# Klasifikacija uzoraka na temelju udaljenosti

Prof. dr. sc. Slobodan Ribarić

Uvod u raspoznavanje uzoraka

ak. god. 2019/2020

# Klasifikacija uzoraka na temelju udaljenosti

- Prostor značajki – metrički prostor s definiranom funkcijom udaljenosti
- Jedna od najjednostavnijih metoda klasifikacije

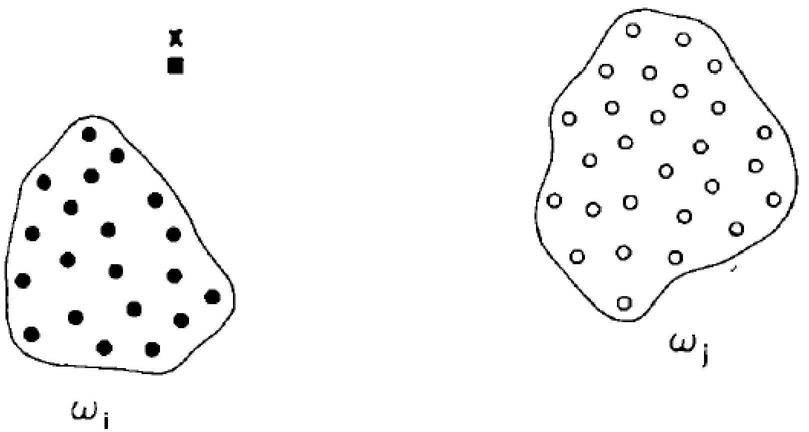
  - intuitivna metoda gdje je udaljenost mjera sličnosti između vektora uzoraka u prostoru značajki

  - vektori uzoraka – točke u Euklidskom prostoru

    slični uzorci se nalaze u neposrednoj blizini

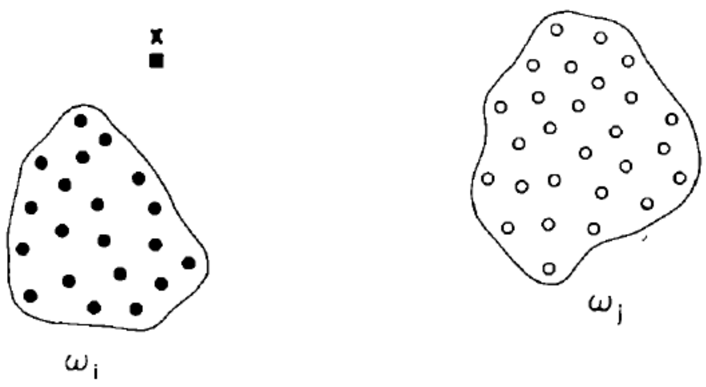
U koji biste razred klasificirali uzorak **x**?

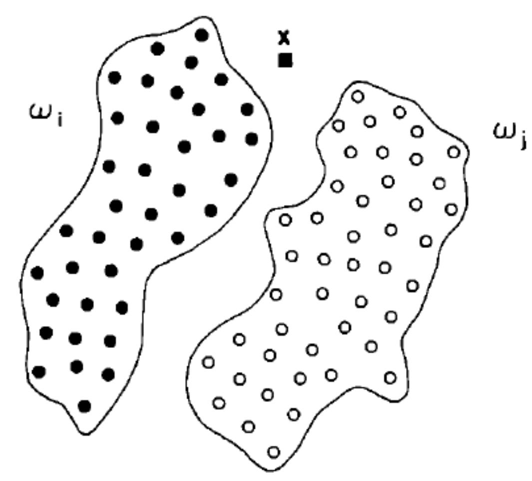
**-** kriterij- neposredna blizina!



# Klasifikacija uzoraka na temelju udaljenosti

- Postupak djelotvoran i daje zadovoljive rezultate samo onda kada uzorci pojedinih razreda pokazuju svojstvo grupiranja (npr. predstavljaju višedimenzionalne kugle)




Uzorak **x** nije moguće lako klasificirati na temelju **neposredne blizine**

# Klasifikacija uzoraka na temelju udaljenosti

- Neposredna blizina:
  – mala udaljenost između uzoraka → velika sličnost uzoraka

**Klasifikacija uzoraka na temelju minimalne udaljenosti**

**(Minimum-distance Pattern Classification)**

- Mjera udaljenosti – bilo koja funkcija koja zadovoljava sljedeće:

a) $$D(\mathbf{x}_\mathrm{k}, \mathbf{x}_l) = 0 \text{ za } \mathbf{x}_\mathrm{k} = \mathbf{x}_l$$

$$D(\mathbf{x}_\mathrm{k}, \mathbf{x}_l) \neq 0 \text{ i } D(\mathbf{x}_\mathrm{k}, \mathbf{x}_l) > 0 \text{ za sve } \mathbf{x}_\mathrm{k} \neq \mathbf{x}_l$$

b) $$D(\mathbf{x}_\mathrm{k}, \mathbf{x}_l) = D(\mathbf{x}_l, \mathbf{x}_\mathrm{k})$$

c) $$D(\mathbf{x}_\mathrm{k}, \mathbf{x}_l) \leq D(\mathbf{x}_\mathrm{k}, \mathbf{x}_j) + D(\mathbf{x}_j, \mathbf{x}_l)$$

# Klasifikacija uzoraka na temelju udaljenosti

Udaljenost Minkowskog:

$$D(\mathbf{x}_{\mathrm{k}}, \mathbf{x}_{l}) = \left( \sum_{j=1}^{n} \left| x_{\mathrm{kj}} - x_{l\mathrm{j}} \right|^{s} \right)^{\frac{1}{s}}$$

- za s= 2: Euklidska udaljenost /$L_2$ distance or $\ell$ 2 norm/

$$D(\mathbf{x}_{\mathrm{k}}, \mathbf{x}_{l}) = \left( \sum_{j=1}^{n} \left| x_{\mathrm{kj}} - x_{l\mathrm{j}} \right|^{2} \right)^{\frac{1}{2}}$$

n- dimenzionalni vektor značajki

Euklidska udaljenost – invarijantna na rotaciju I translaciju

$$D(\mathbf{x}_{\mathrm{k}}, \mathbf{x}_{l}) = \sqrt{(x_{\mathrm{k1}} - x_{l1})^2 + (x_{\mathrm{k2}} - x_{l2})^2 + \cdots + (x_{\mathrm{kn}} - x_{ln})^2}$$

