tog prototipa $\vec{z}_i$

$$D_i = \| \vec{x} - \vec{z}_i \| = \sqrt{(\vec{x} - \vec{z}_i)'(\vec{x} - \vec{z}_i)}$$

PRAVILO:

$$\boxed{\vec{x} \in \omega_i \quad \text{ako} \quad D_i < D_j \quad \text{za sve } j \neq i}$$

$$D_i^2 = \| \vec{x} - \vec{z}_i \|^2 = (\vec{x} - \vec{z}_i)'(\vec{x} - \vec{z}_i) \qquad ' = T$$

$$= \vec{x}'\vec{x} - 2\vec{x}'\vec{z}_i + \vec{z}_i'\vec{z}_i$$

$$= \vec{x}'\vec{x} - 2\left(\vec{x}'\vec{z}_i - \frac{1}{2}\vec{z}_i'\vec{z}_i\right)$$

$$\left.\begin{matrix} \min D_i^2 \\ \min D_i \end{matrix}\right\} \text{budući da su sve udaljenosti pozitivne}$$

$\vec{x}'\vec{x}$ - ne zavisi od i

maksimum $\left(\vec{x}'\vec{z}_i - \frac{1}{2}\vec{z}_i'\vec{z}_i\right)$!!

$D_i = \min$ kod je

S. Ribarić

decizijska funkcija:

$$d_i(\vec{x}) = \vec{x}'\vec{z}_i - \frac{1}{2}\vec{z}_i'\vec{z}_i \qquad i=1,2,\dots,M$$

$$\boxed{\vec{x} \in \omega_i \quad \text{ako je } d_i(\vec{x}) > d_j(\vec{x}) \text{ za sve } j \neq i}$$

$d_i(\vec{x})$ – linearna decizijska funkcija

$z_{ij} \quad j=1,2,\dots,n$ — komponente $\vec{z}_i$.

