

U tekstu vjerojatno ima pogrešaka. Ako ih uočite, molim da mi to javite

Grupiranje podataka u skupine¹

Rudolf Scitovski, Odjela za matematiku, Sveučilište u Osijeku²

1 Formulacija problema

Neka je $S = \{x_i = (x_1^{(i)}, \dots, x_n^{(i)})^T \in \mathbb{R}^n : i = 1, \dots, m\}$ skup vektora iz vektorskog prostora $\mathbb{R}^n$. *Partitivni skup* skupa S je familija svih njegovih podskupova. Uobičajena oznaka je $\mathcal{P}(S)$ (Papić, 2000).

Rastav skupa S na disjunktne podskupove $\pi_1, \dots, \pi_k$, $1 \leq k \leq m$, takve da vrijedi

$$\bigcup_{i=1}^k \pi_i = S, \quad (1)$$

$$\pi_i \cap \pi_j = \emptyset, \quad i \neq j, \quad (2)$$

$$m_j := |\pi_j| \geq 1, \quad j = 1, \dots, k \quad (\text{broj elemenata podskupa } \pi_j), \quad (3)$$

označit ćemo s $\Pi(S) = \{\pi_1, \dots, \pi_k\}$ i zvati *particija skupa* S . Skupove $\pi_1, \dots, \pi_k$ zvat ćemo *klasteri*. Skup svih particija skupa S sastavljenih od k klastera koje zadovoljavaju (1-3) označit ćemo s $\Psi(S, k)$.

Zadatak 1. *Općenito, broj elemenata partitivnog skupa, skupa S od m elemenata jednak je 2^m . Pokažite da je broj svih particija skupa S , koji zadovoljavaju (1-3) jednak*

$$|\Psi(S, k)| = \frac{1}{k!} \sum_{j=1}^k (-1)^{k-j} \binom{k}{j} j^m. \quad (4)$$

Izračunajte broj svih particija skupa S zadanih s (1-3) specijalno za $m = 10, 50, 100, 1000$ i $k = 2, 3, 5, 8, 10$.

$ \Psi(S, k) $	$k = 2$	$k = 3$	$k = 5$	$k = 8$	$k = 10$
$m = 10$	511	9330	42525	750	1
$m = 50$	10^{15}	10^{23}	10^{33}	10^{40}	10^{43}
$m = 10^3$	10^{300}	10^{476}	10^{697}	10^{898}	10^{993}
$m = 10^6$	$10^{301\,029}$	$10^{477\,120}$	$10^{698\,968}$	$10^{903\,085}$	10^{10^6}

The applications are important and very attractive (Biology, Medicine, Robotic, Marketing, Education, ...)

Approaches: Algebraic and Combinatorial Optimization, Statistics, Graph Theory, Fuzzy Clustering

¹Tekst je pisan uglavnom prema (Späth, 1983; Kogan, 2007)

²Sve ilustracije i *Mathematica*-programe koji prate algoritme izradio je Ivan Vazler, asistent Odjela za matematiku, Sveučilišta u Osijeku

Zadatak 2. Ako je $|\pi_j| = m_j$, pokažite da je $c_j = \frac{1}{m_j} \sum_{x_i \in \pi_j} x_i$ **centroid** klastera π_j , tj.

$$\min_{x \in \mathbb{R}^n} \sum_{x_i \in \pi_j} \|x_i - x\|^2 = \sum_{x_i \in \pi_j} \|x_i - c_j\|^2.$$

Least Squares – optimal partition:

$$f(\pi_j) = \sum_{x_i \in \pi_j} \|x_i - c_j\|_2^2,$$

$$c_j = \operatorname{argmin}_{a \in \mathbb{R}^n} \sum_{x_i \in \pi_j} \|x_i - a\|_2^2 = \frac{1}{|\pi_j|} \sum_{x_i \in \pi_j} x_i \in \mathbb{R}^n \quad (\text{centroid of the cluster } \pi_j)$$

$$\begin{aligned} \min_{a \in \mathbb{R}^n} \sum_{x_i \in \pi_j} \|x_i - a\|_2^2 &= \min_{a \in \mathbb{R}^n} \sum_{x_i \in \pi_j} \left(\sum_{s=1}^n (x_s^{(i)} - a_s)^2 \right) \\ &= \min_{a \in \mathbb{R}^n} \sum_{s=1}^n \left(\sum_{x_i \in \pi_j} (x_s^{(i)} - a_s)^2 \right) = \sum_{s=1}^n \min_{a_s \in \mathbb{R}} \sum_{x_i \in \pi_j} (x_s^{(i)} - a_s)^2 \\ &= \sum_{s=1}^n \sum_{x_i \in \pi_j} \left(x_s^{(i)} - A(x_s^{(i)}) \right)^2 = \sum_{x_i \in \pi_j} \sum_{s=1}^n \left(x_s^{(i)} - A(x_s^{(i)}) \right)^2 = \sum_{x_i \in \pi_j} \|x_i - a^*\|_2^2, \\ a^* &= \left(\frac{1}{|\pi_j|} \sum_{x_i \in \pi_j} x_1^{(i)}, \dots, \frac{1}{|\pi_j|} \sum_{x_i \in \pi_j} x_n^{(i)} \right)^T = \frac{1}{|\pi_j|} \sum_{x_i \in \pi_j} x_i \end{aligned}$$

Dakle, suma kvadrata udaljenosti svih elemenata klastera π_j do neke čvrste točke u prostoru $\mathbb{R}^n$ je najmanja onda i samo onda ako je ta čvrsta točka baš centroid c_j klastera π_j .

Zadatak 3. Ako je $\pi_j \in \Pi$ klaster, što je

$$(a) \quad \min_{x \in \mathbb{R}^n} \sum_{x_i \in \pi_j} \|x_i - x\|_1? \quad (b) \quad \min_{x \in \mathbb{R}^n} \sum_{x_i \in \pi_j} \|x_i - x\|_\infty?$$

Least Absolute Deviations – optimal partition>

$$f(\pi_j) = \sum_{x_i \in \pi_j} \|x_i - c_j\|_1,$$

$$c_j = \operatorname{argmin}_{a \in \mathbb{R}^n} \sum_{x_i \in \pi_j} \|x_i - a\|_1 = \left(\operatorname{med}_i x_1^{(i)}, \dots, \operatorname{med}_i x_n^{(i)} \right)^T \in \mathbb{R}^n,$$

$$\begin{aligned} \min_{a \in \mathbb{R}^n} \sum_{x_i \in \pi_j} \|x_i - a\|_1 &= \min_{a \in \mathbb{R}^n} \sum_{x_i \in \pi_j} \left(\sum_{s=1}^n |x_s^{(i)} - a_s| \right) \\ &= \min_{a \in \mathbb{R}^n} \sum_{s=1}^n \left(\sum_{x_i \in \pi_j} |x_s^{(i)} - a_s| \right) = \sum_{s=1}^n \min_{a_s \in \mathbb{R}} \sum_{x_i \in \pi_j} |x_s^{(i)} - a_s| \\ &= \sum_{s=1}^n \sum_{x_i \in \pi_j} |x_s^{(i)} - \operatorname{med}_{i=1, \dots, m} x_s^{(i)}| = \sum_{x_i \in \pi_j} \sum_{s=1}^n |x_s^{(i)} - \operatorname{med}_{i=1, \dots, m} x_s^{(i)}| \\ &= \sum_{x_i \in \pi_j} \|x_i - a^*\|_1, \quad a^* = \left(\operatorname{med}_i x_1^{(i)}, \dots, \operatorname{med}_i x_n^{(i)} \right)^T \in \mathbb{R}^n \end{aligned}$$

Na skupu svih particija $\Psi(S, k)$ skupa S sastavljenih od k klastera definiramo kriterijsku funkciju cilja

$$f : \mathcal{P}(S) \rightarrow [0, +\infty), \quad f(\pi_j) = \sum_{x_i \in \pi_j} \|x_i - c_j\|^2 \quad (5)$$

$$F : \Psi(S, k) \rightarrow [0, +\infty), \quad F(\Pi) = \sum_{j=1}^k f(\pi_j), \quad \Pi = \{\pi_1, \dots, \pi_k\}. \quad (6)$$

Definicija 1. Kažemo da je particija $\Pi^{(o)}$ optimalna u smislu kvadrata l_2 norme ili u smislu najmanjih kvadrata³ onda ako vrijedi

$$F(\Pi^{(o)}) = \min_{\Pi \in \Psi(S, k)} F(\Pi), \quad (7)$$

gdje je funkcija cilja F definirana s (5-6).

Problem određivanja optimalne particije u smislu Varianzkriterium je NP-težak problem (Kogan, 2007).

Primjedba 1. Fermat–Torricelli–Weber problem:

$$C : \sum_{i=1}^3 d_2^2(A_i, T(x, y)) \rightarrow \min_{x, y} \quad M : \sum_{i=1}^3 d_2(A_i, T(x, y)) \rightarrow \min_{x, y}$$

1.1 Klasteri na $\mathbb{R}$

Norma na $\mathbb{R}$: $\|x\| = |x|$ za $x \in \mathbb{R}$;

Udaljenost točaka $x, y \in \mathbb{R}$ definirana je s $d(x, y) = |x - y|$;

$\Pi = \{\pi_1, \dots, \pi_k\}$ particija osnovnog skupa $S = \{x_i \in \mathbb{R} : i = 1, \dots, m\}$ s k klastera.

Definiramo centre c_j klastera π_j , $j = 1, \dots, k$ i vrijednost funkcije cilja na π_j

$$c_j = \operatorname{argmin}_{\alpha} \sum_{x_i \in \pi_j} |x_i - \alpha| = \operatorname{med}_{x_i \in \pi_j} x_i, \quad f(\pi_j) = \sum_{x_i \in \pi_j} |x_i - \operatorname{med}_{x_i \in \pi_j} x_i|,$$

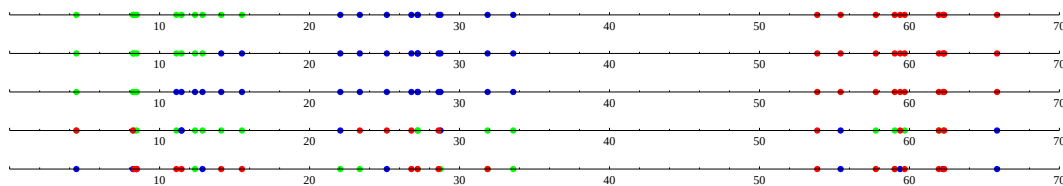
$$c_j^{LS} = \operatorname{argmin}_{\alpha} \sum_{x_i \in \pi_j} (x_i - \alpha)^2 = \frac{1}{|\pi_j|} \sum_{x_i \in \pi_j} x_i, \quad f_{LS}(\pi_j) = \sum_{x_i \in \pi_j} (x_i - c_j^{LS})^2.$$

Zadatak 4. Pokažite da se vrijednost kriterijske funkcije cilja f na nekom klasteru π sa sortiranim elementima $x_1, \dots, x_\mu \in \mathbb{R}$ može dobiti bez određivanja centra (korištenjem jednog teorema iz članka o medianu) na sljedeći način

$$f(\pi) = \sum_{i=1}^r (x_{\mu-i+1} - x_i), \quad r = \lfloor \frac{\mu}{2} \rfloor. \quad (8)$$

Primjer 1. Za $k = 3$ centara: $c_1 = 10$, $c_2 = 30$, $c_3 = 60$ definirano je po 10 normalno distribuiranih podataka $\mathcal{N}(c_i, 4^2)$, $i = 1, 2, 3$. Njihova unija određuje osnovni skup $S = \{x_1, \dots, x_{30}\}$.