# Klasifikacija uzoraka na temelju udaljenosti

- za s = 1: Manhattan ili “block-city” udaljenost / $L_1$ distance, or $\ell$ 1 norm /

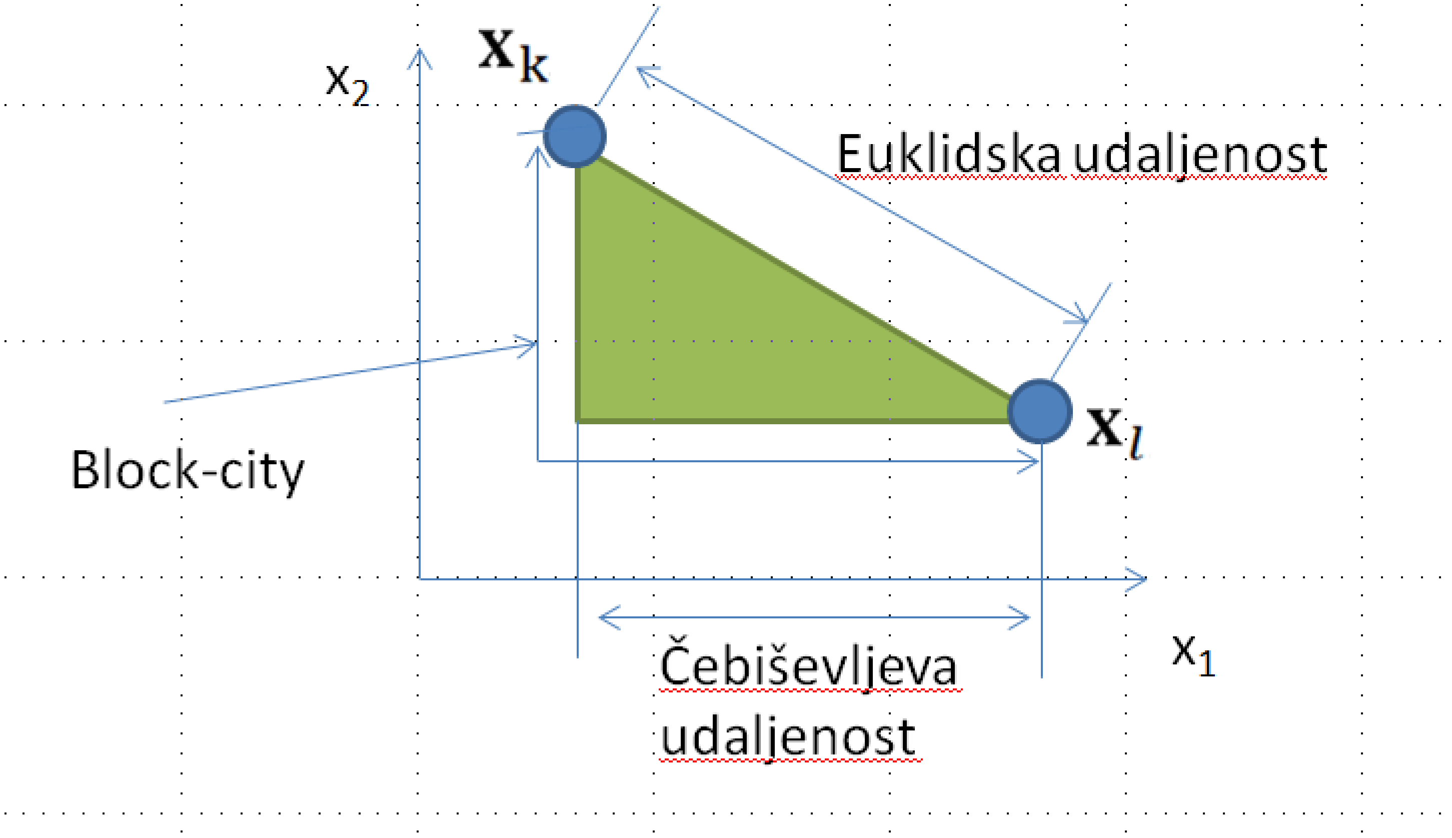
$$D(\mathbf{x}_{\mathrm{k}}, \mathbf{x}_{l}) = \sum_{j=1}^{n} |x_{\mathrm{kj}} - x_{l\mathrm{j}}|$$

- za s→ ∞: Čebiševljeva udaljenost / $L_\infty$ metric /

$$D(\mathbf{x}_{\mathrm{k}}, \mathbf{x}_{l}) = \max\{|x_{\mathrm{kj}} - x_{l\mathrm{j}}|\}$$

# Klasifikacija uzoraka na temelju udaljenosti

- Ilustracija udaljenost za n = 2



# Klasifikacija uzoraka na temelju udaljenosti

Težinska udaljenost Minkowskog

$$D(\mathbf{x}_{\mathrm{k}}, \mathbf{x}_{l}) = \left( \sum_{j=1}^{n} w_j \left| x_{\mathrm{kj}} - x_{l\mathrm{j}} \right|^{s} \right)^{\frac{1}{s}}$$

$w_j$ - težina pojedine značajke

$w_j \geq 1$

Mahalonobisova udaljenost

$$D(\mathbf{x}_{k}, \mathbf{x}_{l}) = (\mathbf{x}_{k} - \mathbf{x}_{l})^{T} \mathbf{C}^{-1} (\mathbf{x}_{k} - \mathbf{x}_{l})$$

$\mathbf{C}$ - kovarijacijska matrica dobivena na temelju skupa za učenje $S_N$

Što ako je $\mathbf{C}$ = I?