$w_{ij} = z_{ij} \qquad j=1,2,\dots n$

$w_{i,n+1} = -\frac{1}{2}\vec{z}_i'\vec{z}_i$

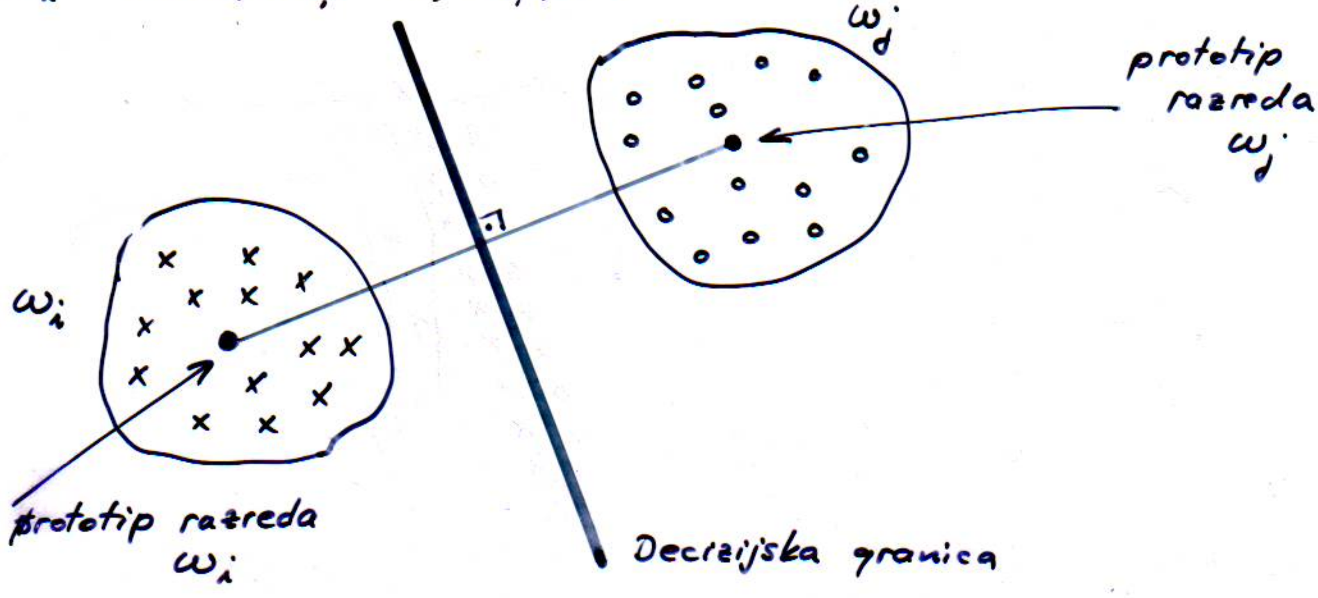
i

$$\vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \\ 1 \end{pmatrix}$$

$$\boxed{d_i(\vec{x}) = \vec{W}_i'\vec{x} \qquad i=1,2,\dots,M}$$

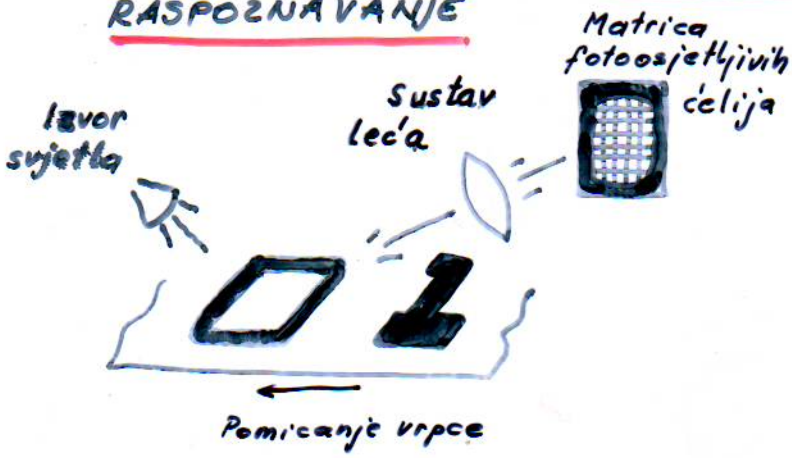
korelacijska metoda (correlation) ili cluster matching

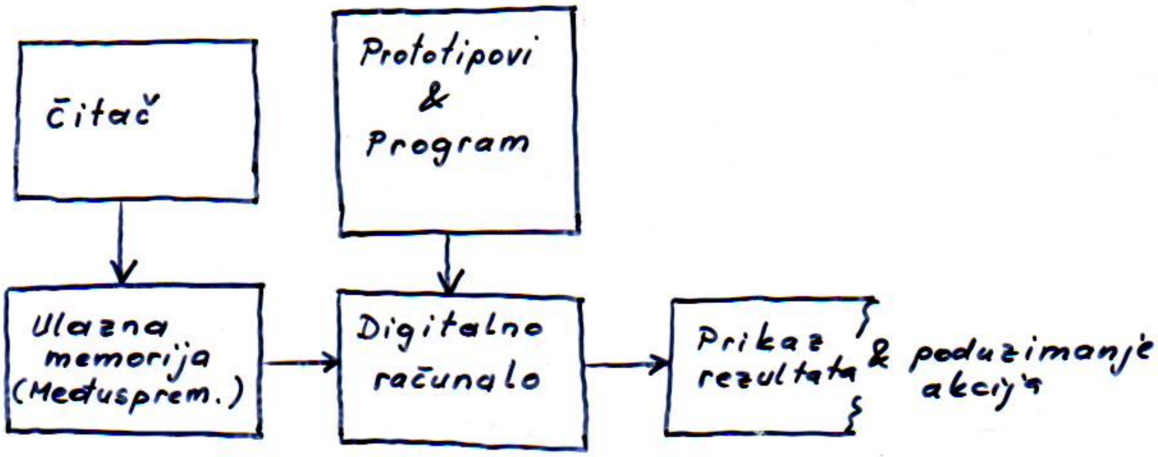
$\vec{W}_i = (w_{i1}, w_{i2} \dots, w_{i,n+1})'$