³Varianzkriterium, Least Squares

Slika 1: Odozgo prema dolje, particije $\Pi_1, \Pi_2, \Pi_3, \Pi_4, \Pi_5$

Particija	p norma				LS			
	c_1	c_2	c_3	F	c_1	c_2	c_3	F
Π_1	11.3	27.2	59.5	77.8	10.7	27.5	59.7	323.7
Π_2	9.8	27.0	59.5	95.3	9.7	25.3	59.7	548.2
Π_3	8.3	24.3	59.5	137.1	7.4	22.0	59.7	1044.3
Π_4	27.2	25.4	27.7	589.3	29.9	32.5	35.4	14656.3
Π_5	28.0	25.2	27.9	569.2	29.8	33.6	33.6	13952.9

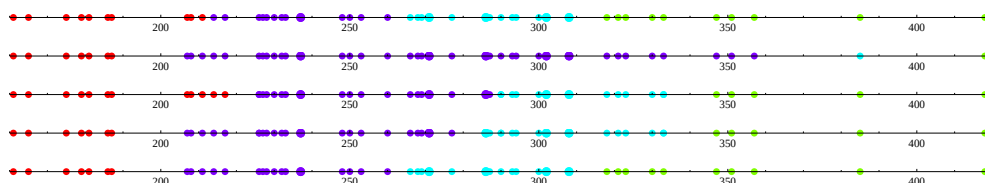
Tablica 1: Centri i funkcije cilja za razne particije

Izabrano je 5 tročlanih particija $\Pi_1, \dots, \Pi_5$, od kojih je Π_1 ona koja je nastala generiranjem podataka (podudara se s particijom koju daje Mathematica naredba **FindClusters**). Za svaku particiju izračunati su centri i vrijednost funkcije cilja u smislu običnih udaljenosti i kvadrata udaljenosti (Tablica 1). Na Slici 1 vide se particije s klasterima i njihovim centrima.

Primjer 2. Svakom studentu pridružujemo ukupan broj bodova postignutih na kolokvijima tijekom semestra. Studente treba razdijeliti u 4 klastera prema postignutom broju bodova: ocjena (5), ocjena (4), ocjena (3) i ocjena (2)).

$$x = \{418, 385, 357, 351, 347, 333, 330, 323, 321, 318, 308, 308, 302, 302, 300, 294, 293, 290, 287, 286, 286, 277, 271, 271, 269, 268, 266, 260, 253, 250, 248, 237, 237, 233, 232, 230, 228, 227, 226, 217, 214, 211, 208, 207, 187, 186, 181, 179, 175, 165, 161\}$$

Izabrano je 5 četveročlanih particija $\Pi_1, \dots, \Pi_5$, od kojih je Π_1 ona koju definira Mathematica naredba **FindClusters** metodom “Optimize”, Π_2 ona koju definira Mathematica naredba **FindClusters** metodom “Agglomerate”, dok su ostale odabrane bez posebnog kriterija.

Slika 2: Odozgo prema dolje, particije $\Pi_1, \Pi_2, \Pi_3, \Pi_4, \Pi_5$

Particija	odličan	vrlo dobar	dobar	dovoljan	F^p	F^{LS}
Π_1	[318-418] (10)	[266-308] (17)	[214-260] (14)	[161-211] (10)	718	17657
Π_2	[418-418] (1)	[385-385] (1)	[207-357] (42)	[161-187] (7)	1545	74369
Π_3	[347-418] (5)	[290-333] (15)	[226-290] (22)	[161-217] (12)	936	22054
Π_4	[347-418] (5)	[286-333] (18)	[207-277] (24)	[161-187] (7)	854	20348
Π_5	[318-418] (10)	[266-308] (17)	[207-260] (17)	[161-187] (7)	701	17187

Tablica 2: Intervali particija s brojem studenata u njima i vrijednosti funkcija cilja

1.2 Klasteri na $\mathbb{R}^2$

Norma na $\mathbb{R}^2$: $\|x\|_p$ za $x \in \mathbb{R}^2$ i $p \in [1, \infty) \cup \{+\infty\}$;

Udaljenost $x, y \in \mathbb{R}^2$ za spomenuti p definirana je s $d_p(x, y) = \|x - y\|_p$;

$\Pi = \{\pi_1, \dots, \pi_k\}$ particija osnovnog skupa $S = \{x_i \in \mathbb{R}^2 : i = 1, \dots, m\}$ s k klastera.

Za $p \in [0, \infty) \cup \{+\infty\}$ definiramo centre klastera π_j , $j = 1, \dots, k$,

$$c_j^p = \operatorname{argmin}_{\alpha \in \mathbb{R}^2} \sum_{x_i \in \pi_j} \|x_i - \alpha\|_p, \quad f_p(\pi_j) = \sum_{x_i \in \pi_j} \|x_i - c_j^p\|_p,$$

$$c_j^{LS} = \operatorname{argmin}_{\alpha \in \mathbb{R}^2} \sum_{x_i \in \pi_j} \|x_i - \alpha\|_2^2 = \frac{1}{|\pi_j|} \sum_{x_i \in \pi_j} x_i, \quad f_{LS}(\pi_j) = \sum_{x_i \in \pi_j} \|x_i - c_j^{LS}\|_2^2.$$

Zadatak 5. Pokažite da se vrijednost f_1 kriterijske funkcije na nekom klasteru π s elementima $x_i = (x_i^{(1)}, \dots, x_i^{(n)})^T$, $i \in \{1, \dots, \mu\}$, može dobiti poopćenjem (8) bez određivanja centra, korištenjem formule

$$f_1(\pi) = \sum_{l=1}^n \sum_{i=1}^r (x_{\mu-i+1}^{(l)} - x_i^{(l)}), \quad r = \lfloor \frac{\mu}{2} \rfloor. \quad (9)$$

uz pretpostavku da su podaci $x_1^{(l)}, \dots, x_\mu^{(l)}$ (l -te komponente vektora klastera π) sortirani za svaki $l = 1, \dots, n$.

Rješenje: Neka je $f_1(\pi) = \sum_{x_i \in \pi} \|x_i - c\|_1$. Sukladno Zadatku 3a je $c = \left(\operatorname{med}_i x_i^{(1)}, \dots, \operatorname{med}_i x_i^{(n)} \right)^T$.

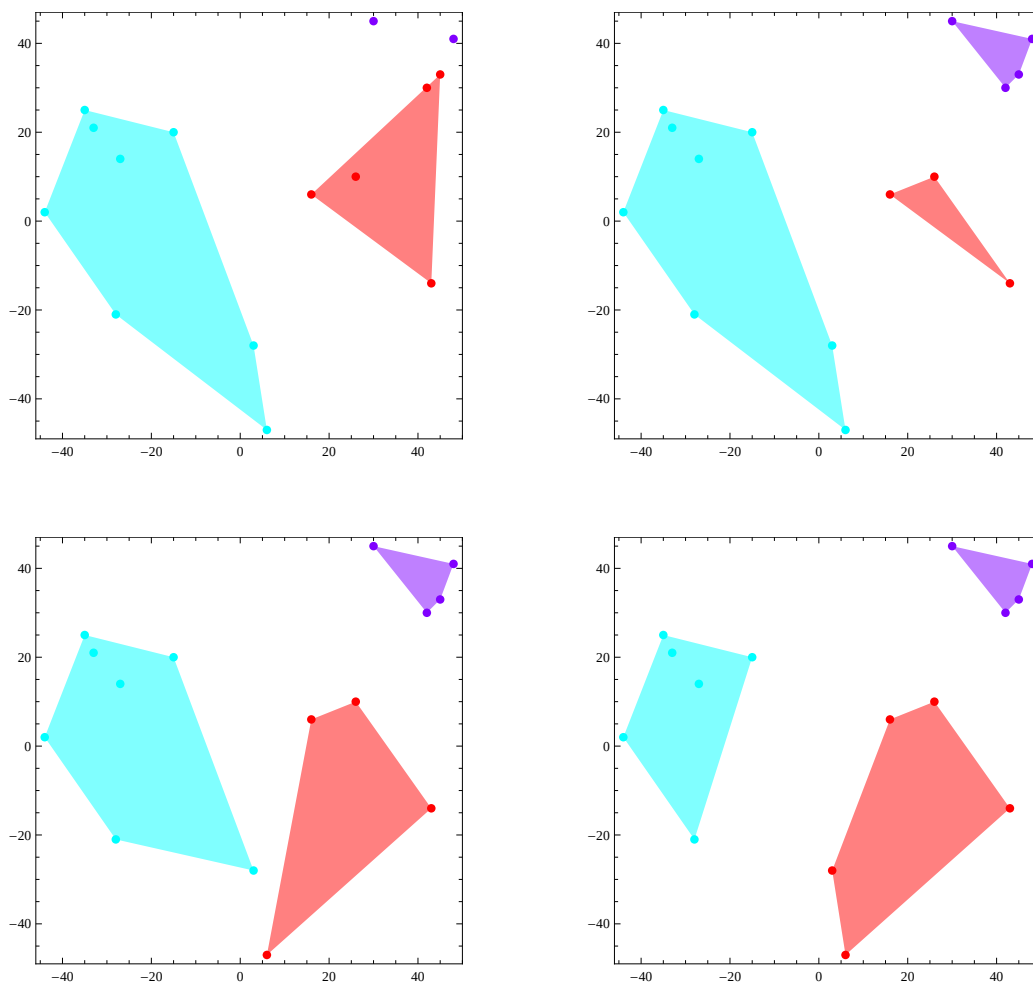
Zato je (korištenjem jednog teorema iz članka o medianu)

$$\begin{aligned} f_1(\pi) &= \sum_{x_i \in \pi} \|x_i - c\|_1 = \sum_{x_i \in \pi} \sum_{l=1}^n |x_i^{(l)} - c^{(l)}| = \sum_{l=1}^n \left(\sum_{x_i \in \pi} |x_i^{(l)} - c^{(l)}| \right) \\ &= \sum_{l=1}^n \left(\sum_{x_i \in \pi} |x_i^{(l)} - \operatorname{med}_i x_i^{(l)}| \right) = \sum_{l=1}^n \sum_{i=1}^r (x_{\mu-i+1}^{(l)} - x_i^{(l)}), \quad r = \lfloor \frac{\mu}{2} \rfloor. \end{aligned}$$

Primjer 3. Za $k = 3$ centara: $c_1 = \{10, 10\}$, $c_2 = \{30, 30\}$, $c_3 = \{45, 15\}$ definirano je po 10 podataka dobivenih pomoću bivarijantnih normalnih distribucija s očekivanjima c_1 , c_2 i c_3 i matricom kovarijacije $\Sigma = \begin{bmatrix} 4^2 & 0 \\ 0 & 4^2 \end{bmatrix}$. Njihova unija određuje osnovni skup $S = \{x_1, \dots, x_{30}\}$.

Izabrano je 4 tročlanih particija $\Pi_1, \dots, \Pi_4$, od kojih je Π_1 ona dobivena samim generiranjem podataka, Π_2 ona koju definira *Mathematica* naredba `FindClusters` metodom “Optimize”, Π_3

ona koju definira *Mathematica* naredba `FindClusters` metodom “Agglomerate”, dok je particija Π_4 odabrane bez posebnog kriterija. Za svaku particiju izračunati su centri i vrijednost funkcije cilja u smislu l_1, l_2, l_∞ udaljenosti i kvadrata udaljenosti (Tablica ??). Na Slici ?? vide se particije s klasterima i njihovim centrima.