# Klasifikacija uzoraka na temelju udaljenosti

Pearsonov korelacijski koeficijent

$$Cor_{nor}(\mathbf{x}_k, \mathbf{x}_l) = \frac{\sum_{j=1}^{n}(x_{kj} - m_k)(x_{lj} - m_l)}{\sqrt{\sum_{j=1}^{n}(x_{kj} - m_k)^2 \sum_{j=1}^{n}(x_{lj} - m_l)^2}}$$

gdje su $\mathbf{x}_k$ i $\mathbf{x}_l$ $n$ – dimenzionalni vektori i

$$m_k = \frac{1}{n}\sum_{i=j}^{n} x_{kj}$$

$$m_l = \frac{1}{n}\sum_{i=j}^{n} x_{lj}$$

# Klasifikacija uzoraka na temelju udaljenosti

- **Mjere sličnosti za binarne uzorke**

$$sličnost \sim D^{-1}(\mathbf{x}_k, \mathbf{x}_l)$$

Vektori $\mathbf{x}_k$ i $\mathbf{x}_l$ - opisani binarnim značajkama (komponente vektora: 0 i 1

**Udaljenost između binarnih uzoraka:**

- Hammingova udaljenost:

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

$$\mathbf{x}_k = \begin{bmatrix} x_{k,1} \\ x_{k,2} \\ \vdots \\ x_{k,n} \end{bmatrix} \quad \mathbf{x}_l = \begin{bmatrix} x_{l,1} \\ x_{l,2} \\ \vdots \\ x_{l,n} \end{bmatrix}$$

$$x_{k,i} \in \{0, 1\}$$

$$x_{l,i} \in \{0, 1\}$$

$$d_j = \begin{cases} 1, ako\ je\ \mathrm{x}_{kj} \neq \mathrm{x}_{lj} \\ 0, ako\ je\ \mathrm{x}_{kj} = \mathrm{x}_{lj} \end{cases}$$

# Klasifikacija uzoraka na temelju udaljenosti

$$D(\mathbf{x}_\mathrm{k}, \mathbf{x}_l) = \sum_{j=1}^{n} w_j\, d_j$$

$w_j \geq 1$ - težina pojedinih značajki

$\mathbf{x}_k$ *i* $\mathbf{x}_l$ opisani binarnim značajkama

$\mathbf{x}_k$

| $\mathbf{x}_l$ | | 1 | 0 |
|---|---|---|---|
| | 1 | a | b |
| | 0 | c | d |

# Klasifikacija uzoraka na temelju udaljenosti

- Russelova i Raova funkcija sličnosti

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

| | | $\mathbf{x}_k$ | |
|---|---|---|---|
| | | 1 | 0 |
| $\mathbf{x}_l$ | 1 | a | b |
| | 0 | c | d |

- Jaccardova i Needhamova funkcija sličnost

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

- Sokalova i Sneathova funkcija sličnosti

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

- Jednostavan koeficijent podudaranja

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

# Klasifikacija uzoraka na temelju udaljenosti

- Rogersova i Tanimotova funkcija sličnosti

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

| | | $\mathbf{x}_k$ | |
|---|---|---|---|
| | | 1 | 0 |
| $\mathbf{x}_l$ | 1 | a | b |
| | 0 | c | d |

- Korelacija

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

**Primjer:** Rogersova i Tanimotova funkcija sličnosti

$$x_1 = (0, 1, 1, 0, 0, 1, 1,1, 0,1)^T$$

$$x_2 = (0, 1, 1, 1, 0, 1, 0, 0, 1,1)^T$$

| | | $\mathbf{x}_1$ | |
|---|---|---|---|
| | | 1 | 0 |
| $\mathbf{x}_2$ | 1 | 4 | 2 |
| | 0 | 2 | 2 |

# Klasifikacija uzoraka na temelju udaljenosti

**Primjer:** Rogersova i Tanimotova funkcija sličnosti

$$x_1 = (0, 1, 1, 0, 0, 1, 1,1, 0,1)^T$$

$$x_2 = (0, 1, 1, 1, 0, 1, 0, 0, 1,1)^T$$

| $\mathbf{x}_2$ \ $\mathbf{x}_1$ | 1 | 0 |
|---|---|---|
| 1 | 4 | 2 |
| 0 | 2 | 2 |

| $\mathbf{x}_l$ \ $\mathbf{x}_k$ | 1 | 0 |
|---|---|---|
| 1 | a | b |
| 0 | c | d |

$$S(\mathbf{x}_k, \mathbf{x}_l) = \frac{a + d}{a + 2(b + c)} = \frac{4 + 2}{4 + 2 + 2(2 + 2)} = \frac{6}{14} = 0.429$$

# Klasifikacija uzoraka na temelju udaljenosti

- Klasifikacija uzoraka na temelju minimalne udaljenosti

**Slučaj 1: Klasifikacija na temelju jednog prototipa – karakterističnog predstavnika razreda**

- Uzorci pojedinih razreda teže grupiranju oko tipičnog (reprezentativnog) uzorka
- za M razreda – reprezentativni uzorci:

$$\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_M$$

Euklidska udaljenost nekog uzorka $\mathbf{x}$ i i-tog prototipa $\mathbf{z}_i$

$$D(\mathbf{x}, \mathbf{z}_i) = \|\mathbf{x} - \mathbf{z}_i\| = \sqrt{(\mathbf{x} - \mathbf{z}_i)^T (\mathbf{x} - \mathbf{z}_i)}$$

**Klasifikacijsko pravilo:**

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

# Klasifikacija uzoraka na temelju udaljenosti

$$D_i^2 = \|\mathbf{x} - \mathbf{z}_i\|^2 = (\mathbf{x} - \mathbf{z}_i)^T(\mathbf{x} - \mathbf{z}_i)$$

$$= \mathbf{x}^T\mathbf{x} - 2\mathbf{x}^T\mathbf{z}_i + \mathbf{z}_i^T\mathbf{z}_i$$

$$= \mathbf{x}^T\mathbf{x} - 2(\mathbf{x}^T\mathbf{z}_i - \frac{1}{2}\mathbf{z}_i^T\mathbf{z}_i)$$

Vrijedi: ako je $D^2$ minimum onda je i $D$ minimum (udaljenost je pozitivna!)

