Algoritmi i strukture podataka *
III. Klasteri

Rudolf Scitovski, Martina Briš Alić

2. lipnja 2014.

Sadržaj

*Izborni predmet u 2. semestru sveučilišnog diplomskog studijskog programa Poslovna informatika Ekonomskog fakulteta u Osijeku (30 sati predavanja, 15 sati seminara i 15 sati vježbi, 5 ECTS bodova)

1 Uvod

Definicija 1. Neka je $\mathcal{A} = \{a_i \in \mathbb{R}^n : i = 1, \ldots, m\}$ skup s $m \geq 2$ elemenata. Rastav skupa $\mathcal{A}$ na $1 \leq k \leq m$ disjunktnih nepraznih podskupova $\pi_1, \ldots, \pi_k$, takvih da bude

$$\begin{aligned} &(i) \quad \bigcup_{j=1}^{k} \pi_j = \mathcal{A}, \\ &(ii) \quad \pi_r \bigcap \pi_s = \emptyset, \quad r \neq s, \\ &(iii) \quad m_j := |\pi_j| \geq 1, \quad j = 1, \ldots, k. \end{aligned}$$

zovemo *particija* Π *skupa* $\mathcal{A}$. Elemente particije $\Pi = \{\pi_1, \ldots, \pi_k\}$ zovemo *klasteri*. Skup svih particija skupa $\mathcal{A}$ sastavljenih od k klastera koje zadovoljavaju *(i)-(iii)* označavamo s $\mathcal{P}(\mathcal{A}; k)$.

Nadalje, kad god budemo govorili o particiji skupa $\mathcal{A}$, podrazumijevat ćemo da je ona sastavljena od ovakvih podskupova skupa $\mathcal{A}$. Na taj način svjesno smo iz razmatranja isključili particije, koje sadržavaju prazan skup ili skup $\mathcal{A}$.

Sinonimi: *grupiranje, segmentiranje, klasifikacija, rangiranje*
En.: cluster analysis, clustering, data mining

Može se pokazati (Veljan, 2001) da je broj svih particija skupa $\mathcal{A}$ iz Definicije 1 jednak Stirlingovom broju druge vrste

$$|\mathcal{P}(\mathcal{A}; k)| = \frac{1}{k!} \sum_{j=1}^{k} (-1)^{k-j} \binom{k}{j} j^m. \tag{1}$$

Specijalno

$$\text{za } k = 2: \quad |\mathcal{P}(\mathcal{A}; 2)| = \tfrac{1}{2}(2^m - 2) = 2^{m-1} - 1, \tag{2}$$

$$\text{za } k = 3: \quad |\mathcal{P}(\mathcal{A}; 3)| = \tfrac{1}{2}\left(1 - 2^m + 3^{m-1}\right), \tag{3}$$

Primjer 1. *Broj svih particija skupa $\mathcal{A}$ koje zadovoljavaju Definiciju 1 specijalno za $m = 10, 50, 10^3, 10^6$ i $k = 2, 3, 5, 8, 10$ Vidljiv je u* Tablici 1

$\lvert\mathcal{P}(\mathcal{A}; k)\rvert$	$k = 2$	$k = 3$	$k = 5$	$k = 8$	$k = 10$
$m = 10$	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}

Tablica 1: Broj particija u ovisnosti o broju elemenata i broju klastera

Iz navedenog primjera vidi se da traženje optimalne particije općenito neće biti moguće provesti pretraživanjem čitavog skupa $\mathcal{P}(\mathcal{A}; k)$. Odmah teba reći da problem traženja optimalne particije spada u NP-teške probleme (Gan et al., 2007) nekonveksne optimizacije općenito nediferencijabilne funkcije više varijabli, koja najčešće posjeduje značajan broj stacionarnih točaka.

Primjene:
poljoprivreda (primjerice, razvrstavanje oranica prema plodnosti zemljišta);
biologija (primjerice, klsasifikacija kukaca u grupe)
medicina (primjerice, analiza rengenskih slika)
promet (primjerice, identifikacija prometnih "čepova")
analiza i pretraživanje teksta
analiza klimatskih kretanja
donošenje raznih odluka u tijelima državne i lokalne administracije.
definiranje izbornih sustava

Programska podrška (Sabo et al., 2010):
`http://www.mathos.hr/oml/software.htm`

2 Motivacija: grupiranje u dva klastera na osnovi jednog obilježja

$\mathcal{A} = \{a_1, \ldots, a_m\} \subset \mathbb{R}$ – podskup realnih brojeva
$\Pi(\mathcal{A}) = \{\pi_1, \pi_2\}$ – particija skupa $\mathcal{A}$, takva da vrijedi

$$\pi_1 \cup \pi_2 = \mathcal{A}, \qquad \pi_1 \cap \pi_2 = \emptyset, \qquad m_1 = |\pi_1| \geq 1, \quad m_2 = |\pi_2| \geq 1.$$

$|\mathcal{P}(\mathcal{A}; k)| = 2^{m-1} - 1$ – broj svih ovakvih particija

Primjer 2. *Neka je* $\mathcal{A} = \{1, 3, 4, 8\}$. *U Tablici 2 navedeno je svih* 7 *particija. Postavlja se pitanje koja od njih je najprirodnija – najbolja u smislu*

- interne kompaktnosti, *tj. u smislu da su svi slični/bliski elementi što više na okupu;*
- eksterne razdvojenosti, *tj. u smislu da su elementi pojedinih klastera što više razdvojeni jedni od drugih.*

Vizualno ta svojstva najbolje ispunjava particija $\Pi_4 = \{\{1, 3, 4\}, \{8\}\}$. *Pokušajmo kvantificirati ove kriterije.*

Neka je $d\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ neka kvazimetrička funkcija. Definirajmo centre klastera

$$c_1 = \operatorname*{argmin}_{x \in \mathbb{R}} \sum_{a_i \in \pi_1} d(x, a_i), \qquad c_2 = \operatorname*{argmin}_{x \in \mathbb{R}} \sum_{a_i \in \pi_2} d(x, a_i,$$

i uvedimo sljedecu kriterijsku funkciju cilja *(objective function)*

$$\mathcal{F}(\Pi) = \sum_{a_i \in \pi_1} d(c_1, a_i) + \sum_{a_i \in \pi_2} d(c_2, a_i),$$

Geometrijski, funkcija $\mathcal{F}$ predstavlja ukupno "rasipanje" - zbroj suma udaljenosti elemenata svakog klastera do njegovog centra. Jasno je da sto je vrijednost kriterijske funkcije $\mathcal{F}$ manja time je "rasipanje" manje, a time su i predthodno navedeni kriteriji kompaktnosti i razdvojenosti bolje ispunjeni.

2.1 Kriterij najmanjih kvadrata

Ako je $d(x, y) = (x - y)^2$ LS-kvazimetrička funkcija, onda je

$$c_1 = \frac{1}{m_1} \sum_{a_i \in \pi_1} a_i, \qquad c_2 = \frac{1}{m_2} \sum_{a_i \in \pi_2} a_i, \qquad \mathcal{F}(\Pi) = \sum_{a_i \in \pi_1} (c_1 - a_i)^2 + \sum_{a_i \in \pi_2} (c_2 - a_i)^2.$$

U ovom slučaju vrijednost kriterijske funkcije $\mathcal{F}$ predstavlja sumu "kvadratnog rasipanja" točaka klastera π_1 do centroida c_1 i točaka klastera π_2 do centroida c_2.

Primjer 3. *Za skup* $\mathcal{A} = \{1, 3, 4, 8\}$ *i sve njegove particije treba odrediti pripadne centroide i vrijednosti funkcije cilja* $\mathcal{F}$.

π_1	π_2	c_1	c_2	$\mathcal{F}(\Pi)$		$\mathcal{G}(\Pi)$	
$\{1\}$	$\{3,4,8\}$	1	5	0+14=	14	9+3=	12
$\{3\}$	$\{1,4,8\}$	3	13/3	0+74/3=	24.67	1+1/3=	1.33
$\{4\}$	$\{1,3,8\}$	4	4	0+26=	26	0+0=	0
$\{8\}$	$\{1,3,4\}$	8	8/3	0+14/3=	4.67	16+16/3=	21.33
$\{1,3\}$	$\{4,8\}$	2	6	2+8=	10	8+8=	16
$\{1,4\}$	$\{3,8\}$	5/2	11/2	9/2+25/2=	17	9/2+9/2=	9
$\{1,8\}$	$\{3,4\}$	9/2	7/2	49/2+1/2=	25	1/2+1/2=	1

Tablica 2: Biranje optimalne particije skupa $\mathcal{A} = \{1, 3, 4, 8\}$ na osnovi LS-kriterija

Vidi se (*Tablica 2*) da particija na kojoj funkcija cilja $\mathcal{F}$ postiže najmanju vrijednost odgovara particiji koju smo i ranije vizualno identificirali kao najbolju.