Partition	Objective function			
	l_1 -norm	l_2 -norm	l_∞ -norm	LS
Π_1	420	326.89	278.0	9666.58
Π_2	376	284.42	241.5	8402.21
Π_3	354	269.42	234.0	7268.14
Π_4	308	248.53	215.5	5603.67

1.3 Least Squares kriterij optimalnosti i trag simetrične matrice

Zadatak 6. Neka je $A \in \mathbb{R}^{n \times n}$ kvadratna matrica s elementima a_{ij} . Broj

$$\text{tr } A := \sum_{i=1}^n a_{ii},$$

nazivamo trag matrice A . Pokažite da vrijedi:

- (i) $\text{tr}(A + B) = \text{tr } A + \text{tr } B$;
- (ii) $\text{tr}(\lambda A) = \lambda \text{tr } A$;
- (iii) $\text{tr}(AB) = \text{tr}(BA)$;
- (iv) ako je A simetrična, onda $\text{tr}(A) = \sum_{i=1}^n \lambda_i$, gdje su λ_i svojstvene vrijednosti matrice A ;
- (v) $x \in \mathbb{R}^n$, $\text{tr}(xx^T) = x^T x = \|x\|^2$;
- (vi) $A \in \mathbb{R}^{m \times n}$, $\text{tr}(AA^T) = \text{tr}(A^T A) = \sum_{i=1}^n \sum_{j=1}^m a_{ij}^2 = \|A\|_F^2$;
- (vii) $A, B \in \mathbb{R}^{n \times n}$, pri čemu je B nesingularna, $\text{tr}(B^{-1}AB) = \text{tr}(A)$.

Zadatak 7. Pokažite da je za $x \neq 0$, $xx^T \in \mathbb{R}^{n \times n}$ simetrična matrica⁴ ranga 1.

Zadatak 8. Neka je π_j neki klaster sastavljen od $m_j = |\pi_j|$ elemenata u particiji $\Pi \in \Psi(S, k)$, gdje je $S = \{x_i = (x_1^{(i)}, \dots, x_n^{(i)})^T \in \mathbb{R}^n : i = 1, \dots, m\}$. Pokažite da za svaki $a \in \mathbb{R}^n$ vrijedi

$$\sum_{x_i \in \pi_j} (x_i - a)(x_i - a)^T = \sum_{x_i \in \pi_j} (x_i - c_j)(x_i - c_j)^T + m_j(c_j - a)(c_j - a)^T, \quad (10)$$

gdje je c_j centroid klastera π_j .

Rješenje:

$$\begin{aligned} \sum_{x_i \in \pi_j} (x_i - a)(x_i - a)^T &= \sum_{x_i \in \pi_j} ((x_i - c_j) + (c_j - a))((x_i - c_j) + (c_j - a))^T \\ &= \sum_{x_i \in \pi_j} (x_i - c_j)(x_i - c_j)^T + m_j(c_j - a)(c_j - a)^T, \end{aligned}$$

$$\text{jer je} \quad \sum_{x_i \in \pi_j} (x_i - c_j)(x_i - a)^T = \left(\sum_{x_i \in \pi_j} (x_i - c_j) \right) (x_i - a)^T = 0,$$

$$\text{jer je} \quad \sum_{x_i \in \pi_j} (x_i - c_j) = 0.$$

Ako u (10) umjesto a stavimo $c = \frac{1}{m} \sum_{i=1}^m x_i$, sve prosumiramo po $j = 1, \dots, k$ i uočimo da lijeva strana ne ovisi o particiji Π , dobivamo

$$\sum_{i=1}^m (x_i - c)(x_i - c)^T = \sum_{j=1}^k \sum_{x_i \in \pi_j} (x_i - c_j)(x_i - c_j)^T + \sum_{j=1}^k m_j(c_j - c)(c_j - c)^T, \quad (11)$$

odnosno

⁴U literaturi (Späth, 1983) može se naći pod nazivom "diadski produkt"

Teorem 1. Uz prethodne oznake vrijedi

$$T = W(\Pi) + B(\Pi), \quad (12)$$

gdje je

$$T = \sum_{i=1}^m (x_i - c)(x_i - c)^T, \quad (13)$$

$$W(\Pi) = \sum_{j=1}^k \sum_{x_i \in \pi_j} (x_i - c_j)(x_i - c_j)^T, \quad (14)$$

$$B(\Pi) = \sum_{j=1}^k m_j (c_j - c)(c_j - c)^T. \quad (15)$$

Primjedba 2. Korištenjem Zadatka 6, (11) možemo zapisati kao

$$T = \sum_{i=1}^m \|x_i - c\|^2 = \sum_{j=1}^k \sum_{x_i \in \pi_j} \|x_i - c_j\|^2 + \sum_{j=1}^k m_j \|c_j - c\|^2 = W(\Pi) + B(\Pi). \quad (16)$$

Primijetite da su matrice $W, B, T \in \mathbb{R}^{n \times n}$ simetrične, da T ne ovisi, a W, B ovise o particiji Π i da se kriterij (7) može zapisati kao

$$F(\Pi^{(o)}) = \min_{\Pi \in \Psi(S, k)} \text{tr } W(\Pi), \quad (17)$$

jer je

$$\text{tr } W(\Pi) = \text{tr} \sum_{j=1}^k \sum_{x_i \in \pi_j} (x_i - c_j)(x_i - c_j)^T = \sum_{j=1}^k \sum_{x_i \in \pi_j} \|x_i - c_j\|^2 =: F(\Pi). \quad (18)$$

Primjedba 3. Primijetite da Teorem 1 ne vrijedi ako bi umjesto $f(\pi_j) = \sum_{x_i \in \pi_j} \|x_i - c_j\|^2$ koristili Kovarianzmatricu $\frac{1}{m_j - 1} f(\pi_j)$ ili matrice $\frac{1}{m_j} f(\pi_j)$. Zato se umjesto (7) ne koristiti kriterij (**obrazložite!**)

$$F(\Pi^{(o)}) = \min_{\Pi \in \Psi(S, k)} \sum_{j=1}^k \frac{1}{m_j} \sum_{x_i \in \pi_j} \|x_i - c_j\|^2.$$

Zadatak 9. Dokažite da vrijedi

$$\sum_{i=1}^m \|x_i - c\|^2 = \sum_{j=1}^k \sum_{x_i \in \pi_j} \|x_i - c_j\|^2 + \sum_{j=1}^k m_j \|c_j - c\|^2,$$

Primjedba 4. Ako su $\varphi, \psi \in C^2(\mathbb{R})$, takve funkcije da je $\varphi(x) + \psi(x) = \text{const}$, onda vrijedi

$$\begin{aligned} x_0 \in \mathbb{R}, \quad \varphi'(x_0) = 0 \quad \& \quad \varphi''(x_0) > 0 \quad \Rightarrow \quad \psi'(x_0) = 0 \quad \& \quad \psi''(x_0) < 0, \\ \min_{x \in \mathbb{R}} \varphi(x) = \varphi(x_0) \quad \Rightarrow \quad \max_{x \in \mathbb{R}} \psi(x) = \psi(x_0) \end{aligned}$$

Teorem 2. Uz prethodne oznake vrijedi

$$\operatorname{argmin}_{\Pi \in \Psi(S,k)} \sum_{j=1}^k \sum_{x_i \in \pi_j} \|x_i - c_j\|^2 = \operatorname{argmax}_{\Pi \in \Psi(S,k)} \sum_{j=1}^k m_j \|c_j - c\|^2 =: \Pi^{(o)}; \quad (19)$$

$$F(\Pi^{(o)}) = \min_{\Pi \in \Psi(S,k)} \sum_{j=1}^k \sum_{x_i \in \pi_j} \|x_i - c_j\|^2 = \sum_{i=1}^m \|x_i - c\|^2 - \max_{\Pi \in \Psi(S,k)} \sum_{j=1}^k m_j \|c_j - c\|^2 \quad (20)$$

Ovaj teorem pokazuje da se rješenje problema (7) može svesti na problem traženja maksimuma težinske sume (s težinama m_j) kvadrata udaljenosti centroida pojedinih klastera c_j do centroida čitavog skupa c . Izradite ilustrativni primjer uz odgovarajući grafički prikaz.

Zadatak 10. Pokažite da ako u (10) umjesto a uvrstimo x_s , dobivenu jednakost sumiramo po svim $s \in \pi_j$ i podijelimo s $2m_j$, dobivamo

$$W_j = \frac{1}{2m_j} \sum_{x_s \in \pi_j} \sum_{x_i \in \pi_j} (x_i - x_s)(x_i - x_s)^T. \quad (21)$$

Također pokažite da vrijedi sljedeći teorem formuliran za srednje vrijednosti.

Teorem 3. Uz prethodne oznake vrijedi

$$F(\Pi^{(o)}) = \min_{\Pi \in \Psi(S,k)} \sum_{j=1}^k \sum_{x_i \in \pi_j} \|x_i - c_j\|^2 = \min_{\Pi \in \Psi(S,k)} \sum_{j=1}^k \frac{1}{m_j} \sum_{x_s \in \pi_j} \sum_{x_i \in \pi_j, i > s} \|x_i - x_s\|^2. \quad (22)$$

Zadatak 11. Neka je $S = \{x_i \in \mathbb{R}^n : i = 1, \dots, m\}$ osnovni skup, $k = 2$, $\Pi = \{\pi_1, \pi_2\}$, $m_1 = |\pi_1|$, $m_2 = |\pi_2|$, a c_1, c_2, c redom centriodi klastera π_1, π_2 i centroid skupa S . Pokažite da je u tom slučaju Varianzkriterium ekvivalentan sljedećem:

$$\max_{\Pi \in \mathcal{P}(S,2)} \Phi(\pi_1, \pi_2), \quad \Phi(\pi_1, \pi_2) = \frac{m_1 m_2}{m} \|c_1 - c_2\|^2.$$

Ako je nadalje, (π_1^o, π_2^o) neka optimalna particija i $m_2 > 1$, onda članovi klastera π_2 leže u kugli zadanoj s

$$x_i \in \pi_2^o \Leftrightarrow \|x_i - \left(c - \frac{m_1 m_2}{m} (c_1 - c_2)\right)\|^2 \leq \frac{m_1 m_2 (m_1 + 1)(m_2 - 1)}{m^2} \|c_1 - c_2\|^2.$$

Možete li na osnovi ovoga konstruirati neki specijalni algoritam za traženje optimalne dvočlane particije?

1.4 Transformacija podataka

Teorem 4. Varianzkriterium je invarijantam na rotaciju i translaciju podataka. Preciznije, neka je:

(i) $S = \{x_i \in \mathbb{R}^n : i = 1, \dots, m\}$ skup, a $\Pi = \{\pi_1, \dots, \pi_k\}$ proizvoljna particija s centroidima $c_1, \dots, c_k$ i duljinama klastera $m_j = |\pi_j|$, $j = 1, \dots, k$;

(ii) $Q \in \mathbb{R}^{n \times n}$ ortogonalna matrica i $b \in \mathbb{R}^n$ neki vektor.

(iii) $S' = \{y_i \in \mathbb{R}^n : y_i = Qx_i + b, i = 1, \dots, m\}$

$\Pi' = \{\pi'_1, \dots, \pi'_k\}$ particija s klasterima $\pi'_j = \{y_i \in S' : x_i \in \pi_j\}$.

Tada vrijedi

$$c'_j = Qc_j + b, \quad \text{gdje je} \quad c'_j = \frac{1}{m_j} \sum_{y_i \in \pi'_j} y_i,$$

$$\min_{\Pi' \in \Psi(S', k)} \sum_{j=1}^k \sum_{y_i \in \pi'_j} \|y_i - c'_j\|^2 = \min_{\Pi \in \Psi(S, k)} \sum_{j=1}^k \sum_{x_i \in \pi_j} \|x_i - c_j\|^2.$$

Dokaz. Vrijedi

$$c'_j = \frac{1}{m_j} \sum_{y_i \in \pi'_j} y_i = \frac{1}{m_j} \sum_{x_i \in \pi_j} (Qx_i + b) = Q \left(\frac{1}{m_j} \sum_{x_i \in \pi_j} x_i \right) + b = Qc_j + b,$$

odakle i ostatak tvrdnje lako slijedi. □

Zadatak 12. Izradite ilustrativni primjer uz odgovarajući grafički prikaz.