$\mathbf{x}^T\mathbf{x}$ - ne zavisi od $i$

Kada će $D^2$, odnosno $D$ biti *minimum?*

# Klasifikacija uzoraka na temelju udaljenosti

$$D_i^2 = \|\mathbf{x} - \mathbf{z}_i\|^2 = (\mathbf{x} - \mathbf{z}_i)^T(\mathbf{x} - \mathbf{z}_i)$$

$$= \mathbf{x}^T\mathbf{x} - 2\mathbf{x}^T\mathbf{z}_i + \mathbf{z}_i^T\mathbf{z}_i$$

$$= \mathbf{x}^T\mathbf{x} - \underline{2(\mathbf{x}^T\mathbf{z}_i - \frac{1}{2}\mathbf{z}_i^T\mathbf{z}_i)}$$

*D*$^2$ minimum onda je i *D* minimum, onda kada je ovo maksimum

**Decizijska funkcija:** $\mathrm{d_i}(\mathbf{x}) = \mathbf{x}^T\mathbf{z}_i - \frac{1}{2}\mathbf{z}_i^T\mathbf{z}_i \qquad i = 1, 2, \ldots, M$

**Decizijsko pravilo:**

$$\mathbf{x} \in \omega_i \text{ akko } \mathrm{d_i}(\mathbf{x}) > \mathrm{d_j}(\mathbf{x}) \text{ za sve } j \neq i$$

# Klasifikacija uzoraka na temelju udaljenosti

- Usporedimo decizijsku funkciju

$$\mathrm{d_i}(\mathbf{x}) = \mathbf{x}^T\mathbf{z}_i - \frac{1}{2}\mathbf{z}_i^T\mathbf{z}_i \qquad i = 1, 2, \dots, M$$

s (linearnom) decizijskom funkcijom (opći oblik)

$$\mathrm{d}(\mathbf{x}) = \mathbf{w}_0^{\mathrm{T}}\,\mathbf{x} + \mathrm{w}_{n+1}$$

$\mathrm{d_i}(\mathbf{x})$ je linearna decizijska funkcija!

$\mathbf{z}_i$ je $\mathbf{w}_0$ a $\mathrm{w}_{n+1} = -\frac{1}{2}\mathbf{z}_i^T\mathbf{z}_i$

- $\mathrm{w}_{ij} = \mathrm{z}_{ij}$; $j$= 1, 2, ..., $n$; $\mathrm{w}_{n+1} = -\frac{1}{2}\mathbf{z}_i^T\mathbf{z}_i$

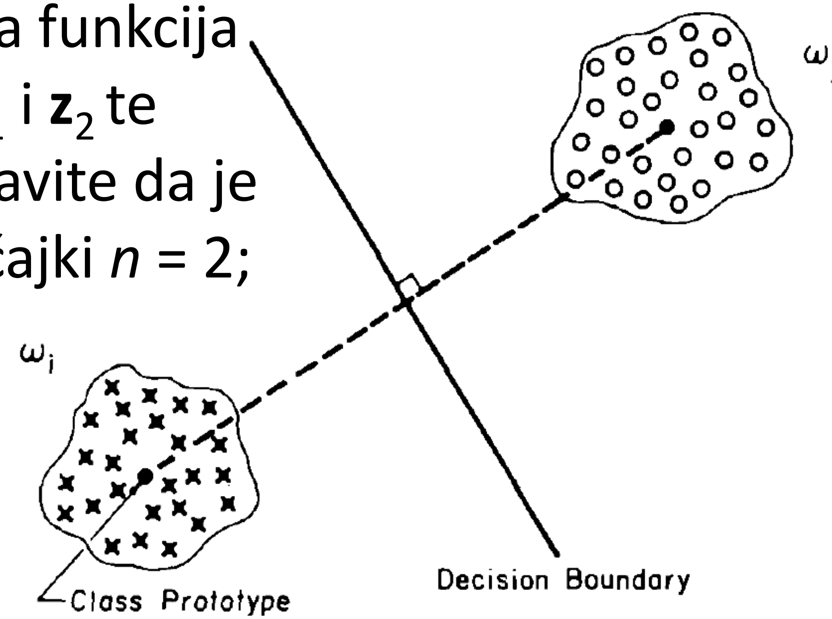
$$\mathbf{x} = \begin{bmatrix} \mathrm{x}_1 \\ \mathrm{x}_2 \\ \mathrm{x}_3 \\ \vdots \\ \mathrm{x}_n \\ 1 \end{bmatrix}$$

# Klasifikacija uzoraka na temelju udaljenosti

- Budući da klasifikator na temelju minimalne udaljenosti razvrstava uzorak na osnovi najboljeg podudaranja uzorka i odnosnog prototipa razreda – taj se pristup naziva i korelacijsko podudaranje (engl. Correlation matching)

# Klasifikacija uzoraka na temelju udaljenosti

Za vježbu pokazati da decizijska funkcija prolazi polovištem spojnice $\mathbf{z}_1$ i $\mathbf{z}_2$ te da je na nju okomita! Pretpostavite da je dimenzionalnost prostora značajki $n = 2$;




Eventualno potrebne jednadžbe:

opća jednadžba pravca: $Ax + By + C = 0$, $A^2 + B^2 \neq 0$

jednadžba pravca kroz dvije točke: $\frac{y-y_1}{y_2-y_1} = \frac{x-x_1}{x_2-x_1}$

okomitost pravaca-pravci su okomiti ako vrijedi: $A_1A_2+B_1B_2 = 0$

# Klasifikacija uzoraka na temelju udaljenosti

**Slučaj 2: Klasifikacija na temelju većeg broja prototipova za jedan razred**

- uzorci razreda $\omega_i$ grupiraju se oko $N_i$ prototipova:

$$\mathbf{z}_i^1, \mathbf{z}_i^2, \ldots, \mathbf{z}_i^{N_i}$$

$N_i$ – broj prototipova za razred $\omega_i$

Označimo udaljenost nekog uzorka **x** i razreda $\omega_i$:

$$D_i = \min_l \left\| \mathbf{x} - \mathbf{z}_i^l \right\| \quad l = 1, 2, \ldots, N_i$$