2.1.1 Dualni problem

Neka je

- $\mathcal{A} = \{a_i \in \mathbb{R} \colon i = 1, \ldots, m\} \subset \mathbb{R}$, $m \geq 2$ — skup;
- $c = \frac{1}{m} \sum\limits_{i=1}^{m} a_i$ — centroid skupa $\mathcal{A}$;
- $\Pi = \{\pi_1, \pi_2\}$ — particija s dva klastera π_1, π_2;
- $m_1 = |\pi_1|, \quad m_2 = |\pi_2|$ — broj elemenata klastera;
- $c_1 = \frac{1}{m_1} \sum\limits_{\pi_1} a_i$ — centroid klastera π_1, $\qquad c_2 = \frac{1}{m_2} \sum\limits_{\pi_2} a_i$ — centroid klastera π_2;

Vrijedi:

$$\sum_{\pi_1} (c_1 - a_i) = 0, \tag{4}$$

$$\sum_{\pi_2} (c_2 - a_i) = 0, \tag{5}$$

$$\sum_{\pi_1} (c - a_i)^2 = \sum_{\pi_1} (c_1 - a_i)^2 + m_1 (c_1 - c)^2, \tag{6}$$

$$\sum_{\pi_2} (c - a_i)^2 = \sum_{\pi_2} (c_2 - a_i)^2 + m_2 (c_2 - c)^2. \tag{7}$$

Dokaz (4):

$$\begin{aligned}\sum_{\pi_1}(c_1 - a_i) = &\quad c_1 \sum_{\pi_1} 1 - \sum_{\pi_1} a_i \\ = &\quad m_1 c_1 - m_1 \tfrac{1}{m_1} \sum_{\pi_1} a_i = m_1 c_1 - m_1 c_1 = 0.\end{aligned}$$

Dokaz (6):

$$\begin{aligned}\sum_{\pi_1}(c - a_i)^2 = &\quad \sum_{\pi_1}((c - c_1) + (c_1 - a_i))^2 \\ = &\quad \sum_{\pi_1}(c - c_1)^2 + \sum_{\pi_1}(c_1 - a_i)^2 + 2\sum_{\pi_1}(c - c_1)(c_1 - a_i) \\ = &\quad \sum_{\pi_1}(c_1 - a_i)^2 + m_1(c - c_1)^2 + 2(c - c_1)\sum_{\pi_1}(c_1 - a_i) \\ \overset{(4)}{=} &\quad \sum_{\pi_1}(c_1 - a_i)^2 + m_1(c - c_1)^2\end{aligned}$$

Zadatak 1. *Dokažite formule* (5) *i* (7).

Zbrajanjem jednakosti (6) i (7) dobivamo

$$\sum_{\pi_1}(c - a_i)^2 + \sum_{\pi_2}(c - a_i)^2 = \sum_{\pi_1}(c_1 - a_i)^2 + \sum_{\pi_2}(c_2 - a_i)^2 + m_1(c - c_1)^2 + m_2(c - c_2)^2$$

odnosno

$$\sum_{i=1}^{m}(c - a_i)^2 = \mathcal{F}(\Pi) + \mathcal{G}(\Pi),$$

gdje je

$$\begin{aligned}\mathcal{F}(\Pi) &= \sum_{\pi_1}(c_1 - a_i)^2 + \sum_{\pi_2}(c_2 - a_i)^2, \\ \mathcal{G}(\Pi) &= m_1(c - c_1)^2 + m_2(c - c_2)^2.\end{aligned}$$

Primjedba 1. Ako je $\Pi^\star = \underset{\Pi \in \mathcal{P}(\mathcal{A};2)}{\operatorname{argmin}}$, tj. ako je $\mathcal{F}(\Pi^\star)$ najmanja vrijednost funkcije $\mathcal{F}$ koja se može postići na skupu svih particija $\mathcal{P}(\mathcal{A}; 2)$, što je $\mathcal{G}(\Pi^\star)$?

Zadatak 2. *Za svaku particiju* Π *skupa* $\mathcal{A}$ *iz Primjera 2, str.4, izračunajte vrijednost funkcije* $\mathcal{G}(\Pi)$*, dopunite Tablicu 2 i pokušajte odgovoriti na pitanje postavljeno u prethodnoj primjedbi.*

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

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

Teorem 1. *Uz prethodne oznake vrijedi:*

$$\operatorname*{argmin}_{\Pi\in\mathcal{P}(\mathcal{A};2)} \mathcal{F}(\Pi) = \operatorname*{argmax}_{\Pi\in\mathcal{P}(\mathcal{A};2)} \mathcal{G}(\Pi).$$

Primjer 4. *Skup* $\mathcal{A} = \{1, 3, 4, 8\}$ *iz Primjera 3, str.4, ima* 7 *različitih particija i za sve njih u Tablici 2 je prikazana vrijednost kriterijske funkcije cilja* $\mathcal{G}$*. Kao što se vidi, funkcija* $\mathcal{G}$ *prima maksimalnu vrijednost* 21.33 *na optimalnoj particiji* $\Pi^\star = \{\{1, 3, 4\}, \{8\}\}$.

Primjer 5. *Zadan je skup* $\mathcal{A} = \{0, 3, 6, 9\}$*. Treba pronaći sve njegove dvočlane particije koje zadovoljavaju Definiciju 1, odrediti pripadne centre i vrijednosti kriterijske funkcije cilja* $\mathcal{F}$ *i* $\mathcal{G}$ *uz primjenu LS-kvazimetričke funkcije.*

π_1	π_2	c_1	c_2	$\mathcal{F}(\Pi)$		$\mathcal{G}(\Pi)$	
{0}	{3,6,9}	0	6	0+18	=18	81/4+27/4	= 27
{3}	{0,6,9}	3	5	0+42	=42	9/4+3/4	= 3
{6}	{0,3,9}	6	4	0+42	=42	9/4+3/4	= 3
{9}	{0,3,6}	9	3	0+18	=18	81/4+27/4	= 27
{0,3}	{6,9}	3/2	15/2	9/2+9/2	=9	18+18	= 36
{0,6}	{3,9}	3	6	18+18	=36	9/2+9/2	= 9
{0,9}	{3,6}	9/2	9/2	81/2+9/2	=45	0+0	= 0

Tablica 3: Particije skupa $\mathcal{A} = \{0, 3, 6, 9\}$

Broj svih dvočlanih particija ovog skupa je $\mathcal{P}(\mathcal{A}; 2) = 2^{4-1} - 1 = 7$, a kao što se vidi iz Tablice 3 LS-optimalna particija u ovom slučaju je $\{\{0, 3\}, \{6, 9\}\}$ jer na njoj funkcija cilja $\mathcal{F}$ postiže globalni minimum, a funkcija cilja $\mathcal{G}$ postiže globalni maksimum. Za koje particije se pripadni klasteri nastavljaju jedan na drugi?

2.2 Kriterij najmanjih apsolutnih odstupanja

Ako je $d(x, y) = |x - y|$ LAD-metrička funkcija, onda je

$$c_1 = \operatorname*{med}_{a_i\in\pi_1} a_i, \qquad c_2 = \operatorname*{med}_{a_i\in\pi_2} a_i, \qquad \mathcal{F}(\Pi) = \sum_{a_i\in\pi_1} |c_1 - a_i| + \sum_{a_i\in\pi_2} |c_2 - a_i|.$$

U ovom slučaju vrijednost kriterijske funkcije $\mathcal{F}$ predstavlja sumu "rasipanja" točaka klastera π_1 do centroida c_1 i točaka klastera π_2 do centroida c_2.

Primjer 6. *Za skup* $\mathcal{A} = \{1, 3, 4, 8\}$ *i sve njegove particije treba odrediti LAD-centre i vrijednosti LAD-funkcije cilja* $\mathcal{F}$ *(vidi* Tablicu 4*).*

π_1	π_2	c_1	c_2	$\mathcal{F}(\Pi)$
$\{1\}$	$\{3,4,8\}$	1	4	0+5= 5
$\{3\}$	$\{1,4,8\}$	3	4	0+7= 7
$\{4\}$	$\{1,3,8\}$	4	3	0+7= 7
$\{8\}$	$\{1,3,4\}$	8	3	0+3= 3
$\{1,3\}$	$\{4,8\}$	1	6	2+4= 6
$\{1,4\}$	$\{3,8\}$	1	3	3+5= 8
$\{1,8\}$	$\{3,4\}$	8	4	7+1= 8

Tablica 4: Biranje optimalne particije skupa $\mathcal{A} = \{1, 3, 4, 8\}$ na osnovi LAD-kriterija

2.3 Formulacija problema grupiranja preko centroida

Za dane realne brojeve $c_1, c_2 \in \mathbb{R}$, $c_1 \neq c_2$, primjenom **principa minimalnih udaljenosti** možemo definirati particiju $\Pi = \{\pi_1, \pi_2\}$ skupa $\mathcal{A}$ na sljedeći način:

$$\pi_1 = \{a_i \in \mathcal{A} : d(a_i, c_1) \leq d(a_i, c_2)\},$$
$$\pi_2 = \{a_i \in \mathcal{A} : d(a_i, c_2) < d(a_i, c_1)\},$$

pri čemu treba voditi računa da svaki element skupa $\mathcal{A}$ pridružimo samo jednom klasteru. Zato se problem traženja optimalne particije skupa $\mathcal{A}$ može razmatrati kao sljedeći optimizacijski problem

$$\min_{c_1,c_2\in\mathbb{R}} F(c_1, c_2), \qquad F(c_1, c_2) = \sum_{i=1}^{m} \min\{d(c_1, a_i), d(c_2, a_i)\}, \tag{8}$$

gdje je $F\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$. Ovaj problem u literaturi se pojavljuje pod nazivom *k-median problem* i ekvivalentan je problemu traženja optimalne particije na kojoj kriterijska funkcija cilja $\mathcal{F}$ postiže globalni minimum.

