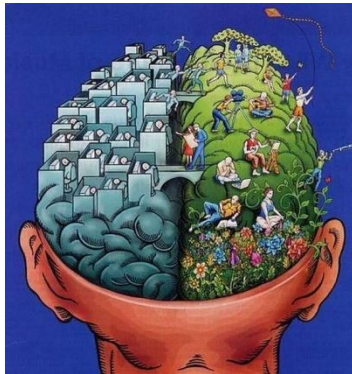




Budapesti Műszaki és Gazdaságtudományi Egyetem

Méréstechnika és Információs rendszerek Tanszék



Nem felügyelt tanulás: klaszterezés

Előadó:

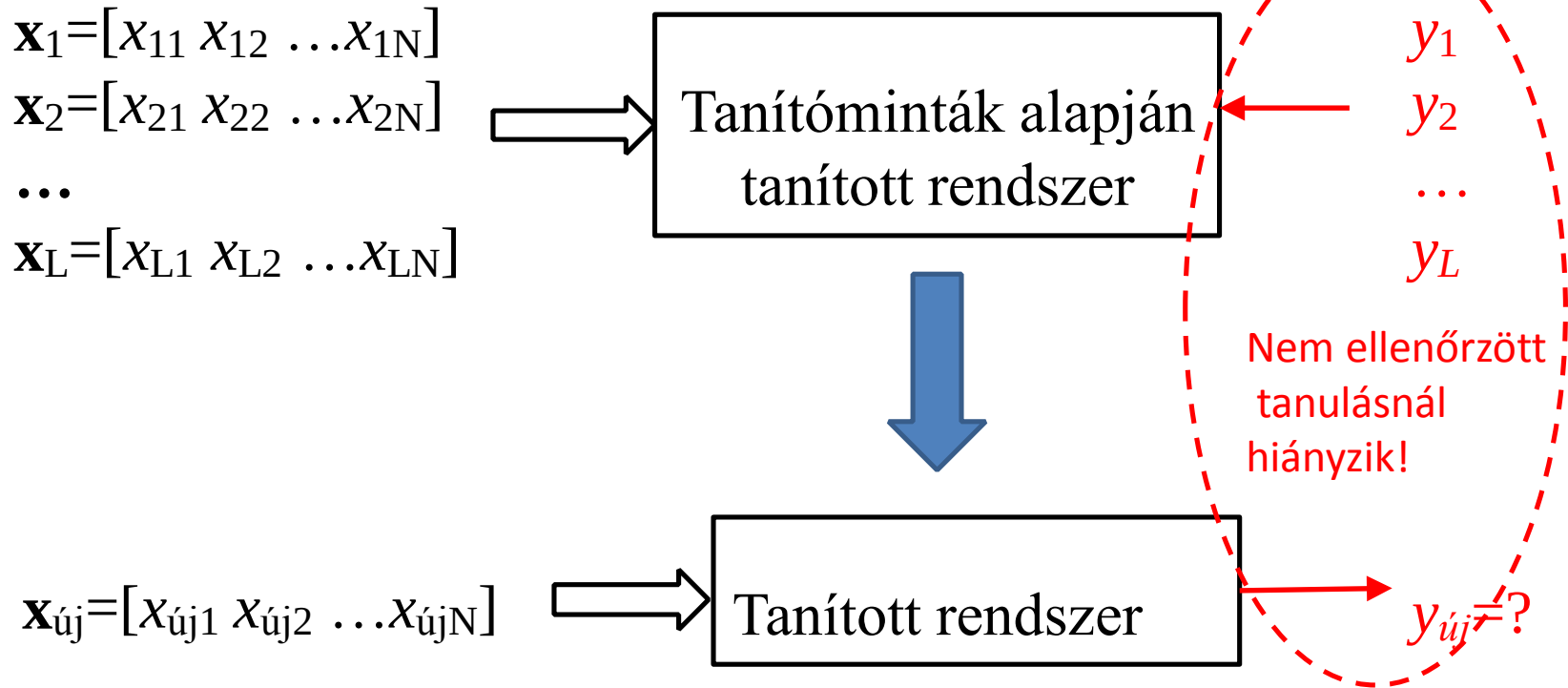
Hullám Gábor

Előadás anyaga:

Pataki Béla

Hullám Gábor

Felügyelt vs nem felügyelt tanulás



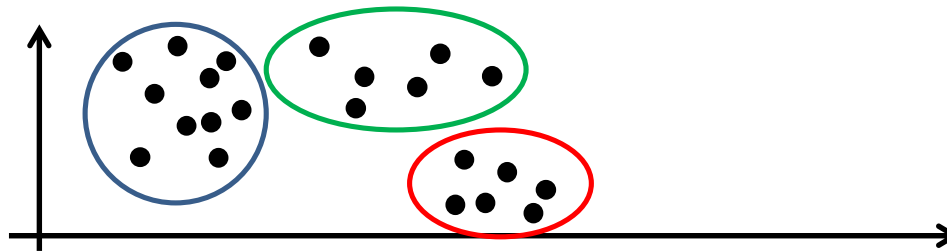
- ▶ Felügyelt tanulásnál a cél: újonnan érkező $\mathbf{x}_{új}$ mintához milyen valószínűséggel tartozik egy y címke
- ▶ $P(y|\mathbf{x})$ eloszlást kéne ismernünk

Felügyelt vs nem felügyelt tanulás

*Felügyelt tanulás*nál a tanulóminta-halmazból becsüljük $P(\mathbf{x}|\mathbf{y})$ -t és $P(\mathbf{x})$ -t, innen becsüljük $P(\mathbf{y}|\mathbf{x})$ -t.

*Nem felügyelt tanulás*nál nem ismerjük az $\mathbf{y}$ címkéket: csak $P(\mathbf{x})$ -t tudjuk becsülni.

Klaszterezés: a $P(\mathbf{x})$ olyan becslése, melyben a mintatérben elkülönülő csoportokra bontjuk a mintákat



A klaszterezés haszna

Miért lehet jó klaszterezni a mintákat?

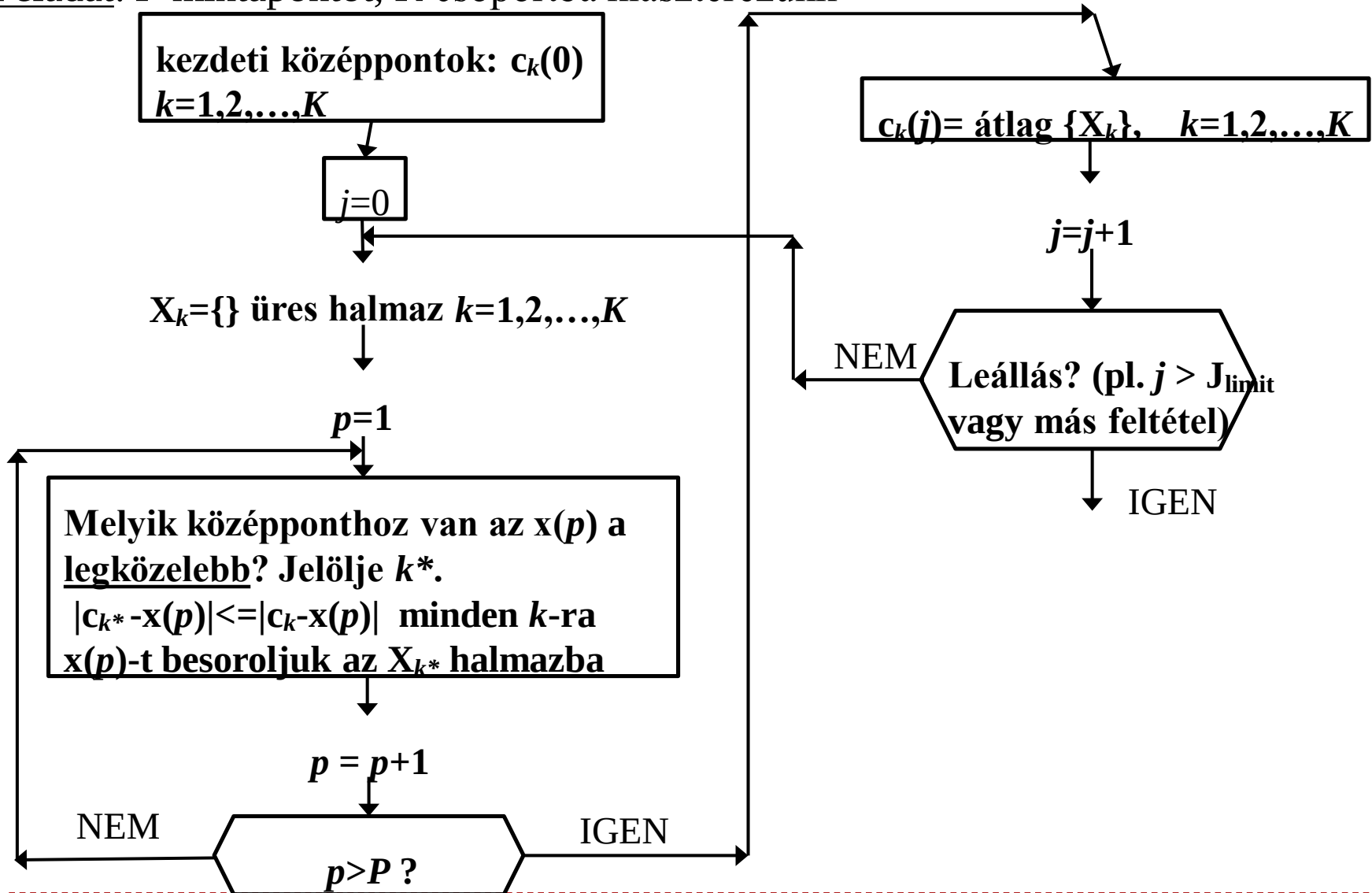
- ▶ **Segíthet a kilógó (outlier) adatok felderítésében**
 - ▶ ha mindegyik kialakuló klasztertől távoli adat érkezik, valószínűleg hibás (kilógó, outlier)
- ▶ Megmutatja, hogy a jelenségnek, folyamatnak vannak-e **tipikus állapotai**
 - ▶ pl. egy gyártási folyamatban, egy elektromos vagy egyéb fogyasztási viselkedésben lehetnek tipikus helyzetek
- ▶ Felhasználható **hiányzó paraméterek pótlására**
 - ▶ ha el tudjuk dönteni, vagy valószínűsíteni tudjuk, hogy melyik klaszterbe tartozik a néhány paraméterében hiányos adat, akkor a klaszter tipikus paraméterei jobb javítást, pótlást tesznek lehetővé