$D_i$ - je najmanja udaljenost uzorka $\mathbf{x}$ od prototipa razreda $\omega_i$

# Klasifikacija uzoraka na temelju udaljenosti

Postupak klasifikacije:

Izračunati $D_i$ za $i = 1, 2, \dots, M$ $\qquad D_i = \min_l \left\| \mathbf{x} - \mathbf{z}_i^l \right\| \quad l = 1, 2, \dots, N_i$

- nepoznati uzorak $\mathbf{x}$ klasificirati:

$$\mathbf{x} \in \omega_i \text{ akko } D_i < D_j \text{ za sve } j \neq i$$

- analogno postupku za jedan prototip dobivamo:

Decizijska funkcija:

$$d_i(\mathbf{x}) = \max_l \{ (\mathbf{x}^T \mathbf{z}_i^l) - \frac{1}{2} (\mathbf{z}_i^l)^T \mathbf{z}_i^l \} \quad l = 1, 2, \dots, N_i$$

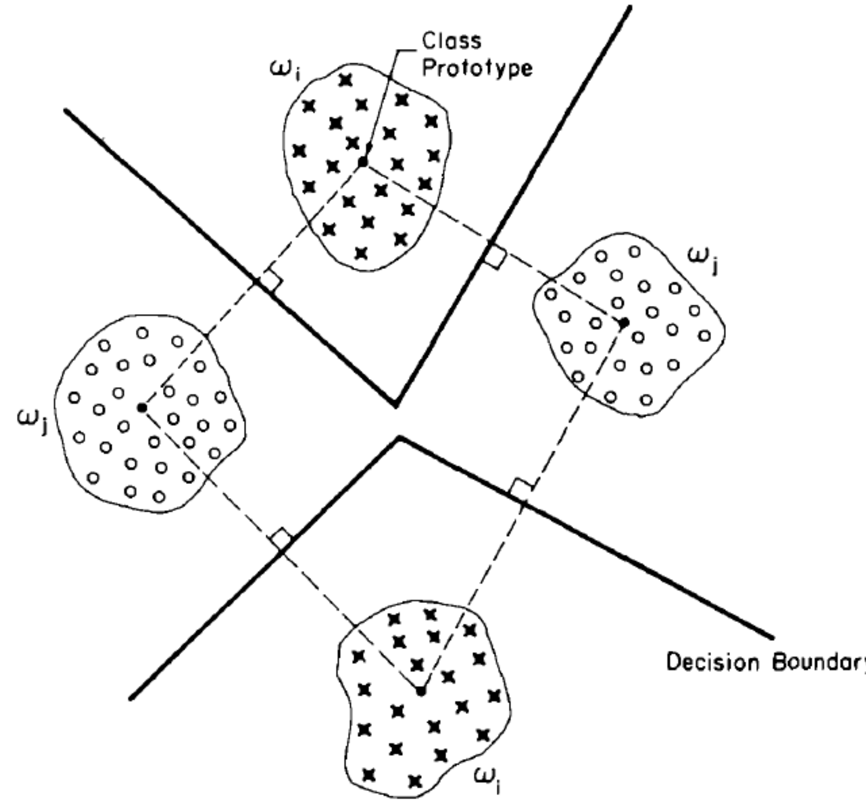
Decizijsko pravilo: $\qquad \mathbf{x} \in \omega_i$ akko je $d_i > d_j$ za sve $j \neq i$

# Klasifikacija uzoraka na temelju udaljenosti

Geometrijska interpretacija (za $n$ = 2 i dva prototipa po razredu)

/po dijelovima linearan klasifikator;

engl. Picewise-linear classifier/



# Klasifikacija uzoraka na temelju udaljenosti

**Proširenje koncepta klasifikacije na temelju udaljenosti**

- zadan je skup označenih uzoraka:

$$S_N = \{\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3, \dots, \mathbf{s}_N\}$$

- svaki od uzoraka pripada nekom od razreda:

$$\omega_1, \omega_2, \omega_3, \dots, \omega_M$$

Klasifikacijsko pravilo najbližeg (jednog) susjeda (1-Nearest Neighbour; 1-NN) /E. Fix et al. 1951/

- nepoznati uzorak $\mathbf{x}$ se razvrstava u razred $\omega_k$, tj. $\mathbf{x} \in \omega_k$ akko je njemu najbliži susjed $\mathbf{s}_i$ iz razreda $\omega_k$

$$D(\mathbf{s}_i, \mathbf{x}) = \min_{l}\{D(\mathbf{s}_l, \mathbf{x})\}, \quad l = 1,2, \dots N$$

# Klasifikacija uzoraka na temelju udaljenosti

## Primjer za 1-NN

(M. N. Murty, V. S. Devi, Pattern Recxognition)