Primjedba 3. Primijetite da klasteri π_1, π_2 ovise o centrima c_1, c_2 i da vrijedi

$$\mathcal{F}(\Pi) := \sum_{a_i\in\pi_1} d(c_1, a_i) + \sum_{a_i\in\pi_2} d(c_2, a_i) = \sum_{i=1}^{m} \min\{d(c_1, a_i), d(c_2, a_i)\} =: F(c_1, c_2). \tag{9}$$

Naime, vrijedi

$$\begin{aligned} F(c_1, c_2) &= \sum_{a_i\in\pi_1(c_1,c_2)} \min\{d(c_1, a_i), d(c_2, a_i) + \sum_{a_i\in\pi_2(c_1,c_2)} \min\{d(c_1, a_i), d(c_2, a_i) \\ &= \sum_{a_i\in\pi_1(c_1,c_2)} d(c_1, a_i) + \sum_{a_i\in\pi_2(c_1,c_2)} d(c_2, a_i) = \mathcal{F}(\Pi). \end{aligned}$$

3 Grupiranje u k klastera na osnovi jednog obilježja

Neka je $\mathcal{A} = \{a_1, \ldots, a_m\}$ skup koji na osnovi jednog obilježja treba grupirati u k klastera $\pi_1, \ldots, \pi_k$, koji zadovoljavaju Definiciju 1, str.2. Primjerice, dane u godini možemo grupirati prema prosječnoj dnevnoj temperaturi izraženoj u °C. Svaki element $a \in \mathcal{A}$ temeljem tog obilježja reprezentirat ćemo jednim realnim brojem, kojeg ćemo također označavati s a. Zato ćemo nadalje govoriti o skupu podataka-realnih brojeva $\mathcal{A} = \{a_1, \ldots, a_m\}$ među kojima može biti i jednakih. Možemo također koristiti i termine: *m-torka realnih brojeva* ili *konačni niz realnih brojeva*.

Ako je zadana neka kvazimetrička funkcija $d\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$, onda svakom klasteru $\pi_j \in \Pi$ možemo pridružiti njegov centar c_j na sljedeći način

$$c_j = c(\pi_j) := \operatorname*{argmin}_{x \in \mathbb{R}} \sum_{a \in \pi_j} d(x, a), \qquad j = 1, \ldots, k. \tag{10}$$

Nadalje, ako na skupu svih particija $\mathcal{P}(\mathcal{A}; k)$ skupa $\mathcal{A}$ sastavljenih od k klastera definiramo kriterijsku funkciju cilja $\mathcal{F} : \mathcal{P}(\mathcal{A}; k) \to \mathbb{R}_+$,

$$\mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{a \in \pi_j} d(c_j, a), \tag{11}$$

onda d-optimalnu particiju $\Pi^\star$ tražimo rješavanjem sljedećeg optimizacijskog problema

$$\mathcal{F}(\Pi^\star) = \min_{\Pi \in \mathcal{P}(\mathcal{A};k)} \mathcal{F}(\Pi). \tag{12}$$

Primijetite da na taj način optimalna particija $\Pi^\star$ ima svojstvo da je suma "rasipanja" (suma odstupanja) elemenata klastera oko svog centra minimalna. Na taj način nastojimo postići što bolju unutrašnju kompaktnost i međusobnu razdvojenost (separiranost) klastera.

3.1 Kriterij najmanjih kvadrata

Ako je $d\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$, $d(x, y) = (x - y)^2$ govorimo o kvazimetričkoj funkciji najmanjih kvadrata (LS-kvazimetrička funkcija)[1]. U tom slučaju centri $c_1, \ldots, c_k$ klastera $\pi_1, \ldots, \pi_k$ određeni su s

$$c_j = \operatorname*{argmin}_{x \in \mathbb{R}} \sum_{a \in \pi_j} (x - a)^2 = \frac{1}{|\pi_j|} \sum_{a \in \pi_j} a, \qquad j = 1, \ldots, k, \tag{13}$$

a funkcija cilja (11), s

$$\mathcal{F}(c_1, \ldots, c_k) = \sum_{j=1}^{k} \sum_{a \in \pi_j} (c_j - a)^2. \tag{14}$$

[1] En.: Least Squares = najmanji kvadrati

3.1.1 Dualni problem

Sljedeća lema pokazuje da je "rasipanje" skupa $\mathcal{A}$ oko njegovog centra c jednako zbroju "rasipanja" klastera π_j, $j = 1, \ldots, k$, oko njihovih centara c_j, $j = 1, \ldots, k$, i težinskoj sumi kvadrata odstupanja centra c od centara c_j, pri čemu su težine određene veličinom skupova π_j.

Lema 1. *Neka je $\mathcal{A} = \{a_1, \ldots, a_m\}$ skup podataka, a $\Pi = \{\pi_1, \ldots, \pi_k\}$ neka particija s klasterima $\pi_1, \ldots, \pi_k$ duljine $m_1, \ldots, m_k$. Neka je nadalje*

$$c = \frac{1}{m} \sum_{i=1}^{m} a_i, \qquad c_j = \frac{1}{m_j} \sum_{a \in \pi_j} a, \quad j = 1, \ldots, k, \tag{15}$$

gdje je $m_j = |\pi_j|$. Tada vrijedi

$$\sum_{i=1}^{m} (c - a_i)^2 = \mathcal{F}(c_1, \ldots, c_k) + \mathcal{G}(c_1, \ldots, c_k), \tag{16}$$

gdje je

$$\mathcal{F}(c_1, \ldots, c_k) = \sum_{j=1}^{k} \sum_{a_i \in \pi_j} (c_j - a_i)^2, \tag{17}$$

$$\mathcal{G}(c_1, \ldots, c_k) = \sum_{j=1}^{k} m_j (c_1, \ldots, c_k)(c_j - c)^2. \tag{18}$$

Dokaz. Primijetimo najprije da za svaki $x \in \mathbb{R}$ vrijedi

$$\sum_{a \in \pi_j} (x - a)^2 = \sum_{a \in \pi_j} (c_j - a)^2 + m_j (c_j - x)^2, \quad j = 1, \ldots, k. \tag{19}$$

Naime, kako je $\sum\limits_{a \in \pi_j} (c_j - a)(c_j - x) = (c_j - x) \sum\limits_{a \in \pi_j} (c_j - a) = 0$, vrijedi

$$\begin{aligned} \sum_{a \in \pi_j} (x - a)^2 &= \sum_{a \in \pi_j} ((x - c_j) + (c_j - a))^2 \\ &= \sum_{a \in \pi_j} (c_j - a)^2 + m_j (c_j - x)^2. \end{aligned}$$

Ako u (19) umjesto x stavimo $c = \frac{1}{m} \sum\limits_{i=1}^{m} a_i$ i zbrojimo sve jednakosti, dobivamo (16). □

Iz Leme 1 neposredno slijedi tvrdnja sljedećeg teorema (Dhillon et al., 2004; Späth, 1983)

Teorem 2. *Uz oznake kao u Lemi 1 vrijedi:*

$$\operatorname*{argmin}_{\Pi \in \mathcal{P}(\mathcal{A};k)} \mathcal{F}(\Pi) = \operatorname*{argmax}_{\Pi \in \mathcal{P}(\mathcal{A};k)} \mathcal{G}(\Pi),$$

$$\operatorname*{argmin}_{c_1, \ldots, c_k \in \mathbb{R}} \mathcal{F}(c_1, \ldots, c_k) = \operatorname*{argmax}_{c_1, \ldots, c_k \in \mathbb{R}} \mathcal{G}(c_1, \ldots, c_k).$$

To znači da u cilju pronalaženja LS-optimalne particije, umjesto minimizacije funkcije $\mathcal{F}$ zadane s (14), odnosno (17), možemo maksimizirati funkciju

$$\mathcal{G}(c_1, \ldots, c_k) = \sum_{j=1}^{k} m_j(c_1, \ldots, c_k)(c_j - c)^2. \tag{20}$$

Primjer 7. *Skup* $A = \{0, 3, 6, 9\}$ *iz Primjera 5, str. 7, ima* 7 *različitih particija i za sve njih u Tablici 3 prikazana je vrijednost kriterijske funkcije cilja* $\mathcal{G}$*. Kao što se vidi iz Tablice 3 funkcija* $\mathcal{G}$ *prima maksimalnu vrijednost* 36 *na optimalnoj particiji* $\{\{0, 3\}, \{6, 9\}\}$*, što je u skladu s Teoremom 1.*

Primjer 8. *Zadan je skup* $\mathcal{A} = \{2, 4, 8, 10, 16\}$*. Treba pronaći sve njegove tročlane particije koje zadovoljavaju Definiciju 1 i koje se nastavljaju jedna na drugu. Za njih treba odrediti pripadne centre i vrijednosti kriterijske funkcije cilja* $\mathcal{F}$ *i* $\mathcal{G}$ *uz primjenu LS-kvazimetričke funkcije cilja.*