Zadatak 13. Pokažite da je Varianzkriterium invarijantam na skaliranje podataka samo u slučaju navedenom u sljedećem teoremu. Izradite ilustrativni primjer uz odgovarajući grafički prikaz.

Teorem 5. Uz oznake kao u Teoremu 4 Varianzkriterium je invarijantam na skaliranje podataka $y_i = Hx_i$, $H = \text{diag}(h_1, \dots, h_n)$ onda i samo onda ako je $h_i = \pm 1$, za sve $i = 1, \dots, n$, tj. vrijedi

$$\min_{\Pi \in \Psi(S, k)} \sum_{j=1}^k \sum_{y_i \in \pi'_j} \|y_i - c'_j\|^2 = \min_{\Pi \in \Psi(S, k)} \sum_{j=1}^k \sum_{x_i \in \pi_j} \|x_i - c_j\|^2, \quad (23)$$

gdje je $c_j = \frac{1}{m_j} \sum_{x_i \in \pi_j} x_i$, $c'_j = \frac{1}{m_j} \sum_{y_i \in \pi'_j} y_i$

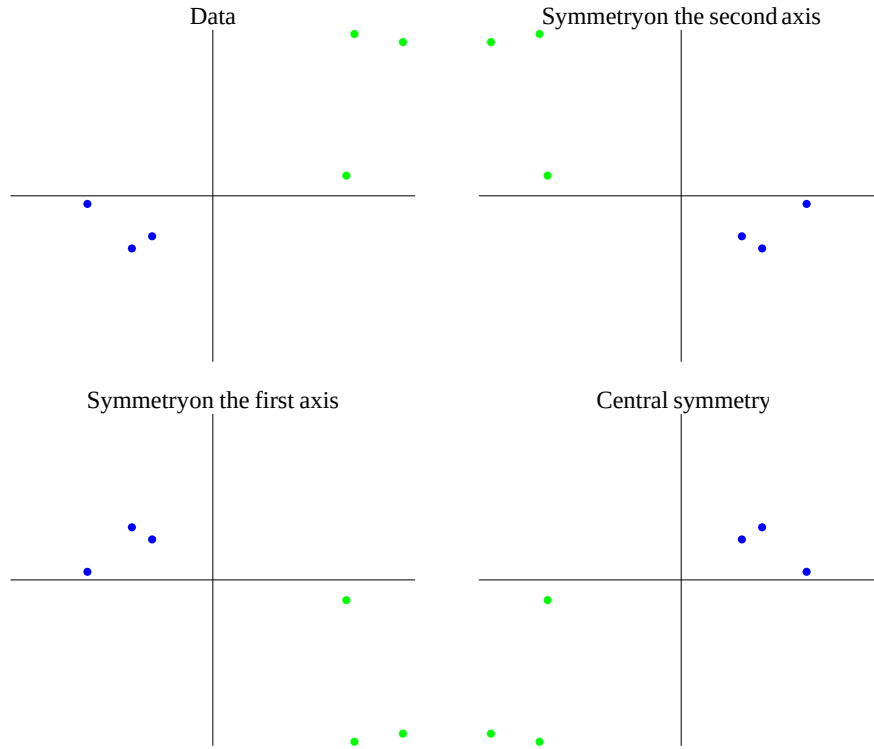
Dokaz.

$$c'_j = \frac{1}{m_j} \sum_{y_i \in \pi'_j} y_i = \frac{1}{m_j} \sum_{x_i \in \pi_j} Hx_i = H \left(\frac{1}{m_j} \sum_{x_i \in \pi_j} x_i \right) = Hc_j.$$

Zato vrijedi

$$\begin{aligned} \sum_{y_i \in \pi'_j} \|y_i - c'_j\|^2 &= \sum_{x_i \in \pi_j} \|H(x_i - c_j)\|^2 = \sum_{x_i \in \pi_j} (x_i - c_j)^T H^T H (x_i - c_j) \\ &= \sum_{x_i \in \pi_j} (x_i - c_j)^T (x_i - c_j) \Leftrightarrow H^T H = I \Leftrightarrow h_i^2 = 1. \end{aligned}$$

□



Slika 3: Simetrije podataka

Transformacija H zapravo predstavlja razne simetrije/refleksije (centralna, osna, ravninska, hiperravninska). Za $h_i = \pm r, r \neq 0$ minimum se postiže na istoj particiji (mada se vrijednost minimuma tada mijenja).

$$\begin{aligned}
 \min_{\Pi} \sum_{y_i \in \pi'_j} \|y_i - c'_j\|^2 &= \min_{\Pi} \sum_{x_i \in \pi_j} (x_i - c_j)^T H^T H (x_i - c_j) \\
 &= r^2 \min_{\Pi} \sum_{x_i \in \pi_j} (x_i - c_j)^T (x_i - c_j) \Leftrightarrow H^T H = r^2 I, \Leftrightarrow h_i^2 = r^2.
 \end{aligned}$$

2 k -means partition (Minimaldistanzpartition)

Definicija 2. Kažemo da je $\hat{\Pi} = \{\hat{\pi}_1, \dots, \hat{\pi}_k\}$ optimalna paricija minimalnih udaljenosti (k -means partition, Minimaldistanzpartition) ako

$$x_i \in \hat{\pi}_j \Leftrightarrow \|x_i - c_j\| = \min_{s=1,k} \|x_i - c_s\|. \quad (24)$$

Standardni k -means algorithm daje optimalnu k -means partition (Minimaldistanzpartition).

Algoritam 1. (k -means algorithm – Minimaldistanzverfahren)

Step 1: Učitaj $1 \leq k \leq m$, it_{max} , $(x_i, i = 1, \dots, m)$, stavi $\mu = 0$, izaberi početnu particiju $\Pi^{(\mu)} = \{\pi_1^{(\mu)}, \dots, \pi_k^{(\mu)}\}$ ⁵ i izračunaj centroide i vrijednost funkcije cilja

$$c_j^{(\mu)} = \frac{1}{m_j^{(\mu)}} \sum_{x_i \in \pi_j^{(\mu)}} x_i, \quad m_j^{(\mu)} = |\pi_j^{(\mu)}|, \quad j = 1, \dots, k;$$

$$F_\mu = F(\Pi^{(\mu)}) = \sum_{j=1}^k \sum_{x_i \in \pi_j^{(\mu)}} \|x_i - c_j^{(\mu)}\|^2;$$

Step 2: Sukladno principu minimalnih udaljenosti (24) definiraj novu particiju

$$\Pi^{(\mu+1)} = \{\pi_1^{(\mu+1)}, \dots, \pi_k^{(\mu+1)}\}$$

$$\pi_1^{(\mu+1)} = \{x_i \in S : \|x_i - c_1^{(\mu)}\| \leq \|x_i - c_l^{(\mu)}\|, l = 1, \dots, k\}$$

$$\pi_j^{(\mu+1)} = \{x_i \in S \setminus \bigcup_{s=1}^{j-1} \pi_s^{(\mu+1)} : \|x_i - c_j^{(\mu)}\| \leq \|x_i - c_l^{(\mu)}\|, \forall l = 1, \dots, k\}, \quad j = 2, \dots, k;$$

Step 3: Izračunaj nove centroide i nove vrijednost funkcije cilja

$$c_j^{(\mu+1)} = \frac{1}{m_j^{(\mu+1)}} \sum_{x_i \in \pi_j^{(\mu+1)}} x_i, \quad j = 1, \dots, k;$$

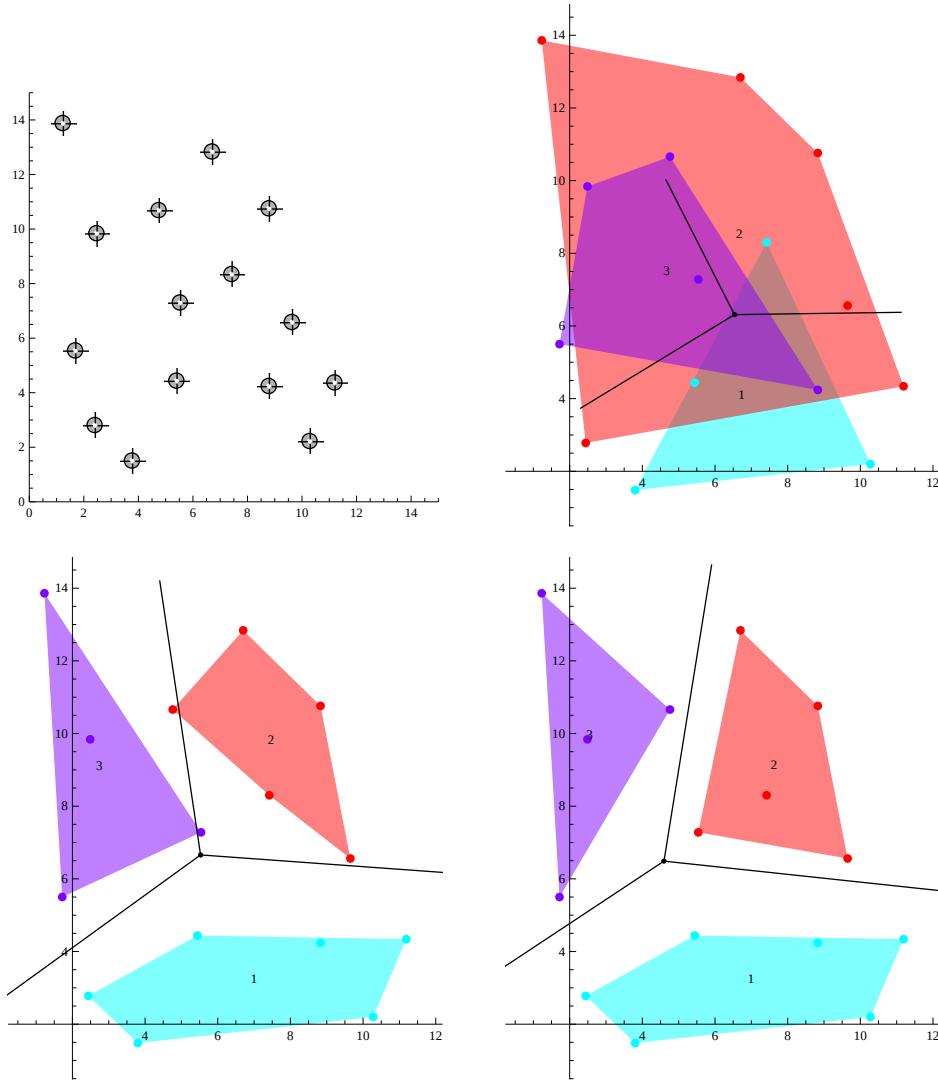
$$F_{\mu+1} = F(\Pi^{(\mu+1)}) = \sum_{j=1}^k \sum_{x_i \in \pi_j^{(\mu+1)}} \|x_i - c_j^{(\mu+1)}\|^2;$$

Step 4: If $F_{\mu+1} < F_\mu$ and if $\mu < it_{max}$, stavi $\mu := \mu + 1$ i idi na Step 2; U protivnom STOP