Klaszterezés típusai

- ▶ Diszkriminatív: az egyes esetek (minták) közti távolságot használjuk fel: a közeli minták tartozzanak egy csoportba, a távoliak külön csoportba.
- ▶ Generatív: Egy, a mintahalmaz létrehozását (generálását) magyarázó modellt használunk. A modell magyarázza az egyes csoportok létrejöttét – a modellparamétereket tanuljuk.

K-közép klaszterezés

Feladat: P mintapontot, K csoportba klaszterezünk



Lényeges jellemzők

- ▶ Kritikus: a jó távolságmérték és jó klaszterszám (K) meghatározása
- ▶ A távolságmérték megváltoztathatja, hogy melyik pont melyik centrumhoz van a legközelebb
- ▶ Szerencsétlen esetben egy rosszul inicializált középpont soha nem jut érvényre (senki se lesz hozzá a legközelebb)

Klaszterezés alapkérdései – I.

2 fő célkitűzés:

- ▶ Az egyes **klasztereken belül** legyen a lehető legnagyobb a homogenitás
 - ▶ = nagy 'hasonlóság' az elemek között
 - ▶ = kis 'távolság' az elemek között
- ▶ Az különböző **klaszterek között** legyen nagy a 'távolság'
 - = kis 'hasonlóság' a különböző klaszterekbe tartozó elemek között
- ▶ A két kritérium között kell egyensúlyt találni



Klaszterezés alapkérdései – II.

- ▶ Mi legyen a **hasonlósági/távolsági mérce** (similarity/distance metric)?
 - ▶ Folytonos és diszkrét numerikus változóknál egyszerűen értelmezhető
 - ▶ Kategorikus változóknál külön megfontolást igényel
- Mekkora legyen a **hasonlósági küszöb**?
 - Túl alacsony: jelentősen különböző elemek kerülnek egy klaszterbe
 - Túl magas: elég hasonló elemeket sem enged egy klaszterbe kerülni



Klaszterezés alapkérdései – III.