Prema Stirlingovoj formul (3), broj svih particija skupa $\mathcal{A}$ s po 3 klastera je 25. Međutim, u slučaju podataka s jednim obilježjem očigledno je da se optimalna particija može očekivati između particija čiji klasteri se nastavljaju jedan na drugi (vidi Sabo et al. (2010), str.161). Broj takvih particija znatno je manji i iznosi

$$\binom{m-1}{k-1}. \tag{21}$$

U ovom slučaju to znači da optimalnu particiju treba tražiti između

$$\binom{5-1}{3-1} = \tfrac{4!}{2!\cdot 2!} = 6$$

particija navedenih u *Tablici 5.*

π_1	π_2	π_3	c_1	c_2	c_3	$\mathcal{F}(\Pi)$	$\mathcal{G}(\Pi)$
{2}	{4}	{8,10,16}	2	4	11.33	0+0+34.67=34.67	36+16+33.33=85.33
{2}	{4,8}	{10,16}	2	6	13	0+8+18=26	36+8+50=94
{2}	{4,8,10}	{16}	2	7.33	16	0+18.67+0=18.67	36+0+64=100
{2,4}	{8}	{10,16}	3	8	13	2+0+18=20	50+0+50=100
{2,4}	{8,10}	{16}	3	9	16	2+2+0=4	50+2+64=116
{2,4,8}	{10}	{16}	4.67	10	16	18.67+0+0=18.67	33.33+4+64=101.33

Tablica 5: LS-particije skupa $\mathcal{A} = \{2, 4, 8, 10, 16\}$

Kao što se vidi iz *Tablice 5* LS-optimalna particija u ovom slučaju je $\{\{2,4\},\{8,10\},\{16\}\}$ jer na njoj funkcija cilja $\mathcal{F}$ zadana s (14) postiže najmanju vrijednost (globalni minimum). Istovremeno kriterijska funkcija $\mathcal{G}$ na toj particiji postiže najveću vrijednost, što je u skladu s Teoremom 2, str.10.

Primjedba 4. Lako se može provjeriti da je veza između centra c čitavog skupa $\mathcal{A}$ i centara c_j pojedinih klastera π_j zadanih s (15) dana s

$$c = \frac{m_1}{m}c_1 + \cdots + \frac{m_k}{m}c_k.$$

Specijalno, za dva disjunktna skupa realnih brojeva $A = \{x_1, \ldots, x_p\}$, $B = \{y_1, \ldots, y_q\}$ aritmetička sredina njihove unije jednaka je ponderiranom zbroju njihovih aritmetičkih sredina, tj. vrijedi

$$\overline{A \cup B} = \frac{p}{p+q}\overline{A} + \frac{q}{p+q}\overline{B}.$$

3.2 Kriterij najmanjih apsolutnih odstupanja

Ako je $d\colon \mathbb{R}\times\mathbb{R} \to \mathbb{R}_+$, $d(x,y) = |x-y|$ govorimo o metričkoj funkciji najmanjih apsolutnih odstupanja (LAD-metrička funkcija)[2]. U tom slučaju centri $c_1, \ldots, c_k$ klastera $\pi_1, \ldots, \pi_k$ određeni su s

$$c_j = \operatorname*{argmin}_{x\in\mathbb{R}} \sum_{a\in\pi_j} |x-a| = \operatorname{med}(\pi_j), \qquad j = 1, \ldots, k, \tag{22}$$

a funkcija cilja (11) s

$$\mathcal{F}(c_1, \ldots, c_k) = \sum_{j=1}^{k} \sum_{a\in\pi_j} |c_j - a|. \tag{23}$$

Ako pri tome iskoristimo (24), onda za izračunavanje funkcije cilja (23) nije potrebno poznavati centre klastera (22), što može značajno ubrzati računski proces.

Primjer 9. *Zadan je skup $\mathcal{A} = \{2,4,8,10,16\}$ kao u Primjeru 8, str.11. Treba pronaći sve njegove tročlane particije koje zadovoljavaju Definiciju 1 i koje se nastavljaju jedna na drugu. Za njih treba odrediti pripadne centre i vrijednosti funkcije cilja $\mathcal{F}$ u smislu najmanjih apsolutnih odstupanja (l_1 udaljenosti) te pronaći globalno optimalnu particiju.*

[2]En.: Least Absolute Deviations = najmanja apsolutna odstupanja

π_1	π_2	π_3	c_1	c_2	c_3	$\mathcal{F}(\Pi)$
{2}	{4}	{8,10,16}	2	4	10	0+0+8=8
{2}	{4,8}	{10,16}	2	5	12	0+4+6=10
{2}	{4,8,10}	{16}	2	8	16	0+6+0=6
{2,4}	{8}	{10,16}	3	8	13	2+0+6=8
{2,4}	{8,10}	{16}	3	8	13	2+2+0=4
{2,4,8}	{10}	{16}	4	10	16	6+0+0=6

Tablica 6: Particije skupa $\mathcal{A} = \{2, 4, 8, 10, 16\}$

Broj svih tročlanih particija koje se nastavljaju je $\binom{m-1}{k-1} = 6$, a kao što se vidi iz *Tablice 6* LAD-optimalna particija u ovom slučaju je $\{\{2, 4\}, \{8, 10\}, \{16\}\}$ jer na njoj funkcija cilja $\mathcal{F}$ zadana s (23) postiže najmanju vrijednost (globalni minimum).

Zadatak 3. *Neka je $\mathcal{A} = \{a_1, \ldots, a_m\}$ konačan niz realnih brojeva. Pokažite da vrijedi*

$$\sum_{i=1}^{m} |a_i - \operatorname{med}(A)| = \sum_{i=1}^{k} (a_{m-i+1} - a_i)\,. \tag{24}$$

3.3 Grupiranje podataka s težinama

Pretpostavimo da je zadan skup podataka $\mathcal{A} = \{a_1, \ldots, a_m\}$, pri čemu je svakom podatku a_i pridužena odgovarajuća težina $w_i > 0$. Kriterijska funkcija cilja (11) sada postaje

$$\mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{a_i \in \pi_j} w_i d(c_j, a_i). \tag{25}$$

Specijalno, kod primjene kriterija LS-optimalnosti centar c_j klastera π_j određen je težinskom aritmetičkom sredinom podataka iz klastera π_j

$$c_j = \frac{1}{\kappa_j} \sum_{a_i \in \pi_j} w_i a_i, \qquad \kappa_j = \sum_{a_i \in \pi_j} w_i, \tag{26}$$

a kod primjene kriterija LAD-optimalnosti centar c_j klastera π_j određen je težinskim medijanom podataka koji pripadaju klasteru π_j (Sabo and Scitovski, 2008; **?**)

$$c_j = \operatorname*{med}_{a_i \in \pi_j}(w_i, a_i). \tag{27}$$

3.4 Formulacija problema grupiranja preko centroida

Za dani skup centara $c_1, \ldots, c_k \in \mathbb{R}$, uz primjenu *principa minimalnih udaljenosti* možemo definirati particiju $\Pi = \{\pi_1, \ldots, \pi_k\}$ skupa $\mathcal{A}$ na sljedeći način:

$$\pi_j = \{a \in \mathcal{A} : d(c_j, a) \leq d(c_s, a),\ \forall s = 1, \ldots, k\}, \qquad j = 1, \ldots, k, \tag{28}$$

pri čemu treba voditi računa o tome da svaki element skupa $\mathcal{A}$ pripadne samo jednom klasteru. Zato se problem traženja optimalne particije skupa $\mathcal{A}$ može svesti na sljedeći optimizacijski problem

$$\min_{c_1,\ldots,c_k \in \mathbb{R}} F(c_1, \ldots, c_k), \qquad F(c_1, \ldots, c_k) = \sum_{i=1}^{m} \min_{j=1,\ldots,k} d(c_j, a_i), \tag{29}$$

gdje je $F\colon \mathbb{R}^k \to \mathbb{R}_+$. Općenito, ova funkcija nije konveksna ni diferencijabilna, a može imati više lokalnih minimuma (Gan et al., 2007; Iyigun and Ben-Israel, 2010; Teboulle, 2007).

Optimizacijski problem (29) u literaturi se može naći pod nazivom *k-median problem* i ekvivalentan je optimizacijskom problemu (12). Naime, vrijedi

$$\begin{aligned} F(c_1, \ldots, c_k) :=& \sum_{i=1}^{m} \min\{d(c_1, a_i), \ldots, d(c_k, a_i)\} \\ =& \sum_{j=1}^{k} \sum_{a_i \in \pi_j} \min\{d(c_1, a_i), \ldots, d(c_k, a_i)\} \\ =& \sum_{j=1}^{k} \sum_{a_i \in \pi_j} d(c_j, a_i) =: \mathcal{F}(\Pi). \end{aligned}$$

4 Traženje lokalno optimalne particije podataka s jednim obilježjem

Problem traženja globalno optimalnog rješenja je složeni problem nediferencijabilne i nekonveksne optimizacije. Zbog toga ćemo se zadovoljiti traženjem lokalno optimalne particije. Najpoznatiji algoritam za traženje lokalno optimalne particije je k-means algoritam. To je iterativni proces koji na osnovi početne aproksimacije (početnih centara ili početne particije) daje lokalno optimalnu particiju. Algoritam daje rješenje u konačno koraka, a u svakom koraku snižava vrijednost funkcije cilja. Takvo rješenje ne mora biti globalno optimalno, što znači da se može dogoditi da ima boljih rješenja od dobivenog. Još jedan nedostatak ove metode je mogućnost da se tijekom iterativnog procesa može dogoditi da neki od klastera postane prazan skup.