⁵Alternativno, izaberi neku početnu aproksimaciju centroida $\{c_1^{(\mu)}, \dots, c_k^{(\mu)}\}$ i u smislu minimalnih udaljenosti (24) definiraj početnu particiju $\Pi^{(\mu)} = \{\pi_1^{(\mu)}, \dots, \pi_k^{(\mu)}\}$,

$$\pi_1^{(\mu)} = \{x_i \in S : \|x_i - c_1^{(\mu)}\| \leq \|x_i - c_l^{(\mu)}\|, l = 1, \dots, k\};$$

$$\pi_j^{(\mu)} = \{x_i \in S \setminus \bigcup_{s=1}^{j-1} \pi_s^{(\mu)} : \|x_i - c_j^{(\mu)}\| \leq \|x_i - c_l^{(\mu)}\|, \forall l = 1, \dots, k\}, \quad j = 2, \dots, k;$$

Slika 4: k -means algorithm

Teorem 6. *Algoritam Minimaldistanzverfahren ne povećava vrijednost minimizirajućeg funkcionala F definiranog sa (7), odnosno (17).*

Dokaz.

$$\begin{aligned}
 F(\Pi^{(\mu)}) &= \sum_{j=1}^k \sum_{x_i \in \pi_j^{(\mu)}} \|x_i - c_j^{(\mu)}\|^2 \\
 &\geq \sum_{j=1}^k \sum_{x_i \in \pi_j^{(\mu)}} \min_{s=1,k} \|x_i - c_s^{(\mu)}\|^2 = \sum_{j=1}^k \sum_{x_i \in \pi_j^{(\mu+1)}} \|x_i - c_j^{(\mu)}\|^2 \\
 &\geq \sum_{j=1}^k \sum_{x_i \in \pi_j^{(\mu+1)}} \|x_i - c_j^{(\mu+1)}\|^2 = F(\Pi^{(\mu+1)})
 \end{aligned}$$

□

Teorem 7. (Teorem o razdvajanju) Neka je $\hat{\Pi} = \{\hat{\pi}_1, \dots, \hat{\pi}_k\}$ optimalna (k -means particija. Tada

- $\text{conv}(\hat{\pi}_i) \cap \text{conv}(\hat{\pi}_j) = \emptyset, \quad \forall 1 \leq i \neq j \leq k,$
- hiperravnina koja razdvaja dva različita klastera $\hat{\pi}_p, \hat{\pi}_q$ s centrima c_p, c_q is zadana je s

$$(c_q - c_p)^T \left(x - \frac{1}{2}(c_q + c_p) \right) = 0.$$

Dokaz. Skup svih točaka $x \in \mathbb{R}^n$ jednako udaljenih od točaka c_p i c_q zadan je s

$$H = \{x \in \mathbb{R}^n : \|x - c_q\| = \|x - c_p\|\}.$$

$$\|x\|^2 + \|c_q\|^2 - 2c_q^T x = \|x\|^2 + \|c_p\|^2 - 2c_p^T x,$$

$$\|c_q\|^2 - \|c_p\|^2 = 2(c_q - c_p)^T x,$$

$$(c_q - c_p)^T (c_q + c_p) = 2(c_q - c_p)^T x.$$

□

Primjedba 5. Za optimalnu k -means particiju spomenute hiperravnine cijeli prostor $\mathbb{R}^n$ dijele na disjunktna područja $\{\Omega_1, \dots, \Omega_k\}$, tako da je za svaki $j = 1, \dots, k$, $\pi_j \subset \Omega_j$, a centroid c_j je najbliži centroid za sve elemente iz Ω_j (Kogan, 2007), str.15. Ta subregionizacija naziva se **Voronoi** ili **Dirichlet** particija ili tesalacija (vidi također *Mathematica 7*).

Zadatak 14. k -means algorithm (*Minimaldistanzverfahren*) u nekim slučajevima ne daje optimalnu particiju, a može dovesti i do toga da neki od klastera bude prazan skup. Konstruirajte primjere koji će potvrditi ove tvrdnje. Kako u tom slučaju treba prilagoditi k -means algorithm?

(Kogan, 2007), str. 18-20.

- k -means algorithm u nekim slučajevima ne daje optimalnu particiju:

$S = \{0, 2, 3\}$ – the set

$\Pi^0 = \{\{0, 2\}, \{3\}\}$ – initial partition

$\Pi^o = \{\{0, 2\}, \{3\}\}$ – optimal k -means partition $F(\Pi^o) = 2$

$\Pi^o = \{\{0\}, \{2, 3\}\}$ – optimal partition $F(\Pi^o) = 1$

- k -means algorithm može smanjiti broj klastera:

$S = \{-7, -5, -4, 4, 5, 7\}$ – the set

$\Pi^0 = \{\{-7, -5\}, \{-4, 4\}, \{5, 7\}\}$ – initial partition

$\Pi^o = \{\{-7, -5, -4\}, \emptyset, \{4, 5, 7\}\}$ – optimal k -means partition

3 Druge metode

3.1 Izmještanje elemenata klastera

Neka je $c = \frac{1}{m} \sum_{i=1}^m x_i$ i $a \in \mathbb{R}^n$. Tada

$$\begin{aligned} \sum_{i=1}^m (x_i - a)(x_i - a)^T &= \sum_{i=1}^m ((x_i - c) + (c - a))((x_i - c) + (c - a))^T \\ &= \sum_{i=1}^m (x_i - c)(x_i - c)^T + m(c - a)(c - a)^T, \\ \text{jer je } 2 \sum_{i=1}^m (x_i - c)(c - a)^T &= 2 \left(\sum_{i=1}^m (x_i - c) \right) (c - a)^T = 0, \\ \text{jer je } \sum_{i=1}^m (x_i - c) &= 0. \end{aligned}$$

Primjenom tr dobivamo,

$$\sum_{i=1}^m \|x_i - a\|^2 = \sum_{i=1}^m \|x_i - c\|^2 + m\|a - c\|^2, \quad \forall a \in \mathbb{R}^n. \quad (25)$$

Na osnovi (25) lako se pokaže da vrijedi

Lema 1. Neka su $A = \{x_1, \dots, x_p\}$ i $B = \{y_1, \dots, y_q\}$ dva disjunktna skupa u $\mathbb{R}^n$. Tada vrijedi

$$f(A \cup B) = f(A) + f(B) + p\|c(A) - c\|^2 + q\|c(B) - c\|^2,$$

gdje je

$$c(A) = \frac{1}{p} \sum_{i=1}^p x_i, \quad c(B) = \frac{1}{q} \sum_{i=1}^q y_i, \quad c = c(A \cup B) = \frac{p}{p+q} c(A) + \frac{q}{p+q} c(B).$$

Dokaz. Vrijedi

$$c = c(A \cup B) = \frac{1}{p+q} \left(\sum_{i=1}^p x_i + \sum_{i=1}^q y_i \right) = \frac{p}{p+q} \frac{1}{p} \sum_{i=1}^p x_i + \frac{q}{p+q} \frac{1}{q} \sum_{i=1}^q y_i.$$

zato zbog (25) za $a = y_i$ vrijedi

$$\begin{aligned} f(A \cup B) &= \sum_{i=1}^m \|x_i - c\|^2 + m\|y_i - c\|^2 \\ &= \sum_{i=1}^p \|x_i - c(A)\|^2 + p\|c - c(A)\|^2 + \sum_{i=1}^q \|y_i - c(B)\|^2 + q\|c - c(B)\|^2. \end{aligned}$$

□

Zadatak 15. Odredite

$$\min_{x \in \mathbb{R}^n} g(x), \quad \text{gdje je} \quad g(x) = p\|c(A) - x\|^2 + q\|c(B) - x\|^2.$$

Na osnovi Leme 1 slijedi

Teorem 8. *Ako je $\Pi = \pi_1 \cup \dots \cup \pi_k$, uz $m_j = |\pi_j|$ i ako je $c_j = c(\pi_j)$, $j = 1, \dots, k$, onda je*

$$c(\Pi) = \frac{m_1}{m}c_1 + \dots + \frac{m_k}{m}c_k, \quad m = m_1 + \dots + m_k.$$

Ako je $\pi_i \cap \pi_j = \emptyset$, za $i \neq j$, onda

$$F(\Pi) = \sum_{j=1}^k f(\pi_j) + \sum_{j=1}^k m_j \|c - c_j\|^2, \quad \text{odnosno,}$$

$$\sum_{i=1}^m \|c(\Pi) - x_i\|^2 = \sum_{j=1}^k \sum_{x_i \in \pi_j} \|x_i - c_j\|^2 + \sum_{j=1}^k m_j \|c(\Pi) - c_j\|^2.$$

Zadatak 16. *Dokažite Teorem 8 metodom matematičke indukcije.*

Također, na osnovi Leme 1 slijedi

Lema 2. *Neka su $A = \{x_1, \dots, x_p\}$ i $B = \{y_1, \dots, y_q\}$ dva disjunktne skupa u $\mathbb{R}^n$, $y_{i_0} \in B$, $A^+ = A \cup \{y_{i_0}\}$, $B^- = B \setminus \{y_{i_0}\}$. Tada vrijedi*

$$f(A^+) = f(A) + \frac{p}{p+1} \|c(A) - y_{i_0}\|^2, \quad (26)$$

$$f(B^-) = f(B) - \frac{q}{q-1} \|c(B) - y_{i_0}\|^2. \quad (27)$$

Dokaz. Kako je $f(\{y_{i_0}\}) = 0$ i kako je prema Lemi 1, $c(A^+) = \frac{p}{p+1}c(A) + \frac{1}{p+1}y_{i_0}$, vrijedi

$$\begin{aligned} f(A^+) &= f(A) + p \|c(A^+) - c(A)\|^2 + \|c(A^+) - y_{i_0}\|^2 \\ &= f(A) + \frac{p}{(p+1)^2} \|c(A) - y_{i_0}\|^2 + \frac{p^2}{(p+1)^2} \|c(A) - y_{i_0}\|^2 \\ &= f(A) + \frac{p}{p+1} \|c(A) - y_{i_0}\|^2. \end{aligned}$$

Analogno, prema Lemi 1 je $c(B) = \frac{q-1}{q}c(B^-) + \frac{1}{q}y_{i_0}$, iz čega slijedi $c(B^-) = \frac{q}{q-1}c(B) - \frac{1}{q-1}y_{i_0}$. Zato vrijedi

$$\begin{aligned} f(B) &= f(B^-) + (q-1) \|c(B) - c(B^-)\|^2 + \|c(B) - y_{i_0}\|^2 \\ &= f(B^-) + \frac{q-1}{(q-1)^2} \|c(B) - y_{i_0}\|^2 + \|c(B) - y_{i_0}\|^2 \\ &= f(B^-) + \frac{q}{q-1} \|c(B) - y_{i_0}\|^2. \end{aligned}$$

□

Tvrđnja Leme 1 može se dobiti i korištenjem prikaza pomoću tragova iz Sekcije 1.3.