▶ Mekkora az **optimális klaszterszám**?

▶ Extrém esetek: → használhatatlan

▶ Annyi klaszter ahány minta

▶ Egy óriási klaszter (benne minden minta)

▶ Túl sok klaszter → nehezen értelmezhető

▶ Túl kevés klaszter → nem különülnek el a csoportok

▶ Nincs általános megoldás

▶ Adott problémára megadható egy ideális klaszterszám



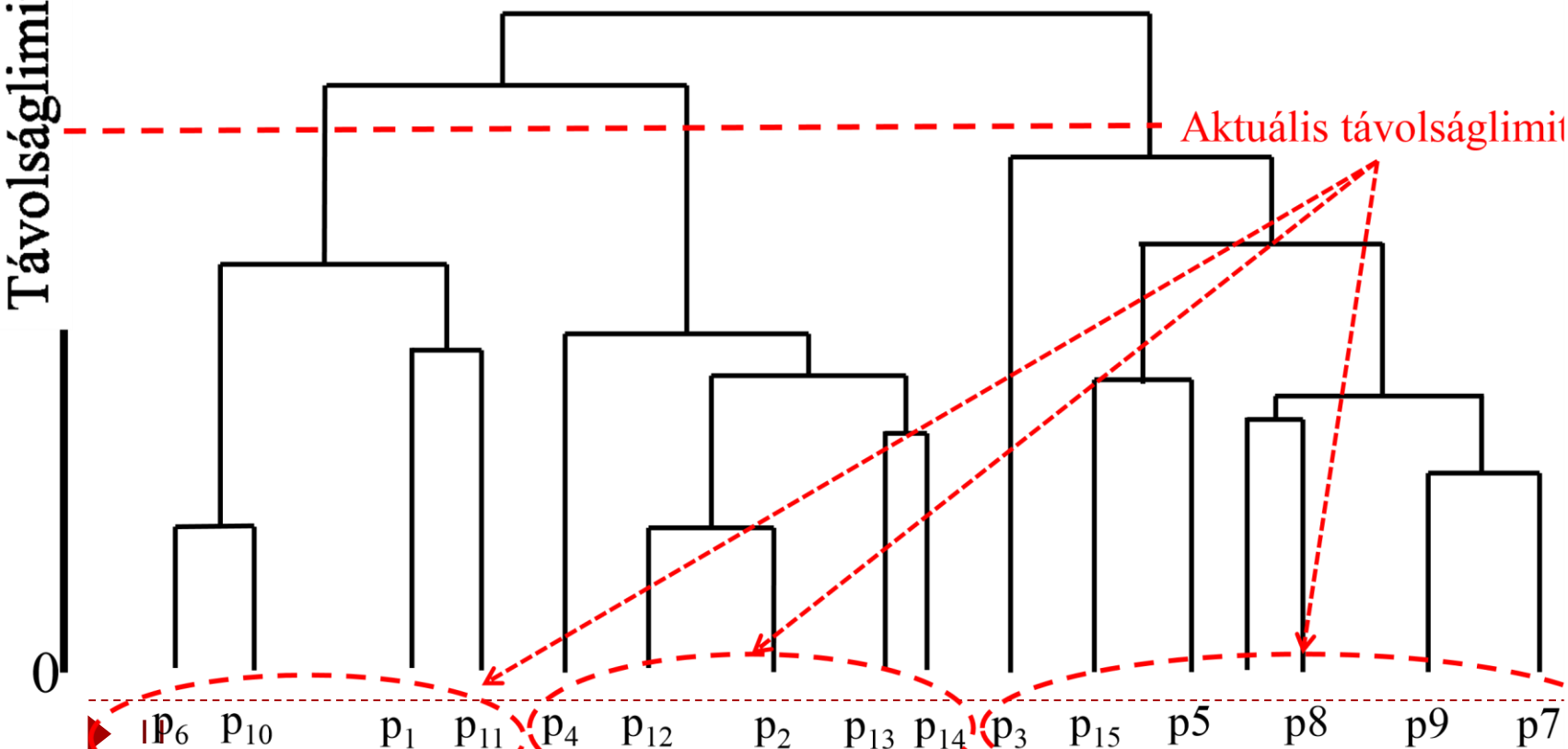
Diszkriminatív klaszterezés – példa #2

Hierarchikus klaszterezés

∞

Távolságlimit

Kritikus: a jó távolságmérték és távolságlimit meghatározása



Top-down tanulás: kezdetben egyetlen klaszter, mindig kettévágjuk

Down-Up: kezdetben minden pont külön klaszter, egyesítjük a legközelebbieket

Távolságmértékek

Euklideszi távolság:

$$\|p_1, p_2\|_2 = \sqrt{\sum_{k=1}^N (p_{1k} - p_{2k})^2}$$

Manhattan távolság:

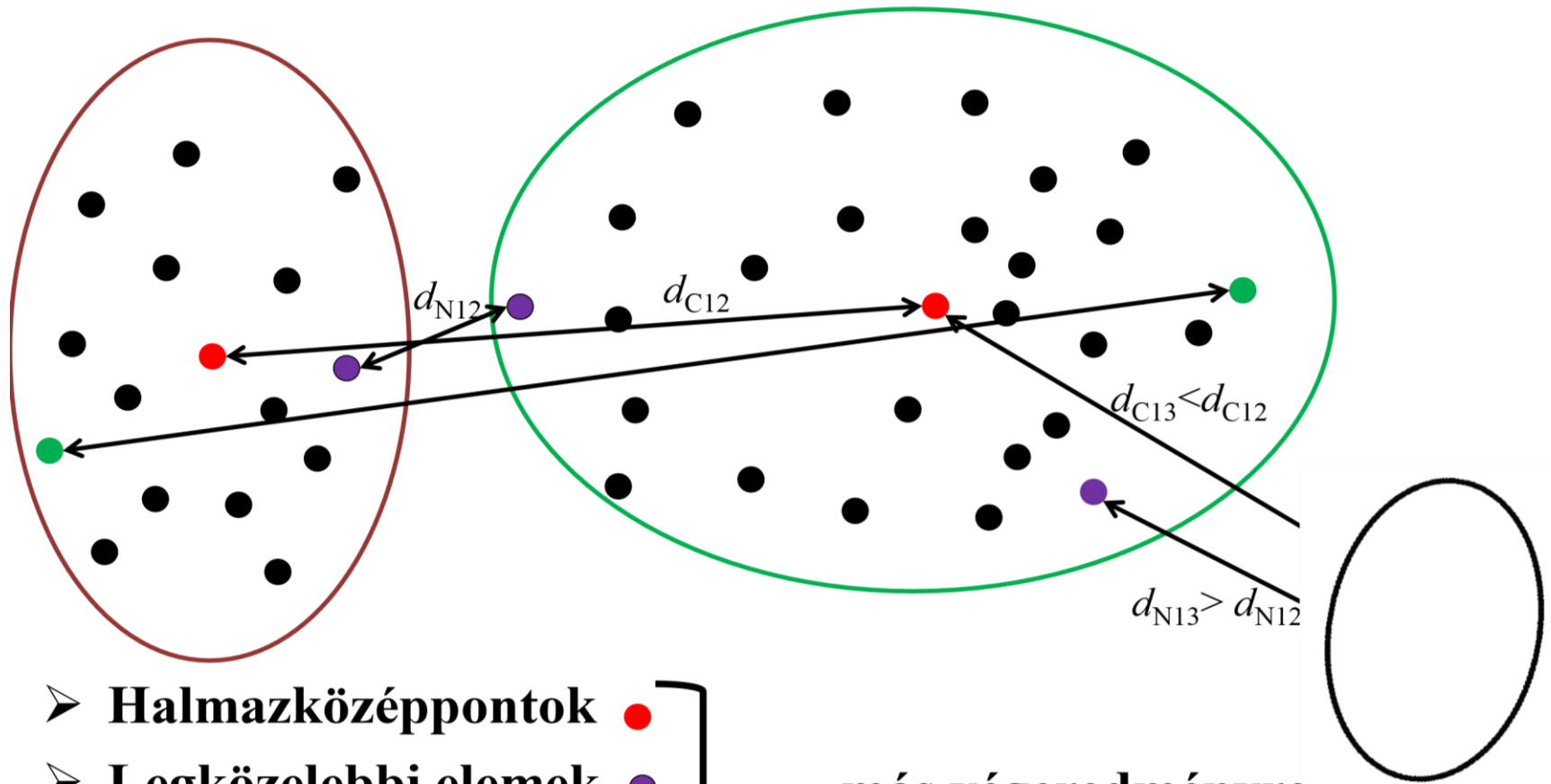
$$\|p_1, p_2\|_1 = \sum_{k=1}^N |p_{1k} - p_{2k}|$$