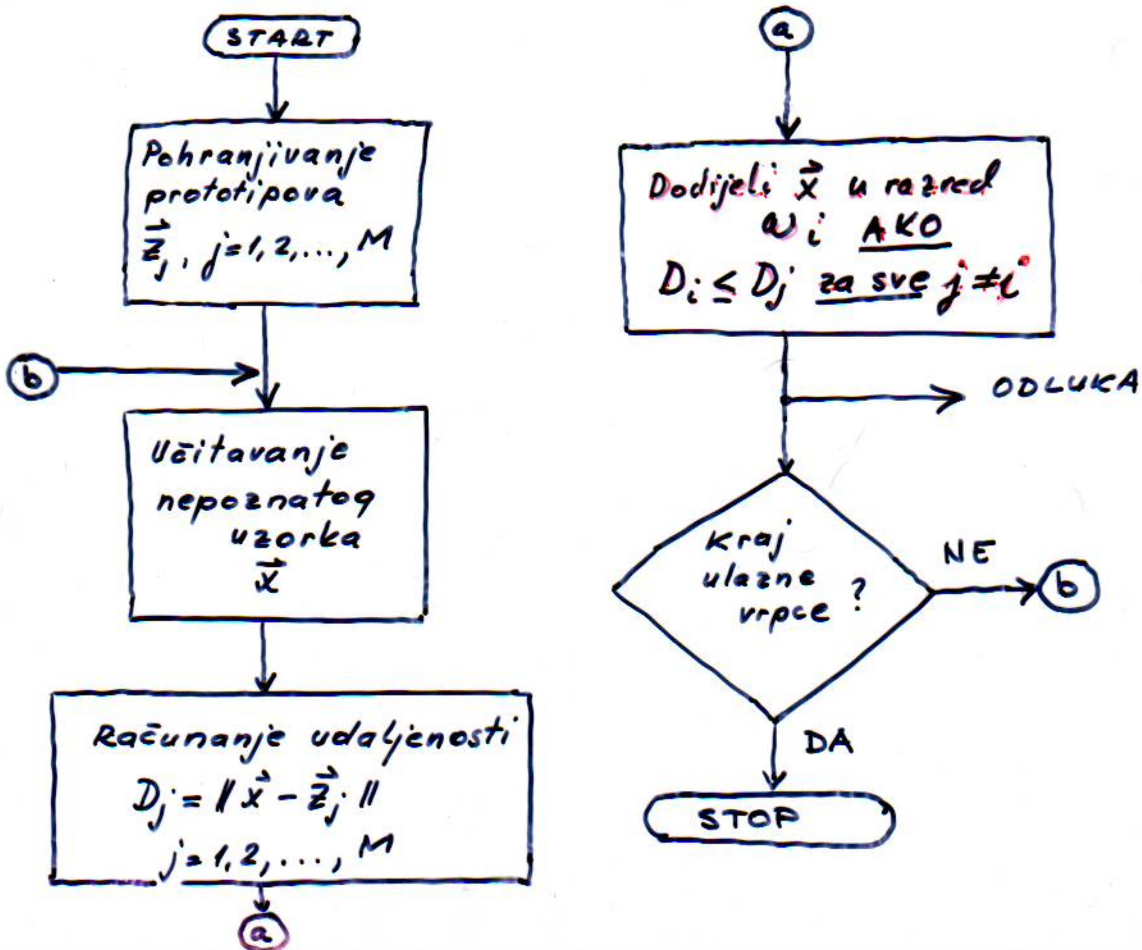
S. Ribarić

PRIMJER IZVEDBE SUSTAVA ZA RASPOZNAVANJE





Organizacija sustava



S. Ribarić

B) KLASIFIKACIJA NA TEMELJU VIŠE PROTOTIPOVA ZA POJEDINI RAZRED

- razred ω_i grupira se oko N_i prototipova

$$\vec{z}_i^{\,1}, \vec{z}_i^{\,2}, \dots, \vec{z}_i^{\,N_i}$$

N_i – broj prototipova za i-ti razred

$$D_i = \min_{l} \| \vec{x} - \vec{z}_i^{\,l} \| \qquad l = 1, 2, \dots, N_i$$

D_i – najmanja udaljenost između $\vec{x}$ i prototipa razreda ω_i.