Zadatak 17. *Pokažite da ako klasteru π_j dodamo neki novi element $\hat{a} \notin \pi_j$, dobijemo novi klaster $\hat{\pi}_j = \pi_j \cup \{\hat{a}\}$ za koji vrijedi:*

$$\begin{aligned} (i) \quad c(\hat{\pi}_j) &= \frac{1}{m_j+1} (m_j c(\pi_j) + \hat{a}), \\ (ii) \quad \text{tr } W(\hat{\pi}_j) &= \text{tr } W(\pi_j) + \frac{m_j}{m_j+1} \|\hat{a} - c(\pi_j)\|^2. \end{aligned}$$

Uputa: najprije izračunajte:

$$\begin{aligned}
 (a) \quad x_i - c(\hat{\pi}_j) &= (x_i - c(\pi_j)) + (c(\pi_j) - c(\hat{\pi}_j)) \\
 &= (x_i - c(\pi_j)) + \frac{1}{m_j+1}(c(\pi_j) - \hat{a}), \quad \forall x_i \in S, \\
 (b) \quad (x_i - c(\hat{\pi}_j))(x_i - c(\hat{\pi}_j))^T &= (x_i - c(\pi_j))(x_i - c(\pi_j))^T + \frac{2}{m_j+1}(x_i - c(\pi_j))(c(\pi_j) - \hat{a}) \\
 &\quad + \frac{1}{(m_j+1)^2}(c(\pi_j) - \hat{a})(c(\pi_j) - \hat{a})^T.
 \end{aligned}$$

Zadatak 18. Pokažite da ako klasteru π_j , $m_j > 1$ oduzmemo neki njegov element $\hat{a} \in \pi_j$, dobivamo novi klaster $\hat{\pi}_j = \pi_j \setminus \{\hat{a}\}$, za koji vrijedi:

$$\begin{aligned}
 (i) \quad c(\hat{\pi}_j) &= \frac{1}{m_j-1}(m_j c(\pi_j) - \hat{a}), \\
 (ii) \quad \text{tr } W(\hat{\pi}_j) &= \text{tr } W(\pi_j) - \frac{m_j}{m_j-1} \|\hat{a} - c(\pi_j)\|^2.
 \end{aligned}$$

Sljedeći teorem pokazuje za koliko će se promijeniti funkcija cilja ako element y_{i_0} iz klastera B prijeđe u klaster A .

Teorem 9. Neka su $A = \{x_1, \dots, x_p\}$ i $B = \{y_1, \dots, y_q\}$, $q > 1$, dva disjunktne skupa u $\mathbb{R}^n$, $y_{i_0} \in B$, $A^+ = A \cup \{y_{i_0}\}$, $B^- = B \setminus \{y_{i_0}\}$. Tada vrijedi

$$F(\{A, B\}) - F(\{A^+, B^-\}) = \frac{q}{q-1} \|c(B) - y_{i_0}\|^2 - \frac{p}{p+1} \|c(A) - y_{i_0}\|^2$$

Dokaz. Kako je

$$\begin{aligned}
 F(\{A, B\}) - F(\{A^+, B^-\}) &= (f(A) + f(B)) - (f(A^+) + f(B^-)) \\
 &= (f(A) - f(A^+)) + (f(B) - f(B^-)),
 \end{aligned}$$

korištenjem Leme 2 neposredno slijedi tvrdnja teorema. □

3.2 Korigirani k -means algoritam

Definicija 3. Neka je $\Pi = \{\pi_1, \dots, \pi_k\}$ neka k -particija skupa S i neka su $i, j \in \{1, \dots, k\}$. Svaku particiju Π' , koja se dobije premještanjem samo jednog vektora $\hat{a} \in \pi_i \in \Pi$ u klaster $\pi_j \in \Pi$, nazivamo **prva varijacija** particije Π .

Za neki vektor $\hat{a} \in \pi_i$, $m_i = |\pi_i| > 1$, prema Teoremu 9 definiramo razliku kriterijske funkcije cilja prije i poslije premještanja tog vektora iz klastera π_i u klaster π_j

$$\Delta_{ij} = F(\Pi) - F(\Pi') = \frac{m_i}{m_i-1} \|c(\pi_i) - \hat{a}\|^2 - \frac{m_j}{m_j+1} \|c(\pi_j) - \hat{a}\|^2,$$

i također pomoćnu razliku

$$\hat{\Delta}_{ij} = \|c(\pi_i) - \hat{a}\|^2 - \|c(\pi_j) - \hat{a}\|^2.$$

Vrijedi

$$\Delta_{ij} - \hat{\Delta}_{ij} = \frac{1}{m_i-1} \|c(\pi_i) - \hat{a}\|^2 + \frac{1}{m_j+1} \|c(\pi_j) - \hat{a}\|^2 \geq 0 \quad \Rightarrow \quad \Delta_{ij} \geq \hat{\Delta}_{ij}.$$

Ako je $\hat{\Delta}_{ij} \geq 0$, bit će i $\Delta_{ij} \geq 0$. To prema Teoremu 9 daje nadu da bi premještanje vektora $\hat{a} \in \pi_i$ u klaster π_j , za koji je $\hat{\Delta}_{ij} \geq 0$, moglo smanjiti vrijednost funkcije cilja. Zato k -means algorithm možemo dopuniti s još jednim korakom prema (Kogan, 2007):

Step 5: Ako postoji vektor $\hat{a} \in S$ iz nekog klastera π_i , $|\pi_i| > 1$, takav da postoji klaster π_j , $j \neq i$, takav da je $\|c(\pi_i) - \hat{a}\|^2 \geq \|c(\pi_j) - \hat{a}\|^2$, onda načinimo prvu varijaciju particije Π premještanjem vektora $\hat{a}$ iz klastera π_i u klaster π_j .

Tako smo konstruirali Korigirani k -means algoritam

Zadatak 19. Kontraprimjerom pokažite da ni korigirani k -means clustering algorithm nužno ne dovodi do optimalne LS particije.

3.3 Metode zamjene

- Pretpostavimo da raspolažemo s nekom particijom $\Pi = \{\pi_1, \dots, \pi_k\}$;
- Neka je $\hat{a}$ prvi element niza $x_1, \dots, x_m \in \mathbb{R}^n$, takav da klaster π_i u kome se on nalazi nije jednočlan;
- Element $\hat{a}$ iz klastera π_i premjestimo redom u sve druge klaster π_j , $j \neq i$ i računamo promjenu vrijednosti funkcije cilja. U tu svrhu s $\Pi^{(j)}$, $j \in \{1, \dots, k\} \setminus \{i\}$ označimo prvu varijaciju particije Π , koja se dobije premještanjem elementa $\hat{a}$ iz klastera π_i u klaster π_j .
- Ako je $\Delta_{ij} \leq 0$ za svaki $j \in \{1, \dots, k\} \setminus \{i\}$, onda se premještanjem elementa $\hat{a}$ ne može postići sniženje vrijednosti funkcije cilja F . U tom slučaju izaberemo sljedeći element u nizu $x_1, \dots, x_m$, takav da klaster u kome se on nalazi nije jednočlan i ponovimo proceduru. Ovo ponavljamo tako dugo dok ne prođemo cijeli niz $x_1, \dots, x_m$. U slučaju da prođemo cijeli niz $x_1, \dots, x_m$ bez smanjenja vrijednosti funkcije cilja, STOP.
- Ako je $\Delta_{ij} > 0$ za neki $j_0 \in \{1, \dots, k\} \setminus \{i\}$, onda se premještanjem elementa $\hat{a}$ iz klastera π_i u klaster π_{j_0} dobiva nova particija $\Pi^{(j_0)}$ na kojoj funkcija cilja F postiže nižu vrijednost. Ako postoji više indeksa $j \in \{1, \dots, k\} \setminus \{i\}$ za koje je $\Delta_{ij} > 0$, možemo izabrati prvi po redu takav j ili onaj koji najviše smanjuje funkciju cilja. Primijetite da u ovom slučaju vrijedi

$$\Delta_{ij} > 0 \Leftrightarrow \frac{m_i}{m_i - 1} \|c(\pi_i) - \hat{a}\|^2 > \frac{m_j}{m_j + 1} \|c(\pi_j) - \hat{a}\|^2. \quad (28)$$

Zadatak 20. Konstruirajte Algoritam zamjene koji će prethodno opisanom metodom od zadane particije $\Pi = \{\pi_1, \dots, \pi_k\}$ odrediti novu bolju particiju $\Pi^{(\nu)}$, ako takva postoji. Izradite odgovarajuće primjere.

Primjedba 6. Ako do kraja provedemo Metodu zamjene, na taj način dobit ćemo optimalnu k -means partition (Minimaldistanzpartition).

Naime, ako je $\hat{a}$ proizvoljni element particije π_i i $\Delta_{ij} \leq 0$ za svaki $j = 1, \dots, k$, tj.

$$\frac{m_i}{m_i - 1} \|c(\pi_i) - \hat{a}\|^2 \leq \frac{m_j}{m_j + 1} \|c(\pi_j) - \hat{a}\|^2, \quad j = 1, \dots, k$$

onda množeći lijevu stranu s $0 < \frac{m_i - 1}{m_i} < 1$, a desnu s $\frac{m_j + 1}{m_j} > 1$, nejdnakost ostaje vrijediti. Zato je

$$\|c(\pi_i) - \hat{a}\|^2 \leq \|c(\pi_j) - \hat{a}\|^2, \quad \forall j = 1, \dots, k,$$

a to znači da je Π optimalna k -means partition (Minimaldistanzpartition).

Teorem 10. *Svaka LS-optimalna (Varianzkriterim) particija je optimalna k-means partition (Minimaldistanzpartition), tj. svojstvo Minimaldistanz (24) je nužan, ali ne i dovoljan uvjet da bi neka particija bila optimalna.*

Dokaz. Za $m = k = 1$ tvrdnja je očigledna.

Neka je $1 < k < m$ i neka je $\Pi^{(o)} = \{\pi_1^o, \dots, \pi_k^o\}$ LS-optimalna particija. Pretpostavimo da $\Pi^{(o)}$ nije i optimalna k-means partition (Minimaldistanzpartition). U tom slučaju postoji $\hat{a}$ iz nekog klastera $\pi_i^o (m_i > 1)$, koji je bliži centru $c(\pi_j)$, $j \neq i$,

$$\|c(\pi_j) - \hat{a}\| < \|c(\pi_i) - \hat{a}\|.$$

Ako ovaj izraz najprije kvadriramo, a zatim lijevu stranu pomnožimo s $0 < \frac{m_j}{m_j+1} < 1$, a desnu s $\frac{m_i}{m_i-1} > 1$, dobivamo nejednakost iz (28), a to znači da primjenom Metode zamjene možemo dobiti bolju particiju, što se protivi pretpostavci da je $\Pi^{(o)}$ optimalna particija. $\square$

4 Još neka svojstva LS-particija

4.1 Monotonost

Teorem 11. *Neka je $1 \leq k < m$ i neka su Π , odnosno Π' LS-optimalne (Varianzkriterijumu) particije duljine k , odnosno duljine $k+1$. Tada vrijedi*

$$F(\Pi') \leq F(\Pi).$$

Dokaz. Primijetite najprije da je za svaku particiju $F(\Pi) \geq 0$.

Pretpostavimo najprije da je $k+1 = m$. U tom slučaju je particija Π' sastavljena samo od jednočlanih skupova, pa je $F(\Pi') = 0$. Zato je u ovom slučaju $F(\Pi') \leq F(\Pi)$.

Pretpostavimo sada da je $k+1 < m$. Neka su $\Pi = \{\pi_1, \dots, \pi_k\}$, odnosno Π' , optimalne particije duljine k , odnosno duljine $k+1$. Zbog $k+1 < m$ postoji $1 \leq j_0 \leq m$, takav da je $m_{j_0} = |\pi_{j_0}| > 1$. Neka je $\hat{a} \in \pi_{j_0}$ i neka je $\hat{\pi}_{j_0} = \pi_{j_0} \setminus \{\hat{a}\} \neq \emptyset$. Prema Lemi 2 vrijedi

$$f(\hat{\pi}_{j_0}) = f(\pi_{j_0}) - \frac{m_{j_0}}{m_{j_0}-1} \|c(\pi_{j_0}) - \hat{a}\|^2 \leq f(\pi_{j_0}) \quad \& \quad f(\hat{\pi}_{j_0}) = f(\pi_{j_0}) \quad \Leftrightarrow \quad \hat{a} = c(\pi_{j_0}). \quad (29)$$

Definiramo pomoćnu particiju (ne nužno LS-optimalnu) duljine $k+1$

$$\Pi'' = \{\pi_1, \dots, \pi_{j_0-1}, \hat{\pi}_{j_0}, \pi_{j_0+1}, \dots, \pi_k, \{\hat{a}\}\}.$$

Kako je $F(\hat{a}) = 0$, zbog (29) i LS-optimalnosti particije Π' dobivamo

$$F(\Pi') \leq F(\Pi'') = \sum_{j=1}^k f(\pi_j'') \leq \sum_{j=1}^k f(\pi_j) \leq F(\Pi).$$

$\square$

Primjedba 7. Primijetite da (29) također pokazuje da povećanjem broja elemenata u klasteru LS-funkcija cilja prima manju vrijednost, osim ako se dodani elementi baš ne podudaraju s centrom klastera. Ovu tvrdnju treba ispitati za druge kriterijske funkcije!!!