Mahalanobis távolság:

$$\|p_1, p_2\|_{2*} = \sqrt{\sum_{k=1}^N (p_{1k} - p_{2k})^T \cdot C^{-1} \cdot (p_{1k} - p_{2k})}$$

► Ahol C^{-1} a kovariancia mátrix inverze

Távolság halmazok között



- Halmazközéppontok ●
 - Legközelebbi elemek ●
 - Legtávolabbi elemek ●
 - stb.
- } $\Rightarrow$ más végeredményre vezethetnek

Condorcet kritérium alapú klaszterezés

- ▶ **n** jelölt : $x^{(i)}$
- ▶ **m** darab szavazó : v (voter)
- ▶ Mindegyik szavazó v_k egy rangsort állít fel π_k (permutáció)
 - ▶ Pl: $x^{(\pi_k(i))} < x^{(\pi_k(j))}$ az i . helyen állót favorizálja j . helyen lévőhöz képest
- ▶ Az a jelölt győz aki a szavazók rangsorát páronként összemérve többször favorizált (paired majority rule)
- ▶ Ez alapján egy globális rangsort hoz létre λ :
 - ▶ $x^{(\lambda(1))} < x^{(\lambda(2))} < \dots$

$$\sum_{1 \leq i < j \leq n} \#\{k | \pi_k^{-1}(\lambda(i)) < \pi_k^{-1}(\lambda(j))\}$$

A szavazás

I. változó	„Ellenfél” klaszter				
„Jelölt” klaszter		1	2	3	4
	1	—	0	0	1
	2	1	—	1	1
	3	1	0	—	1
	4	0	0	0	—

- ▶ Preferencia sorrend (2,3,1,4)
- ▶ 0 jelentése: a jelölt rosszabb, mint az ellenfél
- ▶ 1 jelentése: a jelölt jobb, mint az ellenfél

A szavazás

I I. változó	„Ellenfél” klaszter					(4,1,3,2)
„Jelölt” klaszter		<i>I</i>	<i>2</i>	<i>3</i>	<i>4</i>	
	<i>I</i>	—			0	
	<i>2</i>	0	—	0	0	
	<i>3</i>	0		—	0	
	<i>4</i>				—	

III. változó	„Ellenfél” klaszter					(1,3,2,4)
„Jelölt” klaszter		<i>I</i>	<i>2</i>	<i>3</i>	<i>4</i>	
	<i>I</i>	—				
	<i>2</i>	0	—	0		
	<i>3</i>	0		—		
	<i>4</i>	0	0	0	—	

Végeredmény

Σ	„Ellenfél” klaszter				
„Jelölt” klaszter		1	2	3	4
	1	—	2	2	2
	2	1	—	1	2
	3	1	2	—	2
	4	1	1	1	—

Condorcet kritérium alkalmazása klaszter jelöltek minősítésére

- ▶ **n** jelölt $\rightarrow$ n darab klaszter
- ▶ **m** szavazó $\rightarrow$ m darab változó

Jelölés:

x, y : két összehasonlítandó objektum

v : egy adott változó, $v(x)$: x objektum v változó szerinti értéke

$\mathcal{V}$: az összes változó halmaza

C : egy adott klaszter

$\mathcal{C}$: teljes klaszterezés

$$c(\mathcal{C}) := \sum_{C \in \mathcal{C}} \sum_{\substack{x, y \in C \\ x \neq y}} \sum_{v \in \mathcal{V}} [v(x) \sim v(y)] + \sum_{C \in \mathcal{C}} \sum_{\substack{x \in C \\ y \in \mathcal{O} \setminus C}} \sum_{v \in \mathcal{V}} [v(x) \not\sim v(y)]$$