D_i – izračunati $i = 1, 2, \dots, M$ i

nepoznati uzorak $\vec{x}$ klasificirati u

$$\boxed{\vec{x} \in \omega_i \text{ ako } D_i < D_j \text{ za sve } j \neq i}$$

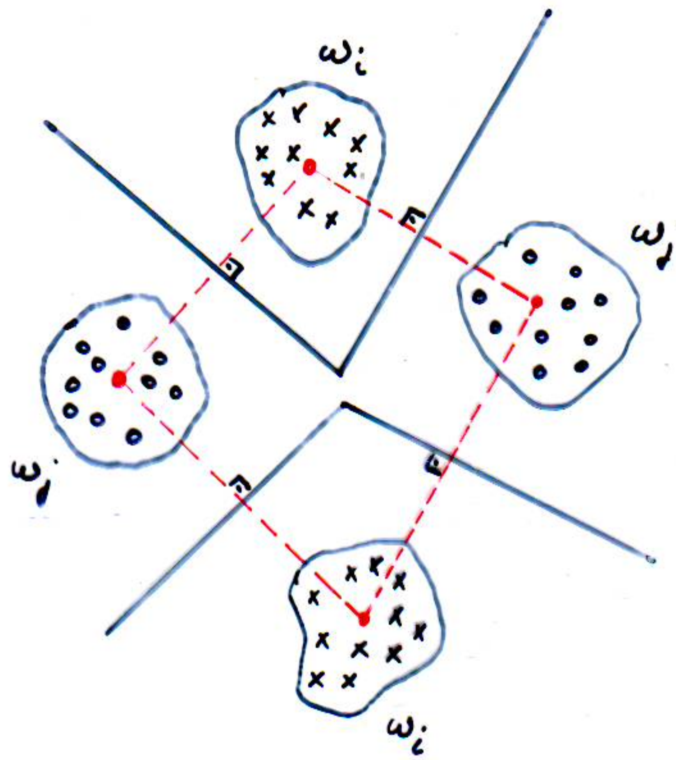
Decizijska funkcija:

$$d_i(\vec{x}) = \max_{l} \{ (\vec{x}' \vec{z}_i^{\,l}) - \tfrac{1}{2} (\vec{z}_i^{\,l})' \vec{z}_i^{\,l} \} \qquad l = 1, 2, \dots, N_i$$

$$\boxed{\vec{x} \in \omega_i \text{ ako je } d_i(\vec{x}) > d_j(\vec{x}) \text{ za sve } j \neq i}$$

S. Ribarić

Primjer :



piecewise - linear classifier :

$$d_i(\vec{x}) = \max_{\ell} \{ d_i^{\ell}(\vec{x}) \} \quad \ell = 1, 2, \dots, N_i \,, \quad i = 1, 2, \dots, M$$

$$d_i^{\ell}(\vec{x}) = w_{i1}^{\ell} x_1 + w_{i2}^{\ell} x_2 + \dots + w_{in}^{\ell} x_n + w_{i,n+1}^{\ell}$$

$$= (\vec{w}_i^{\ell})' \vec{x}$$

- opće iterativne metode (algoritmi) za izračunavanje parametara linearne decizijske funkcije

- opći algoritam za piecewise-linear slučajeve nije poznat!

PROŠIRENJE KONCEPTA KLASIFIKACIJE NA TEMELJU NAJMANJE UDALJENOSTI

- skup uzoraka s poznatom klasifikacijom

  $\{\vec{s}_1, \vec{s}_2, \dots, \vec{s}_N\}$

- svaki uzorak pripada jednom od razreda

  $\omega_1, \omega_2, \dots, \omega_M$

Klasifikacijsko pravilo NAJBLIŽEG SUSJEDA (Nearest Neighbor (NN); 1-NN) (E. Fix, J. L. Hodges, 1951.)

Nepoznati uzorak $\vec{x}$

$\vec{x} \in \omega_k$ ako je $\vec{s}$ najbliži susjed iz razreda ω_k

$$D(\vec{s}_i, \vec{x}) = \min_{\ell} \{ D(\vec{s}_\ell, \vec{x}) \}, \quad \ell = 1, 2, \dots, N$$

$\vec{s}_i \in \{\vec{s}_1, \vec{s}_2, \dots, \vec{s}_N\}$

D - mjera udaljenosti definirana u prostoru značajki

1-NN pravilo — za klasifikaciju koristi samo najbližeg susjeda

q - NN pravilo - za klasifikaciju

koristi q najbližih susjeda; $q > 1$

- na temelju N uzoraka za učenje naći q najbližih susjeda uzorku $\vec{x}$;
  /izaberi q neparan za $M=2$ /broj razreda/ ili neka $q > 1$ ne bude višekratnik broja razreda M/
- na temelju q uzoraka utvrdi broj vektora q_i koji pripadaju razredu ω_i, $i = 1, 2, \ldots, M$.