4.2 Metoda aglomeracije

Metoda aglomeracije spada u hijerarhijske metode. Loša strana hijerarhijskih metoda traženja klastera je nemogućnost vraćanja i popravljivanja prethodnih loših koraka, tj. nakon što se dva podatka spoje u jedan klaster ili razdvoje u dva ta odluka je konačna.

Neka je zadan skup podataka $S = \{x_i = (x_1^{(i)}, \dots, x_n^{(i)})^T \in \mathbb{R}^n : i = 1, \dots, m\}$. Metoda aglomeracije polazi od particije jednočlanih klastera $\Pi^{(m)} = \{\{x_1\}, \dots, \{x_m\}\} \in \Psi(S, m)$. U svakom koraku po nekom kriteriju odabiru se dva klastera koja se spajaju u jedan. Ti kriteriji temelje se na sličnosti (spajamo klastere s najvećim koeficijentom sličnosti) ili na različitosti (spajamo klastere koji se najmanje razlikuju). Neki od kriterija različitosti mogu biti:

$$\begin{aligned} d_{\min}(\pi_i, \pi_j) &= \min_{x \in \pi_i, x' \in \pi_j} d(x, x') \\ d_{\max}(\pi_i, \pi_j) &= \max_{x \in \pi_i, x' \in \pi_j} d(x, x') \\ d_c(\pi_i, \pi_j) &= d(c_i, c_j), \quad c_i, c_j \text{ centroidi klastera } \pi_i, \pi_j \\ d_{\text{avg}}(\pi_i, \pi_j) &= \frac{1}{m_i m_j} \sum_{x \in \pi_i} \sum_{x' \in \pi_j} d(x, x') \end{aligned}$$

U μ -tom koraku dobit ćemo particiju iz skupa $\Psi(S, m - \mu)$. Dakle, ako želimo dobiti k -članu particiju, bit će potrebno provesti $m - k$ koraka algoritma.

Algoritam 2. (Agglomerative Nesting - AGNES)

- Step 1: Učitaj $1 \leq k \leq m$, $(x_i, i = 1, \dots, m)$, postavi $\mu = 0$ i početnu particiju $\Pi^{(m)} = \{\{x_1\}, \dots, \{x_m\}\}$;
- Step 2: Konstruiraj matricu različitosti particije $\Pi^{(m-\mu)}$, $R_\mu \in \mathbb{R}^{(m-\mu) \times (m-\mu)}$ s elementima $r_{ij} = d(\pi_i, \pi_j)$, gdje $d(\pi_i, \pi_j)$ predstavlja različitost klastera π_i i π_j ⁶.
- Step 3: Pronađi najmanji element $r_{i_0 j_0}^*$ matrice R_μ , i načini particiju $\Pi^{(m-\mu-1)} = (\Pi^{(m-\mu)} \setminus \{\pi_{i_0}, \pi_{j_0}\}) \cup \{\pi_{i_0} \cup \pi_{j_0}\}$.
- Step 4: Ako je $\mu < m - k$, postavi $\mu = \mu + 1$ i idi na Step 2; Inače STOP;

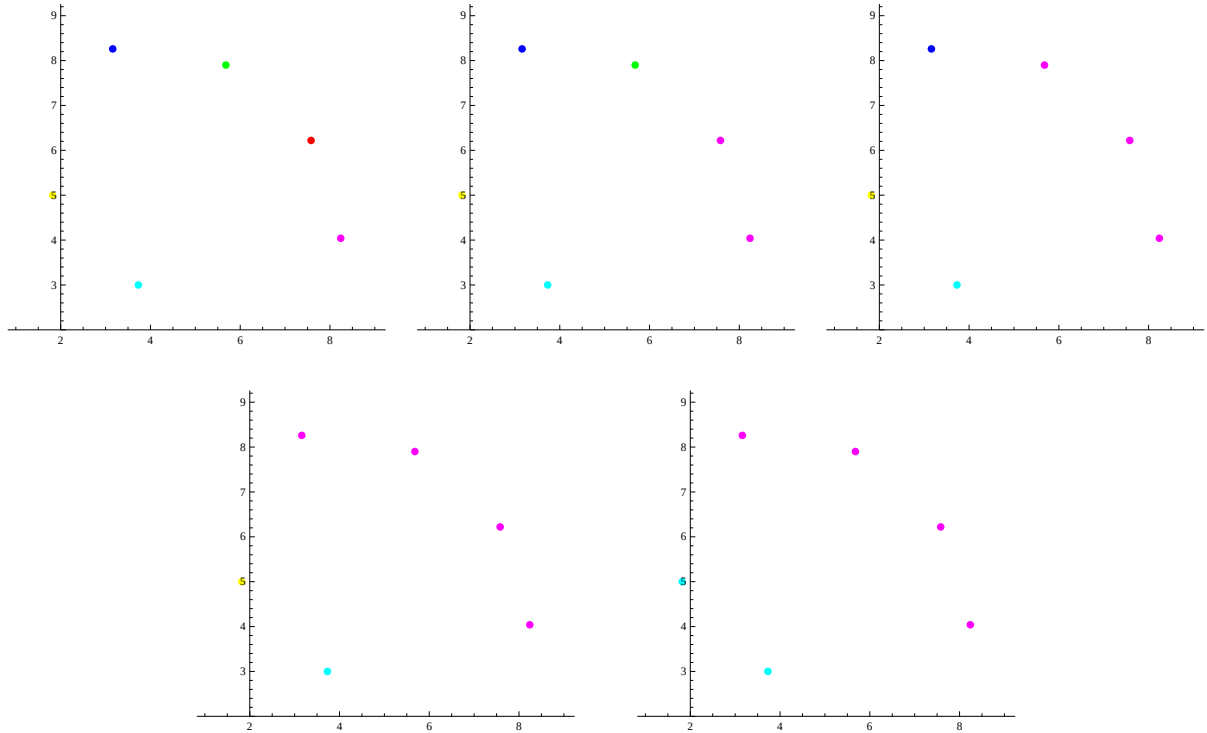
Ovisno o odabiru kriterija, matrice različitosti se nakon prvog koraka mogu mnogo brže računati.

Primjer 4. (Ilustracija metode aglomeracije). Za zadani skup točaka treba potvrditi optimalnu LS-optimalnu particiju sastavljenu od dva klastera.

$$S = \{\{1.82, 5.\}, \{5.68, 7.9\}, \{3.73, 3.\}, \{3.16, 8.26\}, \{8.24, 4.04\}, \{7.58, 6.22\}\}$$

treba

⁶Primjetite da je matrica R simetrična, pa je zbog toga dovoljno pratiti samo njene iznaddijagonalne (ili ispodijagonalne) elemente.



Slika 5: Ilustracija metode aglomeracije

$$\Pi^o = \{\{\{1.82, 5.\}, \{3.73, 3.\}\}, \{\{5.68, 7.9\}, \{8.24, 4.04\}, \{7.58, 6.22\}, \{3.16, 8.26\}\}\}$$

4.3 Stabilnost particije

Definicija 4. Neka je CA neki clustering algorithm. Kažemo da je neka particija Π CA-stabilna onda ako je primjena algoritma CA ne mijenja.

Lema 3. Ako je particija $\Pi = \{\pi_1, \pi_2\}$ stabilna u odnosu na Korigirani k-means algorithm, onda je $c(\pi_1) \neq c(\pi_2)$.

Dokaz. Pretpostavimo suprotno: $c := c(\pi_1) = c(\pi_2)$. Prema Teoremu 9 da bi dvočlana particija $\Pi = \{\pi_1, \pi_2\}$, čiji klasteri imaju isti centroid, bila stabilna, za svaki $y_{i_0} \in \pi_1 \cup \pi_2$ mora biti

$$F(\Pi) - F(\Pi') = \left(\frac{q}{q-1} - \frac{p}{p+1} \right) \|c - y_{i_0}\|^2 = 0,$$

a to je ispunjeno ako je $p + q = 0$, tj. ako je $p = q = 0$, što se protivi pretpostavci da klasteri moraju imati barem jedan element. $\square$

Iz Leme 3 slijedi

Teorem 12. Ako je particija $\Pi = \{\pi_1, \dots, \pi_k\}$ stabilna u odnosu na Korigirani k-means algorithm, onda je $c(\pi_i) \neq c(\pi_j)$ za $i \neq j$.

4.4 Udaljenost dviju particija

Definirat ćemo produkt Π_{12} particija Π_1 i Π_2 skupa S tako da kažemo da je neki klaster element particije Π_{12} onda ako je on neprazni presjek nekog klastera iz Π_1 i nekog klastera iz Π_2 . Preciznije

Definicija 5. Kažemo da je particija Π_{12} produkt particija $\Pi_1 = \{\pi_1^{(1)}, \dots, \pi_q^{(1)}\}$ i $\Pi_2 = \{\pi_1^{(2)}, \dots, \pi_q^{(2)}\}$ skupa S ako su klasteri iz Π_{12} svi neprazni presjeci $\pi_i^{(1)} \cap \pi_j^{(2)}$ klastera iz Π_1 i Π_2 .

Primjer 5. Za skup $S = \{1, 2, 6, 7, 9\}$ definiramo

$$\begin{aligned}\Pi_1 &= \{\{1, 2, 6\}, \{7, 9\}\} \\ \Pi_2 &= \{\{1, 2\}, \{6, 7\}, \{9\}\} \\ \Pi_{12} &= \{\{1, 2\}, \{6\}, \{7\}, \{9\}\}\end{aligned}$$

Primjedba 8. Iz Teorema 8 slijedi $F(\Pi_1) - F(\Pi_{12}) \geq 0$. Stroga nejednakost slijedi iz sljedećeg teorema.

Teorem 13. Neka su $\Pi_1 = \{\pi_1^{(1)}, \dots, \pi_q^{(1)}\}$ i $\Pi_2 = \{\pi_1^{(2)}, \dots, \pi_q^{(2)}\}$ dvije k -means stabilne particije skupa S . Ako je $\Pi_1 \neq \Pi_{12}$, onda je $F(\Pi_1) - F(\Pi_{12}) > 0$.

Dokaz. (Kogan, 2007) □

Sada možemo definirati udaljenost particija Π_1 i Π_2 na sljedeći način (Kogan, 2007)

$$d(\Pi_1, \Pi_2) := (F(\Pi_1) - F(\Pi_{12})) + (F(\Pi_2) - F(\Pi_{12})). \quad (30)$$

Zadatak 21. Pokažite da ako su Π_1 i Π_2 dvije k -means stabilne particije skupa S , onda vrijedi

$$\begin{aligned}(i) \quad & d(\Pi_1, \Pi_2) \geq 0 \text{ i } d(\Pi_1, \Pi_2) = 0 \Leftrightarrow \Pi_1 = \Pi_2 \\ (i) \quad & d(\Pi_1, \Pi_2) = d(\Pi_2, \Pi_1).\end{aligned}$$

Vrijedi li i nejednakost trokuta ?

5 Generalizirani LS-kriterij

Definiramo normu

$$\|\cdot\|_G : \mathbb{R}^n \rightarrow [0, +\infty), \quad \|x\|_G^2 = x^T G x$$

i udaljenost

$$d_G : \mathbb{R}^n \times \mathbb{R}^n \rightarrow [0, +\infty), \quad d_G(x, y) = (x - y)^T G (x - y).$$

Što je jedinična kugla radijusa 1 sa središtem u ishodištu u ovoj metrici ?