Ako k-means algoritam pokrenemo više puta s različitom početnom aproksimacijom i izaberemo najbolje od dobivenih rješenja (Leisch, 2006), možemo se nadati da ćemo se približiti globalno optimalnom rješenju.

4.1 k-means algoritam za dva klastera

Potražit ćemo lokalno optimalnu particiju skupa $\mathcal{A}$ s m elemena koja se sastoji od 2 klastera. k-means algoritam pokrenut ćemo zadavanjem početne particije $\Pi = \{\pi_1, \pi_2\}$.

Algoritam 1. `(Izbor početne particije)`

`Korak 1:` Inicijalizacija: učitati m i elemente skupa $\mathcal{A}$;
Izabrati početnu particiju: $\Pi = \{\pi_1, \pi_2\}$;

`Korak 2:` Priduživanje (assignment step)

$$c_1 = \operatorname*{argmin}_{x\in\mathbb{R}} \sum_{a\in\pi_1} d(x,a), \qquad c_2 = \operatorname*{argmin}_{x\in\mathbb{R}} \sum_{a\in\pi_2} d(x,a),$$
$$\mathcal{F}(\Pi) = \sum_{a\in\pi_1} d(c_1,a) + \sum_{a\in\pi_2} d(c_2,a);$$

`Korak 3:` Korekcija (update step)

$$\nu_1 = \{a \in \mathcal{A} : d(c_1,a) \le d(c_2,a)\},$$
$$\nu_2 = \{a \in \mathcal{A} : d(c_2,a) < d(c_1,a)\};$$

`Korak 4:` Ispitivanje optimalnosti: Ako je $\nu_1 \neq \pi_1$ ili $\nu_2 \neq \pi_2$, staviti

$$\pi_1 = \nu_1 \quad \text{i} \quad \pi_2 = \nu_2,$$

i prijeći na `Korak 2`; Inače `STOP`.

Primjedba 5. Iterativni proces traje tako dugo dok se particije ne počnu ponavljati (u tom slučaju i vrijednost funkcije cilja prestane opadati). Primijetite da se u `Korak 3` novi klasteri ν_1, ν_2 geometrijski mogu odrediti tako da odredimo simetralu spojnice centroida. Tada svi elementi lijevo od simetrale pripadaju novom klasteru ν_1, a svi elementi desno od simetrale pripadaju novom klasteru ν_2. AKo se neki element pojavi baš na simetrali, sukladno `Koraku 3` svrstat ćemo ga u lijevi klaster

Primjer 10. *Treba pronaći LS-optimalnu dvočlanu particiju skupa* $\mathcal{A} = \{1,2,6,7,9\}$ *primjenom k-means Algoritma 1 uz početnu particiju* $\Pi = \{\{1,2,6\},\{7,9\}\}$. *Postupak je vidljiv u* Tablici 7.

Iteracija	π_1	π_2	c_1	c_2	Funkcija cilja
1	{1,2,6}	{7,9}	3	8	16
2	{1,2}	{6,7,9}	$\frac{3}{2}$	$\frac{22}{3}$	$\frac{31}{6} \approx 5.167$
3	{1,2}	{6,7,9}	$\frac{3}{2}$	$\frac{22}{3}$	$\frac{31}{6} \approx 5.167$

Tablica 7: Traženje LS-optimalne dvočlane particije skupa $\mathcal{A} = \{1, 2, 6, 7, 9\}$

Sljedeći primjer pokazuje da standardni k-means algoritam ne daje uvijek optimalno rješenje.

Primjer 11. *Treba pronaći LS-optimalnu dvočlanu particiju skupa* $\mathcal{A} = \{0, 2, 3\}$, $k = 2$ *primjenom k-means Algoritma 1 uz početnu particiju* $\Pi = \{\{0, 2\}, \{3\}\}$.

Kao što se vidi u *Tablici 8* k-means Algoritma 1 ne može pronaći bolju particiju od početne. Međutim, bolja particija u ovom slučaju je particija $\Pi^\star = \{\{0\}, \{2, 3\}\}$ jer je $F(\Pi^\star) = 0.5$.

Iteracija	π_1	π_2	c_1	c_2	Funkcija cilja
1	{0,2}	{3}	1	3	2
2	{0,2}	{3}	1	3	2

Tablica 8: Traženje LS-optimalne dvočlane particije skupa $\mathcal{A} = \{0, 2, 3\}$

k-means algoritam također se može pokrenuti i izborom početnih centara. U tom smislu niže je navedena varijacija Algoritma 1.

Algoritam 2. `(Izbor početnih centara)`

`Korak 1:` Inicijalizacija: učitati m i elemente skupa $\mathcal{A}$;
Izabrati početne centre: $\zeta_1 < \zeta_2$;

`Korak 2:` Korekcija (update step)

$$\pi_1 = \{a \in \mathcal{A} : d(\zeta_1, a) \leq d(\zeta_2, a)\},$$
$$\pi_2 = \{a \in \mathcal{A} : d(\zeta_2, a) < d(\zeta_1, a)\};$$

`Korak 3:` Priduživanje (assignment step)

$$c_1 = \underset{x \in \mathbb{R}}{\operatorname{argmin}} \sum_{a \in \pi_1} d(x, a), \qquad c_2 = \underset{x \in \mathbb{R}}{\operatorname{argmin}} \sum_{a \in \pi_2} d(x, a),$$

$$\mathcal{F}(\Pi) = \sum_{a \in \pi_1} d(c_1, a) + \sum_{a \in \pi_2} d(c_2, a);$$

`Korak 4:` Ispitivanje optimalnosti: Ako je $c_1 \neq \zeta_1$ ili $c_2 \neq \zeta_2$, staviti

$$\zeta_1 = c_1 \quad \text{i} \quad \zeta_2 = c_2,$$

i prijeći na `Korak 2`; Inače `STOP`.

Primjer 12. *Treba pronaći LS-optimalnu dvočlanu particiju skupa* $\mathcal{A} = \{1, 2, 6, 7, 9\}$ *primjenom k-means Algoritma 2 uz početne centre* $\zeta_1 = 4$, $\zeta_2 = 8$.

Primjenom principa minimalnih udaljenosti na osnovi početnih centara $\zeta_1 = 4$, $\zeta_2 = 8$ (`Korak 2`) dobivamo početnu particiju $\Pi = \{\{1, 2, 6\}, \{7, 9\}\}$. Daljnji tijek iterativnog procesa podudara se s Algoritmom 1 i prikazan je u *Tablici 9.*

Iteracija	π_1	π_2	c_1	c_2	$\mathcal{F}$
1	{1,2,6}	{7,9}	3	8	16
2	{1,2}	{6,7,9}	1.5	7.33	5.17
3	{1,2}	{6,7,9}	1.5	7.33	5.17

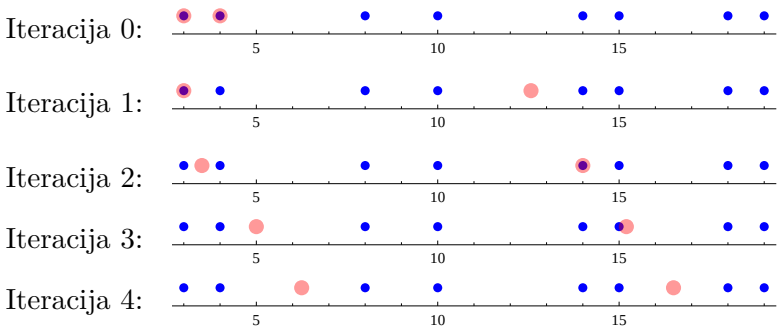
Tablica 9: Traženje LS-optimalne dvočlane particije skupa $\mathcal{A} = \{1, 2, 6, 7, 9\}$

Primjer 13. *Treba pronaći LS-optimalnu dvočlanu particiju skupa* $\mathcal{A} = \{3, 4, 8, 10, 14, 15, 18, 19\}$ *primjenom k-means Algoritma 2 uz početne centre* $\zeta_1 = 3$, $\zeta_2 = 4$.

Primjenom principa minimalnih udaljenosti na osnovi početnih centara $\zeta_1 = 3$, $\zeta_2 = 4$ (`Korak 2`) dobivamo početnu particiju $\Pi = \{\{3\}, \{4, 8, 10, 14, 15, 18, 19\}\}$. Daljnji tijek iterativnog procesa podudara se s Algoritmom 1 i može se pratiti u *Tablici 10* ili na Slici 1.