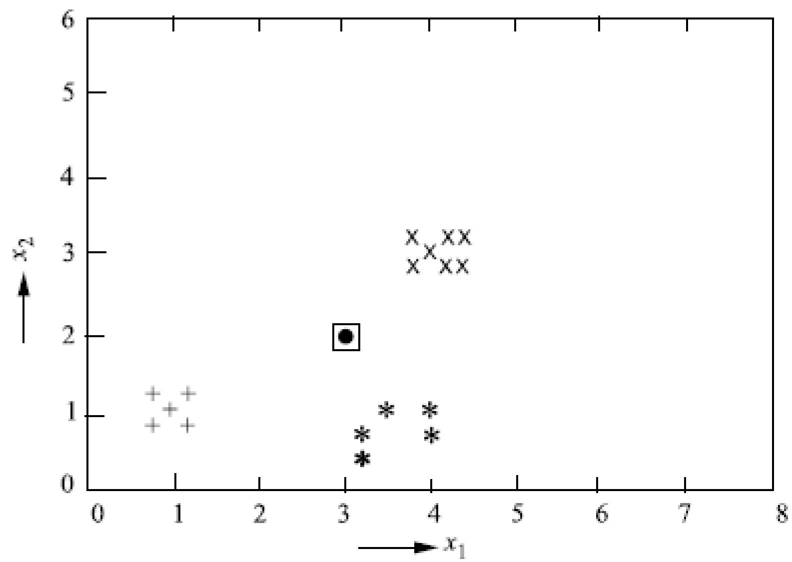
skup označenih uzoraka

Oznaka razreda    Oznaka razreda    Oznaka razreda

$X_1 = (0.8, 0.8, 1)$, $X_2 = (1.0, 1.0, 1)$, $X_3 = (1.2, 0.8, 1)$
$X_4 = (0.8, 1.2, 1)$, $X_5 = (1.2, 1.2, 1)$, $X_6 = (4.0, 3.0, 2)$
$X_7 = (3.8, 2.8, 2)$, $X_8 = (4.2, 2.8, 2)$, $X_9 = (3.8, 3.2, 2)$
$X_{10} = (4.2, 3.2, 2)$, $X_{11} = (4.4, 2.8, 2)$, $X_{12} = (4.4, 3.2, 2)$
$X_{13} = (3.2, 0.4, 3)$, $X_{14} = (3.2, 0.7, 3)$, $X_{15} = (3.8, 0.5, 3)$
$X_{16} = (3.5, 1.0, 3)$, $X_{17} = (4.0, 1.0, 3)$, $X_{18} = (4.0, 0.7, 3)$

nepoznati uzorak $\mathbf{x} = (3, 2)^T$
/označen kao ▣ /

+ – oznaka za $\omega_1$
x – oznaka za $\omega_2$
$*$ – oznaka za $\omega_3$




Kojem je uzorku iz skupa označenih uzoraka uzorak **x** najbliži?
U koji se razred razvrstava uzorak **x**?

# Klasifikacija uzoraka na temelju udaljenosti

Klasifikacijsko pravilo *k* najbližih susjeda (*k*-Nearest Neighbour)
- Umjesto nalaženja samo jednog najbližeg susjeda (1-NN) sada tražimo nepoznatom uzorku $\mathbf{x}$ *k* najbližih označenih susjeda
- nepoznati uzorak $\mathbf{x}$ razvrstava se u razred $\omega_k$ ako većina između *k* najbližih (označenih) susjeda pripada razredu $\omega_k$
- Izbor vrijednosti za *k* presudan – izborom prave vrijednosti za *k* postiže se veća točnost klasifikacije negoli za 1-NN
- za M = 2 razreda treba izabrati *k* neparan
- preporuka – neka *k ne bude* višekratnik broja razreda *M*

# Klasifikacija uzoraka na temelju udaljenosti

Primjer za 1-NN i *k*-NN:

Oznaka razreda Oznaka razreda Oznaka razreda

nepoznati uzorak

$\mathbf{x} = (4.2,\ 1.8)^T$

$X_1 = (0.8, 0.8, 1)$, $X_2 = (1.0, 1.0, 1)$, $X_3 = (1.2, 0.8, 1)$
$X_4 = (0.8, 1.2, 1)$, $X_5 = (1.2, 1.2, 1)$, $X_6 = (4.0, 3.0, 2)$
$X_7 = (3.8, 2.8, 2)$, $X_8 = (4.2, 2.8, 2)$, $X_9 = (3.8, 3.2, 2)$
$X_{10} = (4.2, 3.2, 2)$, $X_{11} = (4.4, 2.8, 2)$, $X_{12} = (4.4, 3.2, 2)$
$X_{13} = (3.2, 0.4, 3)$, $X_{14} = (3.2, 0.7, 3)$, $X_{15} = (3.8, 0.5, 3)$
$X_{16} = (3.5, 1.0, 3)$, $X_{17} = (4.0, 1.0, 3)$, $X_{18} = (4.0, 0.7, 3)$

1- NN

$\mathbf{x} \in \omega_3$

5-NN

?

$+$ – oznaka za $\omega_1$

x – oznaka za $\omega_2$

$*$ – oznaka za $\omega_3$

# Klasifikacija uzoraka na temelju udaljenosti

Modificirano pravilo *k*-najbližih susjeda (M *k*-NN)

- postupak sličan *k*-NN (uzima u obzir *k* najbližih susjeda ALI…)
- utjecaj tih *k* susjeda doprinosi odluci na temelju pridruženih težina koje su određena u skladu s njihovom udaljenosti od nepoznatog uzorka!
- algoritam se naziva i „distance-weighted *k*-NN”

# Klasifikacija uzoraka na temelju udaljenosti

Zamisao:

Svakom od susjeda nepoznatog uzorka pridružena je težina *w:*

$$w_j = \begin{cases} \dfrac{d_k - d_j}{d_k - d_1} & \text{ako je } d_k \neq d_1 \\ 1 & \text{ako je } d_k = d_1 \end{cases},$$

gdje je $j = 1, \ldots, k$ ; $d_1$ - najmanja udaljenost nepoznatog uzorka do jednog od *k susjeda;* $d_k$ - najveća udaljenost

- na temelju izračunatih težina $w_j$ M*k*-NN dodjeljuje nepoznati uzorak u razred čija je suma težina najveća
- umjesto jednostavnog pravila glasova većine (*k*-NN) ovdje se primjenjuje (utežnosno) ponderirano pravilo većine.