Suma kvadrata G -udaljenosti točke $a \in \mathbb{R}^n$ do svih točaka klastera π_j zadana je s

$$F_G(a) = \sum_{x_i \in \pi_j} d_G^2(a, x_i) = \sum_{x_i \in \pi_j} (x_i - a)^T G (x_i - a).$$

Ako je G pozitivno definitna matrica ($G > 0$), onda funkcional F_G u točki

$$c_j = \frac{1}{m_j} \sum_{x_i \in \pi_j} x_i,$$

postiže svoj globalni minimum jer je

$$\begin{aligned}\frac{\partial F_G(c_j)}{\partial a} &= -2 \sum_{x_i \in \pi_j} G(x_i - c_j) = 0 \\ \frac{\partial^2 F_G(c_j)}{\partial a^2} &= 2m_j G > 0.\end{aligned}$$

a generalizirani LS-kriterij *Varianzkriterium* definira se s (Späth, 1983)

$$F_G(P) = \sum_{j=1}^k \sum_{x_i \in \pi_j} (x_i - c_j)^T G(x_i - c_j). \quad (31)$$

6 Neke primjene

6.1 Drugi objekti kao centri klastera

Umjesto točke $c \in \mathbb{R}^n$ (centroid) kao centar klastera može se pojaviti i neki drugi objekti, primjerice: pravci, ravnine, hiperravnine, kružnice itd. Ako primjerice, centar klastera u $\mathbb{R}^2$ treba biti pravac $\mathbf{l}$ (Kogan, 2007), onda udaljenost neke točke $a \in \mathbb{R}^2$ do pravca $\mathbf{l}$ definiramo kao najkraću udaljenost (ortogonalnu projekciju) točke $a \in \mathbb{R}^2$ do pravca $\mathbf{l}$. Za zadani skup točaka $S = \{x_i \in \mathbb{R}^2 : i = 1, \dots, m\}$, $m \geq 2$, pravac $\mathbf{L}$ kao centar klastera $\pi \subseteq S$ definiran je s

$$\sum_{x_i \in \pi} d(\mathbf{L}, x_i) \leq \sum_{x_i \in \pi} d(\mathbf{l}, x_i), \quad \text{za svaki pravac } \mathbf{l} \quad (32)$$

6.2 Sustav linearnih jednadžbi

Specijalno možemo promatrati i sustav linearnih jednadžbi $A\xi = \eta$, $A \in \mathbb{R}^{m \times n}$, $\eta \in \mathbb{R}^m$ (Späth, 1983). Sada bi klaster bio podsustav sustava linearnih jednadžbi odabran prema određenom principu.

Primjer 6. Za dane podatke (t_i, z_i) , $i = 1, \dots, m$ treba pronaći particiju $\Pi = \{\pi_1, \dots, \pi_k\}$, sastavljenu od k najboljih pravaca $t \mapsto \alpha_j^* + \beta_j^* t$. Na osnovi danih podataka formiramo sustav

$$\alpha + \beta t_i = z_i, \quad i = 1, \dots, m.$$

Općenito, sustav $A\xi = \eta$ razdijelimo na k podsustava (klastera) $A^{(j)}\xi = \eta^{(j)}$, $j = 1, \dots, k$ i za svaki podsustav potražimo njegovo LS-rješenje $\tilde{\xi}^{(j)}$ minimizacijom funkcionala

$$\xi \mapsto \|A^{(j)}\xi - \eta^{(j)}\|^2. \quad (33)$$

Vektor $\tilde{\xi}^{(j)}$ ujedno predstavlja centar klastera jednadžbi π_j . Kriterijska funkcija cilja sada je

$$F(\Pi) = \sum_{j=1}^k \|A^{(j)}\tilde{\xi}^{(j)} - \eta^{(j)}\|^2.$$

Minimaldistanzkriterium (k-means) u ovom slučaju znači

$$i\text{-ta jedn. je u } \pi_j \Leftrightarrow \left(\sum_{l=1}^n a_{il}\tilde{\xi}_l^{(j)} - \eta_i \right)^2 \leq \left(\sum_{l=1}^n a_{il}\tilde{\xi}_l^{(s)} - \eta_i \right)^2, \quad \forall s = 1, \dots, k. \quad (34)$$

Rješenje linearnog problema najmanjih kvadrata (33) možemo potražiti rješavanjem sustava normalnih jednadžbi. Zbog jednostavnost promatramo sustav

$$(A^T A)\xi = A^T \eta.$$

Uz pretpostavku da je A punog ranga ($\text{rank } A = n$) dobivamo

$$\xi = (A^T A)^{-1} A^T \eta. \quad (35)$$

Ovaj dio teksta vjerojatno sadrži dosta pogrešaka

Problem ispuštanja ili dodavanja jednadžbe možemo riješiti primjenom Sherman–Morrison–Woodbury leme (see e. g. (Dennis and Schnabel, 1996)).

Lema 4. (Sherman–Morrison–Woodbury) *Neka su $u, v \in \mathbb{R}^n$, $\beta \in \mathbb{R}$ i neka je $A \in \mathbb{R}^{n \times n}$ nesingularna matrica. Tada*

$$(a) \det(A + \beta u^T v) = \det A (1 + \beta v^T A^{-1} u);$$

$$(b) \text{ Matrica } A + \beta u^T v \text{ je nesingularna onda i samo onda ako } 1 + \beta v^T A^{-1} u \neq 0 \text{ i vrijedi}$$

$$(A + \beta u^T v)^{-1} = A^{-1} - \frac{\beta A^{-1} u v^T A^{-1}}{1 + \beta v^T A^{-1} u}. \quad (36)$$

Pretpostavimo najprije da sustavu $A\xi = \eta$ dodamo jednadžbu $u^T \xi = v$, $u \in \mathbb{R}^n, v \in \mathbb{R}$. LS-rješenje ξ' novog sustava dobivamo iz

$$\min_{\xi' \in \mathbb{R}^n} \left\| \begin{bmatrix} A \\ u \end{bmatrix} \xi' - \begin{bmatrix} \eta \\ v \end{bmatrix} \right\|^2,$$

$$\xi' = \left([A^T, u] \begin{bmatrix} A \\ u \end{bmatrix} \right)^{-1} \left([A^T, u] \begin{bmatrix} \eta \\ v \end{bmatrix} \right) = (A^T A + uu^T)^{-1} (A^T \eta + vu),$$

primjenom Sherman–Morrison–Woodbury leme

$$\xi' = \tilde{\xi} - \frac{\tilde{u} u^T \tilde{x}}{1 + u^T \tilde{u}}, \quad \text{gdje je } \tilde{u} = (A^T A)^{-1} u, \quad \tilde{\xi} = \xi + v \tilde{u}.$$

Ako je $A > 0$, onda je $u^T \tilde{u} > 0$. Ako je $A^T A > 0$, onda je $A^T A + uu^T > 0$. Ako sustavu $A\xi = \eta$ oduzmemo jednadžbu $u^T \xi = v$, $u \in \mathbb{R}^n, v \in \mathbb{R}$, dobivamo

$$\xi' = (A^T A - uu^T)^{-1} (A^T \eta - vu).$$

Primjenom Sherman–Morrison–Woodbury leme dobivamo

$$(A^T A - uu^T)^{-1} = (A^T A)^{-1} + \frac{1}{1 + u^T (A^T A)^{-1} u} \tilde{u} \tilde{u}^T, \quad \tilde{u} = (A^T A)^{-1} u.$$

Pri tome $(A^T A - uu^T)^{-1}$ postoji i pozitivno je definitna ako

$$0 \leq u^T (A^T A)^{-1} u < 1.$$

Vrijedi

$$\xi' = \tilde{\xi} - \frac{\tilde{u}u^T\tilde{\xi}}{1 + u^T\tilde{u}}, \quad \tilde{\xi} = \xi - v\tilde{u}.$$

Druga mogućnost je primjena QR-dekompozicije s pivotiranjem (Bartels et al., 1978; Golub and Van Loan, 1996).

U (Späth, 1986) razmatra se *least absolute deviations* kao kriterij.

Literatura

- M. S. Aldenderfer, R. K. Blashfield, *Cluster analysis*, Sage, Newbury Park, 1984.
- P. Bose, A. Maheshwari, P. Morin, *Fast approximations for sum of distances, clustering and the Fermat–Weber problem*, Computational Geometry **24**(2003) 135–146
- R. H. BARTELS, A. R. CONN, J. W. SINCLAIR, *Minimization techniques for piecewise differentiable functions: The l_1 solution to an overdetermined linear system*, SIAM J. Numer. Anal. **15**(1978), 224–241
- R. Cupec, R. Grbić, K. Sabo, R. Scitovski, *Three points method for searching the best least absolute deviations plane*, Applied Mathematics and Computation, **215**(2009) 983–994
- J. E. Dennis, Jr, R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Non-linear Equations*, SIAM, Philadelphia, 1996.
- Inderjit S. Dhillon, Yuqiang Guan and J. Kogan, *Iterative Clustering of High Dimensional Text Data Augmented by Local Search*, 2002
- I. S. Dhillon, Y. Guan, B. Kulis, *Kernel k -means, spectral clustering and normalized cuts*, 2004.
- B. S. Everitt, S. Landau, M. Leese, *Cluster analysis*, Wiley, London, 2001. (OM)
- G. Gan, C Ma, J. Wu, *Data clustering : theory, algorithms, and applications*, SIAM, Philadelphia, 2007. (OM)
- G. H. GOLUB, C. F. VAN LOAN, *Matrix Computation*, The John Hopkins University Press, London, 1996.
- J. Han, M. Kamber, *Data mining: concepts and techniques*, Morgan Kaufmann, 2006.
- J. A. Hartigan, *Clustering Algorithms*, Wiley, 1975.
- L. Kaufman, P. J. Rousseeuw, *Finding groups in data : an introduction to cluster analysis*, John Wiley & Sons, Hoboken, 2005. (OM)
- J. Kogan, *Introduction to Clustering Large and High-Dimensional Data*, Cambridge University Press, 2007.
- K. Jajuga, A. Sokolowski, H. H. Bock, *Classification, clustering and data analysis*, Springer, Berlin, 2002. (OM)

- D. MacKay, Information Theory, Inference and Learning Algorithms, Chapter 20. An Example Inference Task: Clustering, Cambridge University Press, 2003, pp. 284–292
- P. Papić, *Uvod u teoriju skupova*, Hrvatsko matematičko društvo, Zagreb, 2000.
- J. Reese, *Solution methods for the p -median problem: an annotated bibliography*, Networks, Wiley, 2006, 125–142, Takodjer (InterScience www.interscience.wiley.com)
- H. C. Romesburg, *Cluster analysis for researchers*, Lulu Press, Nort Carolina, 2004. (OM)
- K. Sabo, R. Scitovski, *The best least absolute deviations line – properties and two efficient methods*, ANZIAM Journal **50**(2008), 185–198
- H. Späth, *Cluster-Formation und- Analyse*, R. Oldenburg Verlag, München, 1983.
- H. Späth, *Clusterwise linear least absolute deviations regression*, Computing **37**(1986) 371–378
- R. Swiniarski, K. Cios, W. Pedrycz, Witold, *Data mining methods for knowledge discovery*, Kluwer Academic, 1998.
- Stella X. Yu, Jianbo Shi, Multiclass Spectral Clustering (članak i slides), 2003.
- Hongyuan Zha, Xiaofeng He, Chris Ding, Horst Simon, Ming Gu, Spectral Relaxation for K-means Clustering, 2002.