Iteracija	π_1	π_2	c_1	c_2	$\mathcal{F}$
1	{3}	{4, 8, 10, 14, 15, 18, 19}	3	12.57	179.7
2	{3, 4}	{8,10, 14, 15, 18, 19}	3.5	14	94.5
3	{3, 4, 8}	{10,14, 15, 18, 19}	5	15.2	64.8
4	{3, 4, 8, 10}	{14, 15, 18, 19}	6.25	16.5	49.75
5	{3, 4, 8, 10}	{14, 15, 18, 19}	6.25	16.5	49.75

Tablica 10: Traženje LS-optimalne dvočlane particije skupa $\mathcal{A} = \{3, 4, 8, 10, 14, 15, 18, 19\}$



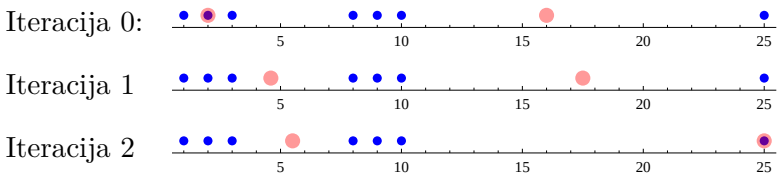
Slika 1: k-means iterativni proces

Primjer 14. *Treba pronaći LS-optimalnu dvočlanu particiju skupa* $\mathcal{A} = \{1, 2, 3, 8, 9, 10, 25\}$ *primjenom k-means Algoritma 2 uz početne centre* $\zeta_1 = 2$, $\zeta_2 = 16$. *Iterativni proces može se pratiti u tablici ili na Slici 2.*

Primjenom principa minimalnih udaljenosti na osnovi početnih centara $\zeta_1 = 2$, $\zeta_2 = 16$ (`Korak 2`) dobivamo početnu particiju $\Pi = \{\{1, 2, 3, 8, 9\}, \{10, 25\}\}$. Daljnji tijek iterativnog procesa podudara se s Algoritmom 1 i može se pratiti u *Tablici 11* ili na Slici 2.

Iteracija	π_1	π_2	c_1	c_2	$\mathcal{F}$
1	{1, 2, 3, 8, 9}	{10, 25}	4.6	17.5	165.7
2	{1, 2, 3, 8, 9,10}	{25}	5.5	25	77.5
3	{1, 2, 3, 8, 9,10}	{25}	5.5	25	77.5

Tablica 11: Traženje LS-optimalne dvočlane particije skupa $\mathcal{A} = \{1, 2, 3, 8, 9, 10, 25\}$



Slika 2: k-means iterativni proces

Broj svih dvočlanih particija ovog skupa je $2^{7-1} - 1 = 63$, ali postoji samo 6 particija čiji klasteri se nastavljaju. Odmah uočavamo da skup $\mathcal{A}$ sadrži dvije značajno različite skupine realnih brojeva $\mathcal{A}_1 = \{1, 2, 3\}$ te $\mathcal{A}_2 = \{8, 9, 10\}$. Također, skup $\mathcal{A}$ sadrži i element 25, kojeg možemo shvatiti kao jako stršeći podatak nastao zbog određene pogreške, a prirodno dolazi iz skupine $\mathcal{A}_2$. Primjenom Algoritma 2 uz početne centre $\zeta_1 = 2$ i $\zeta_2 = 16$, dobivamo početnu particiju $\Pi = \{\pi_1, \pi_2\}$, $\pi_1 = \{1, 2, 3, 8, 9\}$, $\pi_2 = \{10, 25\}$. Direktnom provjerom svih particija može se pokazati da je Algoritam 2 pronašao upravo optimalnu particiju. Iz ovog primjera vidljivo je da k-means algoritam u smislu LS-optimalnosti daje particiju, koja značajno ovisi o stršećem podatku, tako da upravo stršeći podatak čini zaseban klaster (vidi Sliku 2).

Primjedba 6. U slučaju izbora LAD-kriterija optimalnosti u Koraku 2 Algoritma 2 može se dogoditi da neki centroid c_j, može biti proizvoljan broj iz nekog intervala $[\alpha, \beta] \subset \mathbb{R}$. U tom slučaju najbolje je uzeti $c_j = \frac{\alpha+\beta}{2}$.

Primjer 15. *Treba pronaći LAD-optimalnu dvočlanu particiju skupa* $\mathcal{A} = \{1, 2, 6, 7, 9\}$ *primjenom k-means Algoritma 1 uz početnu particiju* $\Pi = \{\{1, 2, 6\}, \{7, 9\}\}$.

Iteracija	π_1	π_2	c_1	c_2	Funkcija cilja
1	{1,2,6}	{7,9}	2	8	5+2=7
2	{1,2}	{6,7,9}	1.5	7	1+3=4
3	{1,2}	{6,7,9}	1.5	7	1+3=4

Tablica 12: Traženje LS-optimalne dvočlane particije skupa $\mathcal{A} = \{1, 2, 6, 7, 9\}$

Primjer 16. *Treba pronaći LAD-optimalnu dvočlanu particiju skupa* $\mathcal{A} = \{1, 2, 3, 8, 9, 10, 25\}$ *primjenom k-means Algoritma 2 uz početne centre* $\zeta_1 = 2$, $\zeta_2 = 15$.

Primjenom principa minimalnih udaljenosti na osnovi početnih centara $\zeta_1 = 2$, $\zeta_2 = 15$ (`Korak 2`) dobivamo početnu particiju $\Pi = \{\{1, 2, 3, 8\}, \{9, 10, 25\}\}$. Daljnji tijek iterativnog procesa podudara se s Algoritmom 1 i može se pratiti u *Tablici 13*.

Iteracija	π_1	π_2	c_1	c_2	Funkcija cilja
1	{1,2,3,8}	{9,10,25}	2.50	10.00	24
2.	{1,2,3}	{8,9,10,25}	2.00	9.50	20
3.	{1,2,3}	{8,9,10,25}	2.00	9.50	20

Tablica 13: Traženje LAD-optimalne dvočlane particije skupa $\mathcal{A} = \{1, 2, 3, 8, 9, 10, 25\}$

Direktnom provjerom može se pokazati da je algoritam pronašao upravo LAD-optimalnu particiju.

4.2 k-means algoritam za k klastera

Potražit ćemo lokalno optimalnu particiju skupa $\mathcal{A}$ s m elemenata koja se sastoji od $1 \leq k \leq m$ klastera. k-means algoritam pokrenut ćemo zadavanjem početne particije $\Pi = \{\pi_1, \ldots, \pi_k\}$.

Algoritam 3. `(Izbor početne particije)`

`Korak 1:` Inicijalizacija: učitati m, k i elemente skupa $\mathcal{A}$;
Izabrati početnu particiju: $\Pi = \{\pi_1, \ldots, \pi_k\}$;

`Korak 2:` Priduživanje (assignment step)

$$c_j = \operatorname*{argmin}_{x \in \mathbb{R}} \sum_{a \in \pi_j} d(x, a), \quad j = 1, \ldots, k,$$

$$\mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{a \in \pi_j} d(c_j, a);$$

`Korak 3:` Korekcija (update step)

$$\nu_j = \{a \in \mathcal{A} : d(c_j, a) \leq d(c_s, a), \forall s = 1, \ldots, k\}, \quad j = 1, \ldots, k$$

`Korak 4:` Ispitivanje optimalnosti: Ako je $\nu_j = \pi_j$ za svaki $j = 1, \ldots, k$, `STOP`; U protivnom staviti

$$\pi_j = \nu_j \quad j = 1, \ldots, k,$$

i prijeći na `Korak 2`.

Primjedba 7. Iterativni proces traje tako dugo dok se particije ne počnu ponavljati (u tom slučaju i vrijednost funkcije cilja prestane opadati). Primijetite da se u `Korak 3` novi klasteri ν_j geometrijski mogu odrediti tako da odredimo simetrale spojnica centroida.

Primjer 17. *Treba pronaći LS-optimalnu tročlanu particiju skupa* $\mathcal{A} = \{0, 2, 4, 8, 9, 10, 12, 16\}$ *primjenom k-means Algoritma 3 uz početnu particiju* $\Pi = \{\{0, 2, 4\}, \{8, 9\}, \{10, 12, 16\}\}$.

Iteracija	π_1	π_2	π_3	c_1	c_2	c_3	Funkcija cilja
1	{0,2,4}	{8,9}	{10, 12, 16}	2	8.5	12.67	27.17
2	{0,2,4}	{8,9,10}	{12, 16}	2	9	14	18
3	{0,2,4}	{8,9,10}	{12, 16}	2	9	14	18

Tablica 14: Traženje LS-optimalne tročlane particije skupa $\mathcal{A} = \{0, 2, 4, 8, 9, 10, 12, 16\}$

Primjer 18. *Treba pronaći LS-optimalnu tročlanu particiju skupa* $\mathcal{A} = \{1, 2, 4, 8, 9, 10, 12, 15, 18, 20\}$ *primjenom k-means Algoritma 3 uz početnu particiju* $\Pi = \{\{1, 2, 4\}, \{8, 9\}, \{10, 12, 15, 18, 20\}\}$.