  Vrijedi $$\sum_i q_i = q$$
- razvrstaj $\vec{x}$ u razred ω_k za koji vrijedi da je q_k maksimalan!

Mogu se koristiti različite mjere udaljenosti, npr. Euklidska udaljenost, Mahalanobisova udaljenost, Čebiševljeva udaljenost, ...

Mahalanobisova udaljenost :

$$D(\vec{x}_k, \vec{x}_l) = (\vec{x}_k - \vec{x}_l)^T C^{-1} (\vec{x}_k - \vec{x}_l)$$

gdje je C kovarijantna matrica uzoraka iz skupa za učenje

$$C = \frac{1}{N} \sum_{i}^{N} (\vec{x}_i - \vec{m}_x)(\vec{x}_i - \vec{m}_x)^T$$

Mahalanobisova udaljenost uzima u obzir korelacije između značajki uzoraka. Ako je $C = I$ onda je Mahalanobisova udaljenost jednaka kvadratu Euklidske udaljenosti

Ocjena pogreške

- vjerojatnost pogreške klasifikatora NN za $N \to \infty$

P_{NN} :

$$P_B \le P_{NN} \le P_B \left(2 - \frac{M P_B}{M-1}\right) \le 2P_B$$

M - broj razreda

P_B - optimalna Bayesova pogreška

- Pogreška NN klasifikatora je (asimptotički) najviša dvostrukoj pogrešci optimalnog klasifikatora

S. Ribarić

Asimptotski $\Rightarrow$ performansa qNN klasifikatora bolja je od NN klasifikatora.

Npr. za $M=2$

$$P_B \le P_{qNN} \le P_B + \sqrt{\frac{2P_{NN}}{q}}$$

ako $q \to \infty$ performansa qNN klasifikatora teži optimalnoj.

- za veliki N i male Bayesove pogreške očekuje se npr. da za 3NN klasifikator dobijemo performansu skoro jednaku onoj Bayesovog klasifikatora:

$$P_{3NN} \approx P_B + 3(P_B)^2$$

za $P_B \ll 1$ i $N \gg 1$

$$P_{3NN} \dot{\approx} P_B$$

S. Ribarić

Pravila klasifikacije NN i qNN su za slučaj kada je skup uzoraka za učenje vrlo velik → vrlo djelotvorna

PROBLEM: kompleksnost izračuna udaljenosti i traženja q (ili 1) najbližih udaljenosti !

- Ocjena složenosti $O(kN)$ traženja najbližih susjeda
- Problem je još izraženiji za veliku dimenzionalnost prostora značajki $n >> 1$

Rješenje problema:

- predstavljanje razreda karakterističnim uzorkom (!?)
- uređivanje uzoraka iz skupa za učenje (npr. izlučivanje uzoraka jednog razreda koji se miješaju u prostoru značajki s uzorcima iz drugog razreda)
  (N. Pavešić, 1992)
- postupci zgušćivanja
- vektorska kvantizacija
  (S. Theodoridis, K. Koutroumbas, 2006)

Inačice metode NN

(q, l)-NN pravilo (M. Hellman, 1970)

pretpostavlja da se u skupu q najbližih susjeda trebaju pojaviti barem l uzoraka koji pripadaju nekom razredu ω_i

npr. $l = (2/3)\,q$

- u suprotnom uzorak se ne klasificira ili ga pokušavamo razvrstati pomoću nekog drugog pravila

(q, l_i)-NN (P.A. Devijer, 1977)

l_i ; $l_1 = 2/3\, q$, ... $l_M = 3/4\, q$

l_i - se određuje na temelju apriorne vjerojatnosti razreda uzoraka

apriorne vjerojatnosti razreda uzoraka – različite !

Taksonomija klasifikacijskih metoda na temelju udaljenosti

- s obzirom na način zapisa uzoraka iz skupa za učenje:
  - u memoriji klasifikatora pohranjeni su SVI uzorci iz skupa za učenje;
  - u memoriji su pohranjeni samo karakteristični predstavnici razreda

- s obzirom na broj (najbližih) susjeda:
  - 1 najbliži susjed
  - q najbližih susjeda; $q > 1$
  - (q, l)- i (q, l_i)-NN