# Klasifikacija uzoraka na temelju udaljenosti

Primjer: M *5*-NN

$\mathbf{x} = (4.2,\ \ 1.8)^T$

Oznaka razreda Oznaka razreda Oznaka razreda

| | | |
|---|---|---|
| $X_1 = (0.8, 0.8, 1)$, | $X_2 = (1.0, 1.0, 1)$, | $X_3 = (1.2, 0.8, 1)$ |
| $X_4 = (0.8, 1.2, 1)$, | $X_5 = (1.2, 1.2, 1)$, | $X_6 = (4.0, 3.0, 2)$ |
| $X_7 = (3.8, 2.8, 2)$, | $X_8 = (4.2, 2.8, 2)$, | $X_9 = (3.8, 3.2, 2)$ |
| $X_{10} = (4.2, 3.2, 2)$, | $X_{11} = (4.4, 2.8, 2)$, | $X_{12} = (4.4, 3.2, 2)$ |
| $X_{13} = (3.2, 0.4, 3)$, | $X_{14} = (3.2, 0.7, 3)$, | $X_{15} = (3.8, 0.5, 3)$ |
| $X_{16} = (3.5, 1.0, 3)$, | $X_{17} = (4.0, 1.0, 3)$, | $X_{18} = (4.0, 0.7, 3)$ |

5 najbližih susjeda uzorku **x** su uzorci: $X_{17} \in \omega_3;\ X_8 \in \omega_2;\ X_{11} \in \omega_2;\ X_{16} \in \omega_3;\ X_7 \in \omega_2$

„obično" *k*-NN pravilo: $\mathbf{x} \in \omega_2$

# Klasifikacija uzoraka na temelju udaljenosti

Primjer (nastavak):

M *k*-NN pravilo; *k* = 5

$$X_{17} \in \omega_3;\ X_8 \in \omega_2;\ X_{11} \in \omega_2;\ X_{16} \in \omega_3;\ X_7 \in \omega_2$$

- Izračunajmo udaljenosti između nepoznatog uzorka i *k* najbližih susjeda:

$$d(\mathbf{x}, X_{17}) = 0.83;\ d(\mathbf{x}, X_8) = 1.0;\ d(\mathbf{x}, X_{11}) = 1.02;\ d(\mathbf{x}, X_{16}) = 1.06;\ d(\mathbf{x}, X_7) = 1.08$$

$d_1$ = 0.83 $d_k$ = 1.08

Za uzorak $X_{17}$ :

$w_1 = 1$ jer je $d(\mathbf{x}, X_{17}) = d_1$

Za uzorak $X_8$:

$$w_2 = \frac{d_k - d_j}{d_k - d_1} = \frac{1.08 - 1.0}{1.08 - 0.83} = 0.32$$

Za uzorak $X_{11}$:

$$w_3 = \frac{d_k - d_j}{d_k - d_1} = \frac{1.08 - 1.02}{1.08 - 0.83} = 0.24$$

Za uzorak $X_{16}$:

$$w_4 = \frac{d_k - d_j}{d_k - d_1} = \frac{1.08 - 1.06}{1.08 - 0.83} = 0.08$$

Za uzorak $X_7$:

$w_5 = 0$

# Klasifikacija uzoraka na temelju udaljenosti

Primjer (nastavak):

M *k*-NN pravilo; *k* = 5

Za uzorak $X_{17}$ :

$w_1 = 1$ jer je $d(\mathbf{x}, X_{17}) = d_1$

Za uzorak $X_8$:

$$w_2 = \frac{d_k - d_j}{d_k - d_1} = \frac{1.08 - 1.0}{1.08 - 0.83} = 0.32$$

Za uzorak $X_{11}$:

$$w_3 = \frac{d_k - d_j}{d_k - d_1} = \frac{1.08 - 1.02}{1.08 - 0.83} = 0.24$$

Za uzorak $X_{16}$:

$$w_4 = \frac{d_k - d_j}{d_k - d_1} = \frac{1.08 - 1.06}{1.08 - 0.83} = 0.08$$

Za uzorak $X_7$:

$w_5 = 0$

# Klasifikacija uzoraka na temelju udaljenosti

Primjer (nastavak):
Zbrojimo težinske vrijednosti za razrede:

Za $\omega_3$imamo: 1 ($X_{17}$ pripada razredu $\omega_3$) $+ 0.08(X_{16}$ pripada razredu $\omega_3) = 1.08$

Za $\omega_2$imamo: 0.32($X_8$ pripada razredu $\omega_2$) $+ 0.24$ ($X_{11}$ pripada razredu $\omega_2$) $+ 0(X_7$ pripada razredu $\omega_2) = 0.56$

Odluka: $x \in \omega_3$

# Klasifikacija uzoraka na temelju udaljenosti

Primjer (nastavak):
Zbrojimo težinske vrijednosti za razrede:

Za $\omega_3$imamo: 1 ($X_{17}$ pripada razredu $\omega_3) + 0.08(X_{16}$ pripada razredu $\omega_3) = 1.08$
Za $\omega_2$imamo: 0.32($X_8$ pripada razredu $\omega_2) + 0.24\ (X_{11}$ pripada razredu $\omega_2) + 0(X_7$ pripada razredu $\omega_2) = 0.56$

Odluka: $x \in \omega_3$

# Klasifikacija uzoraka na temelju udaljenosti

Klasifikacijsko pravilo $(k, l)$ -NN

Pravilo klasifikacije pretpostavlja da se u skupu od *k* najbližih susjeda treba pojaviti barem $l$ uzoraka koji pripadaju nekom razredu $\omega_i$ da bi se nepoznati uzorak klasificirao u taj razred *(npr.* $l$ *= 2/3 k).*

- ako to nije zadovoljeno uzorak ostavimo neklasificiranog ili ga pokušamo klasificirati nekim drugim klasifikacijskim pravilom;
- podsjetimo se – sustav za raspoznavanje osoba na temelju karakteristika dlana (uvodno predavanje) koristio je klasifikacijsko pravilo (3, 3)-NN

# Klasifikacija uzoraka na temelju udaljenosti

Inačica $(k, l)$ -NN pravila: $(k, l_i) - NN$ ( M. Hellerman, 1977)

- postupak kojim dopuštamo klasifikaciju nepoznatog uzorka u razred $\omega_i$ ako je potrebna većina susjeda za klasifikaciju u taj razred $l_i$
- npr. $l_1 = 2/3\, k; \ldots, l_M = 3/4k$

- Inačicu možemo upotrijebiti kadu su poznate apriorne vjerojatnosti pojavljivanja razreda različite!

# Klasifikacija uzoraka na temelju udaljenosti

Općenito, mogu se upotrijebiti različite mjere udaljenosti, npr. Euklidska, Čebiševljeva, “block-city”, Mahalanobisova, itd.