Iteracija	π_1	π_2	π_3	c_1	c_2	c_3	Funkcija cilja
1	{1,2,4}	{8,9}	{10, 12, 15, 18, 20}	2.33	8.5	15	73.17
2	{1,2,4}	{8,9,10}	{12, 15, 18, 20}	2.33	9	16.25	43.42
3	{1,2,4}	{8,9,10,12}	{15, 18, 20}	2.33	9.75	17.67	26.08
4	{1,2,4}	{8,9,10,12}	{15, 18, 20}	2.33	9.75	17.67	26.08

Tablica 15: Traženje LS-optimalne tročlane particije skupa $\{1, 2, 4, 8, 9, 10, 12, 15, 18, 20\}$

Zadatak 4. *Odredite LS-optimalnu vrijednost dualne funkcije cilja* $\mathcal{G}$ *tročlane particije skupova iz Zadatka* **??** *i Zadatka* **??**.

Zadatak 5. *Pronađite LAD-optimalnu tročlanu particiju skupova iz Zadatka* **??** *i Zadatka* **??**. *Kolike su optimalne vrijednosti funkcije cilja?*

Algoritam se također može pokrenuti izborom početnih centara. U tom smislu niže je navedena varijacija Algoritma 3

Algoritam 4. `(Izbor početnih centara)`

`Korak 1:` Inicijalizacija: učitati m, k i elemente skupa $\mathcal{A}$;
Izabrati početne centre: $\zeta_1 < \cdots < \zeta_k$;

`Korak 2:` Korekcija (update step)

$$\pi_j = \{a \in \mathcal{A} : d(\zeta_j, a) \leq d(\zeta_s, a),\ \forall s = 1, \ldots, k\}, \quad j = 1, \ldots, k$$

`Korak 3:` Pridruživanje (assignment step)

$$c_j = \underset{x\in\mathbb{R}}{\operatorname{argmin}} \sum_{a\in\pi_j} d(x,a), \quad j = 1,\ldots,k,$$

$$\mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{a\in\pi_j} d(c_j,a);$$

`Korak 4:` Ispitivanje optimalnosti: ako je $c_j = \zeta_j$ za svaki $j = 1,\ldots,k$, `STOP`;
U protivnom staviti

$$\zeta_j = c_j, \quad j = 1,\ldots,k,$$

i prijeći na `Korak 2`.

Primjer 19. *Zadan je skup* $\mathcal{A} = \{2,4,8,10,16\}$ *iz* Primjera 9*, str.12. Za početne centre* $\zeta_1 = 3$*,* $\zeta_2 = 8$*,* $\zeta_3 = 10$ *primjenom k-means algoritma treba pronaći lokalno optimalnu LS i LAD particiju. Također treba odrediti LS-optimalnu vrijednost dualne funkcije cilja* $\mathcal{G}$*.*

Primjenom principa minimalnih udaljenosti na osnovi početnih centara $\zeta_1 = 3$, $\zeta_2 = 8$, $\zeta_3 = 10$ (`Korak 2`) dobivamo početnu particiju $\Pi = \{\{2,4\},\{8\},\{10,16\}\}$. Daljnji tijek iterativnog procesa podudara se s Algoritmom 3 i može se pratiti u *Tablici 16.*

Iteracija	Klasteri			Centri			Funkcija cilja	Funkcija cilja
	π_1	π_2	π_3	c_1	c_2	c_3	$\mathcal{F}$	$\mathcal{G}$
1	{2,4}	{8}	{10,16}	3	8	13	20	6+9+128=143
2	{2,4}	{8,10}	{16}	3	9	16	4	8+32+121=161
3	{2,4}	{8,10}	{16}	3	9	16	4	8+32+121=161

Tablica 16: Traženje LS-optimalne tročlane particije skupa $\mathcal{A} = \{2,4,8,10,16\}$

Primjer 20. *Zadan je skup* $\mathcal{A} = \{2,3,5,7,9,12,13,15\}$*. Za početne centre* $\zeta_1 = 4$*,* $\zeta_2 = 10$*,* $\zeta_3 = 14$ *primjenom k-means algoritma treba pronaći lokalno optimalnu LAD particiju.*

Primjenom principa minimalnih udaljenosti na osnovi početnih centara $\zeta_1 = 4$, $\zeta_2 = 10$, $\zeta_3 = 14$ (`Korak 2`) dobivamo početnu particiju $\Pi = \{\{2,3,5,7\},\{9,12\},\{13,15\}\}$. Daljnji tijek iterativnog procesa podudara se s Algoritmom 3 i može se pratiti u *Tablici 17.*

Iteracija	Klasteri			Centri			Funkcija cilja
	π_1	π_2	π_3	c_1	c_2	c_3	$\mathcal{F}$
1	{2,3,5,7}	{9,12}	{13,15}	4	10.5	14	7+3+2=12
2	{2,3,5}	{7,9}	{12,13,15}	3	8	13	3+2+3=8
3	{2,3,5}	{7,9}	{12,13,15}	3	8	13	3+2+3=8

Tablica 17: Traženje LAD-optimalne tročlane particije skupa $\mathcal{A} = \{2, 3, 5, 7, 9, 12, 13, 15\}$

Primjer 21. *Zadan je skup* $\mathcal{A} = \{2, 4, 8, 10, 16\}$ *iz Primjera 9, str.12, s odgovarajućim težinama* $w_i = \{2, 1, 4, 2, 2\}$. *Za početne centre* $\zeta_1 = 3$, $\zeta_2 = 8$, $\zeta_3 = 10$ *primjenom k-means algoritma treba pronaci lokalno optimalnu LS i LAD particiju. Odredite također i LS-optimalnu vrijednost dualne funkcije cilja* $\mathcal{G}$.

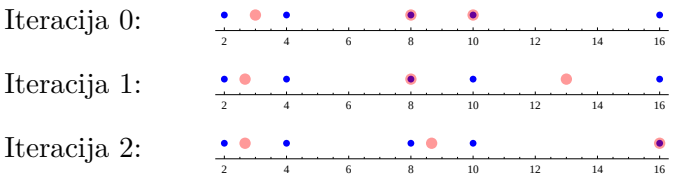
Primjenom principa minimalnih udaljenosti na osnovi početnih centara $\zeta_1 = 3$, $\zeta_2 = 8$, $\zeta_3 = 10$ (`Korak 2`) dobivamo početnu particiju. Daljnji tijek iterativnog procesa podudara se s Algoritmom 3 i može se pratiti u *Tablici 18* i na *Slici 3.*

Iteracija	π_1	π_2	π_3	c_1	c_2	c_3	$\mathcal{F}$	Φ
1	$\{2, 4\}$	$\{8\}$	$\{10, 16\}$	8/3	8	13	38.67	75
2	$\{2, 4\}$	$\{8, 10\}$	$\{16\}$	8/3	26/3	16	8	28.67
3	$\{2, 4\}$	$\{8, 10\}$	$\{16\}$	8/3	26/3	16	8	8

Tablica 18: Traženje LAD-optimalne tročlane particije skupa s težinama

Prosječno težinsko kvadratno rasipanje po klasterima (varijanca): $\{.89, .89, 0\}$

Prosječno težinsko rasipanje po klasterima (standardna devijacija): $\{0.94, 0.94, 0.\}$



Slika 3: k-means iterativni proces

5 Grupiranje u dva klastera na osnovi dva obilježja

$\mathcal{A} = \{a^1, \ldots, a^m\} \subset \mathbb{R}^2$ – skup točaka u ravnini.
$\Pi(\mathcal{A}) = \{\pi_1, \pi_2\}$ – particija skupa $\mathcal{A}$, takva da vrijedi

$$\pi_1 \cup \pi_2 = \mathcal{A}, \qquad \pi_1 \cap \pi_2 = \emptyset, \qquad m_1 = |\pi_1| \geq 1, \quad m_2 = |\pi_2| \geq 1.$$

$|\mathcal{P}(\mathcal{A};k)| = 2^{m-1}-1$ – broj svih ovakvih particija

Akup $\mathcal{A}$ geupirat ćemo u dva klastera tako da budu što bolje razdvojeni i i što više kompaktni. Zbog toga uvedimo neku kvazimetričku funkciju $d\colon \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}_+$, definirajmo centre klastera

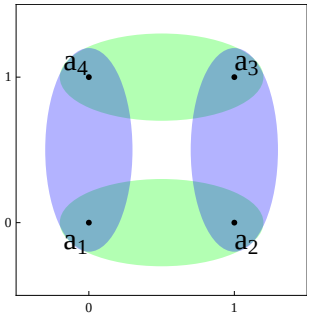
$$c_1 = \operatorname*{argmin}_{x\in\mathbb{R}^2} \sum_{a\in\pi_1} d(x,a), \qquad c_2 = \operatorname*{argmin}_{x\in\mathbb{R}^2} \sum_{a\in\pi_2} d(x,a),$$

i uvedimo sljedeću kriterijsku funkciju cilja *(objective function)*

$$\mathcal{F}(\Pi) = \sum_{a\in\pi_1} d(c_1,a) + \sum_{a\in\pi_2} d(c_2,a),$$

Geometrijski, funkcija $\mathcal{F}$ predstavlja ukupno "rasipanje": zbroj suma udaljenosti elemenata svakog klastera do njegovog centra. Jasno je da što je vrijednost kriterijske funkcije $\mathcal{F}$ manja time je "rasipanje" manje, a time su i predthodno navedeni kriteriji kompaktnosti i razdvojenosti bolje ispunjeni.