Mahalanobisova:

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

$\mathbf{C}$ - kovarijacijska matrica uzoraka iz skupa za učenje:

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

$$\mathbf{m}_x = \frac{1}{N} \sum_{i}^{N} \mathbf{x}_i$$

# Klasifikacija uzoraka na temelju udaljenosti

**Ocjena pogreške 1-*NN* klasifikatora**

- Pravila klasifikacije 1-NN, *k*-NN i različite inačice NN su vrlo djelotvorna u smislu točnosti klasifikacije ONDA kada je skup označenih uzoraka vrlo velik!

$N \to \infty$

- vjerojatnost pogreške

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

$P_{NN}$ - pogreška NN klasifikatora

$P_B$ - optimalna pogreška (Bayesov klasifikator)

$M$ - broj razreda

# Klasifikacija uzoraka na temelju udaljenosti

Problem: složenost računanja udaljenosti i traženja *k* (ili 1) najbližih susjeda (vremenski zahtjevna metoda!)

- ocjena složenosti traženja najbližih susjeda O($kN$)
- problem je još naglašeniji kada je dimenzionalnost prostora značajki $n >> 1$

Moguća rješenja:

- Različiti algoritmi pretprocesiranja podataka

(preliminarno uređivanje označenih uzoraka na temelju intra-set udaljenosti; vektorska kvantizacija prostora značajki; Voronoiev mozaik, uporaba prototipova za razrede)

# Klasifikacija uzoraka na temelju udaljenosti

Još jedan važan detalj: **Normalizacija podataka**

- U mnogim stvarnim situacijama u oblikovanju sustava za raspoznavanje susrećemo se sa značajkama čije su vrijednosti unutar različitih dinamičkih opsega!

Značajke s velikim vrijednostima mogu imati puno veći utjecaj na ishod klasifikacije negoli one izražene malim vrijednostima (iako nije nužno da su značajke izražene velikim vrijednostima važnije i značajnije za oblikovanje klasifikatora!)

*npr.* Klasifikacija (silicijskih) pločica na temelju površine i debljine (odnosi značajki mogu biti nekoliko razreda veličine!)

# Klasifikacija uzoraka na temelju udaljenosti

Problem se može riješiti normalizacijom značajki tako da njihove vrijednosti leže u sličnim granicama

- za *N* raspoloživih (označenih) vektora za neku k-tu značajku izračunamo

$$\bar{x}_k = \frac{1}{N}\sum_{i=1}^{N} x_{ik} \quad ; \quad k = 1, 2, \dots n$$

$$\sigma_k^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_{ik} - \bar{x}_k)^2$$

$$\tilde{x}_{ik} = \frac{x_{ik} - \bar{x}_k}{\sigma_k}$$

- sve normalizirane značajke sad imaju srednju vrijednost 0 i jediničnu varijancu!

# Klasifikacija uzoraka na temelju udaljenosti

Primjer:

$N = 3$

$\mathbf{x}_1 = (1200; 0.2)^T$

$\mathbf{x}_2 = (1000; 0.1)^T$

$\mathbf{x}_3 = (2000; 0.4)^T$

$$\bar{x}_1 = \frac{1}{N}\sum_{i=1}^{N} x_{i1} = \frac{1}{3}\,(1200 + 1000 + 2000) = 1400$$

$$\bar{x}_2 = \frac{1}{N}\sum_{i=1}^{N} x_{i2} = \frac{1}{3}\,(0.2 + 0.1 + 0.4) = 0.233$$

# Klasifikacija uzoraka na temelju udaljenosti

$\sigma_1^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_{i1} - \bar{x}_1)^2 = \frac{1}{2}[(1200-1400)^2 + (1000-1400)^2 + (2000-1400)^2]$

$\sigma_1^2 = 280000$

$\sigma_1 = 529.15$

$\sigma_2^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_{i1} - \bar{x}_1)^2 = \frac{1}{2}[(0.2-0.233)^2 + (0.1-0.233)^2 + (0.4-0.233)^2]=$
$\frac{1}{2}[(0.033)^2 + (0.133)^2 + (0.167)^2]= 0.0225$

$\sigma_2 = 0.15$

$$\tilde{x}_{11} = \frac{x_{11} - \bar{x}_1}{\sigma_1} = \frac{1200 - 1400}{529.15} = \frac{-200}{529.15} = -0.378$$

$$\tilde{x}_{21} = \frac{x_{21} - \bar{x}_1}{\sigma_1} = \frac{1000 - 1400}{529.15} = \frac{-400}{529.15} = -0.756$$

$$\tilde{x}_{31} = \frac{x_{31} - \bar{x}_1}{\sigma_{k1}} = \frac{2000 - 1400}{529.15} = \frac{600}{529.15} = 1.133$$

# Klasifikacija uzoraka na temelju udaljenosti

$$\tilde{x}_{12} = \frac{0.2-0.233}{0.15} = \frac{-0.033}{0.15} = -0.22$$

$$\tilde{x}_{22} = \frac{0.1-0.233}{0.15} = \frac{-0.133}{0.15} = -0.886$$

$$\tilde{x}_{32} = \frac{0.4-0.233}{0.15} = \frac{0.167}{0.15} = 1.11$$

Normalizirani vektori:

$\mathbf{x}_1 = (-0.378; -0.22)^T$

$\mathbf{x}_2 = (-0.756; -0.886)^T$

$\mathbf{x}_3 = (1.13; 1.11)^T$

Provjeriti da li je standardna devijacija 1 i da li je srednja vrijednost značajki 0!

# Klasifikacija uzoraka na temelju udaljenosti

**Taksonomija klasifikacijskih metoda na temelju udaljenosti**

- **obzirom na način zapisa uzoraka iz skupa za učenje:**

  a) u memoriji klasifikatora pohranjeni su svi uzorci iz skupa za učenje;

  b) u memoriji su pohranjeni samo karakteristični predstavnici razreda;

- **obzirom na broj (najbližih) susjeda:**

  a) jedan najbliži susjed (1-NN)

  b) k najbližih susjeda (k-NN)

  c) $(k,l)$ -NN i $(k,l_i) - NN$

- **obzirom na način modifikacije:**

  a) M*k*-NN

  b) neizraziti k-NN (Fuzzy k-NN)