Primjer 22. *Skup* $\mathcal{A} = \{a^1 = (0,0),\, a^2 = (1,0),\, a^3 = (1,1),\, a^4 = (0,1)\}$ *prikazan je Slici 4. Broj svih dvočlanih particija ovog skupa je* $\mathcal{P}(\mathcal{A};2) = 2^{4-1}-1 = 7$.



Slika 4: LS-optimalne particije

5.1 Princip najmanjih kvadrata (LS)

$a = (x_1, x_2)$, $b = (y_1, y_2)$
$d(a,b) = \|a-b\|_2^2 = (x_1-y_1)^2 + (x_2-y_2)^2$ - LS-kvazimetrička funkcija

$$c_1 = \tfrac{1}{|\pi_1|}\sum_{a\in\pi_1} a, \qquad c_2 = \tfrac{1}{|\pi_2|}\sum_{a\in\pi_2} a, \qquad \mathcal{F}(\Pi) = \sum_{a\in\pi_1}\|c_1-a\|^2 + \sum_{a\in\pi_2}\|c_2-a\|^2.$$

U ovom slučaju vrijednost kriterijske funkcije $\mathcal{F}$ predstavlja sumu "kvadrata udaljenosti" točaka klastera π_1 do njegovog centroida c_1 i točaka klastera π_2 do njegovog centroida c_2.

Primjer 23. *Izmedju svih particija skupa* $\mathcal{A}$ *iz* Primjera 22 *potražimo optimalnu. Kao što se vidi iz Tablice 19, dvije particije:* $\{\{a^1,a^2\},\{a^3,a^4\}\}$ *i* $\{\{a^1,a^4\},\{a^2,a^3\}\}$ *su optimalne jer na njima kriterijska funkcija cilja* $\mathcal{F}$ *postiže globalni minimum (vidi Sliku 4).*

π_1	π_2	$\mathbf{c}_1$	$\mathbf{c}_2$	$F(\Pi)$		$\mathcal{G}(\Pi)$
$\{a^1\}$	$\{a^2, a^3, a^4\}$	a^1	$\left(\frac{2}{3}, \frac{2}{3}\right)^T$	$0 + \frac{4}{3}$	≈ 1.3	$\frac{1}{2} + \frac{1}{6} \approx 0.6$
$\{a^2\}$	$\{a^1, a^3, a^4\}$	a^2	$\left(\frac{1}{3}, \frac{2}{3}\right)^T$	$0 + \frac{4}{3}$	≈ 1.3	$\frac{1}{2} + \frac{1}{6} \approx 0.6$
$\{a^3\}$	$\{a^1, a^2, a^4\}$	a^3	$\left(\frac{1}{3}, \frac{1}{3}\right)^T$	$0 + \frac{4}{3}$	≈ 1.3	$\frac{1}{2} + \frac{1}{6} \approx 0.6$
$\{a^4\}$	$\{a^1, a^2, a^3\}$	a^4	$\left(\frac{2}{3}, \frac{1}{3}\right)^T$	$0 + \frac{4}{3}$	≈ 1.3	$\frac{1}{2} + \frac{1}{6} \approx 0.6$
$\{a^1, a^2\}$	$\{a^3, a^4\}$	$\left(\frac{1}{2}, 0\right)^T$	$\left(\frac{1}{2}, 1\right)^T$	$\frac{1}{2} + \frac{1}{2}$	$= 1$	$\frac{1}{2} + \frac{1}{2} = 1$
$\{a^1, a^4\}$	$\{a^2, a^3\}$	$\left(0, \frac{1}{2}\right)^T$	$\left(1, \frac{1}{2}\right)^T$	$\frac{1}{2} + \frac{1}{2}$	$= 1$	$\frac{1}{2} + \frac{1}{2} = 1$
$\{a^1, a^3\}$	$\{a^2, a^4\}$	$\left(\frac{1}{2}, \frac{1}{2}\right)^T$	$\left(\frac{1}{2}, \frac{1}{2}\right)^T$	$1 + 1$	$= 2$	$0 + 0 = 0$

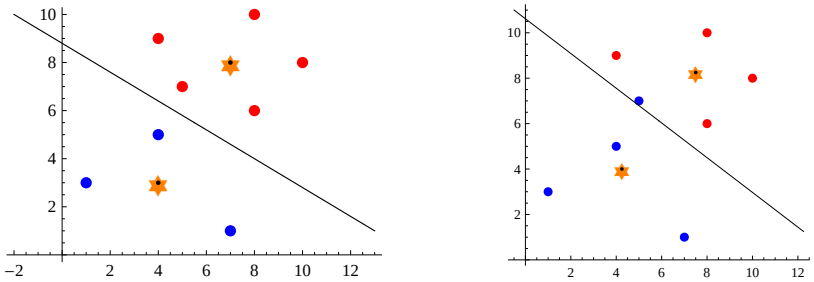
Tablica 19: Particije, centri i funkcije cilja $\mathcal{F}$ i $\mathcal{G}$

Primjer 24. *Skup* $\mathcal{A} = \{a^i = (x_i, y_i) \colon i = 1, \ldots, 8\}$ *zadan je s*

i	1	2	3	4	5	6	7	8
x	1	4	7	5	8	4	8	10
y	3	5	1	7	6	9	10	8

Uz primjenu LS-kvazimetričke funkcije za particije prikazane na Slici 5*, čiji su klasteri označeni plavom, odnosno crvenom bojom treba odrediti centroide i odgovarajuće vrijednosti funkcija cilja* $\mathcal{F}$ *i* $\mathcal{G}$*.*

(a) Particija Π_1 (b) Particija Π_2



Slika 5: Dvije particije skupa $\mathcal{A}$

Za particiju Π_1 dobivamo: $c_1 = (4, 3)$, $c_2 = (7, 8)$, $\mathcal{F} = 26 + 34 = 60$, a za particiju Π_2 dobivamo: $c_1 = (4.25, 4)$, $c_2 = (7.5, 8.25)$, $\mathcal{F} = \frac{155}{4} + \frac{111}{4} = \frac{133}{2}$. Dakle bolja particija je Π_1. Može se pokazati da je to ujedno i globalno optimalna particija. Primijetite da ukupno postoji $2^7 - 1 = 127$ različitih dvočlanih particija.

5.1.1 Dualni LS-problem za podatke s 2 obilježja

Analogno, kao u jednodimenzionalnom slučaju može se pokazati da vrijedi

$$\sum_{i=1}^{m} \|a^i - \mathbf{c}\|_2^2 = \sum_{a\in\pi_1} \|\mathbf{c}_1 - a\|_2^2 + \sum_{a\in\pi_2} \|\mathbf{c}_2 - a\|_2^2 + m_1\|\mathbf{c}_1 - \mathbf{c}\|_2^2 + m_2\|\mathbf{c}_2 - \mathbf{c}\|_2^2,$$

gdje je $\mathbf{c} = \frac{1}{m}\sum_{i=1}^{m} a^i$ centar skupa $\mathcal{A}$. Zato umjesto minimizacije funkcije $\mathcal{F}$ zadane s (34) optimalnu LS-particiju možemo tražiti također i maksimizacijom funkcije

$$\mathcal{G}(\Pi) = m_1\|\mathbf{c}_1 - \mathbf{c}\|_2^2 + m_2\|\mathbf{c}_2 - \mathbf{c}\|_2^2.$$

Određenim prilagođavanjem (Dhillon et al., 2004) problem se svodi na poznate probleme i metode linearne algebre.

Primjer 25. *U* Primjeru 23*, str.24, može se razmatrati i dualni problem. Rezultati su prikazani u Tablici 19, str.25 plavom bojom. Za svaku particiju u tablici je prikazana vrijednost kriterijske funkcije cilja $\mathcal{G}$. Kao što se vidi, funkcija $\mathcal{G}$ prima maksimalnu vrijednost na optimalnim particijama $\{\{a^1, a^2\}, \{a^3, a^4\}\}$ i $\{\{a^1, a^4\}, \{a^2, a^3\}\}$.*

Centroid čitavog skupa $\mathcal{A}$ iz Primjera 24 *je $c = \left(\frac{47}{8}, \frac{49}{8}\right)$. Vrijednost funkcije cilja $\mathcal{G}$ na particiji Π_1 je $\mathcal{G}(\Pi_1) = \frac{255}{4}$, a na particiji Π_2, $\mathcal{G}(\Pi_2) = \frac{229}{4}$, što opet potvrđuje da je Π_1 bolja particija.*

5.2 Princip najmanjih apsolutnih odstupanja (LAD)

$a = (x_1, x_2)$, $b = (y_1, y_2)$
$d(a, b) = \|a - b\|_1 = |x_1 - y_1| + |x_2 - y_2|$ - LAD-kvazimetrička funkcija

$$c_1 = \operatorname*{med}_{a\in\pi_1} a, \qquad c_2 = \operatorname*{med}_{a\in\pi_2} a, \qquad \mathcal{F}(\Pi) = \sum_{a\in\pi_1} \|c_1 - a\| + \sum_{a\in\pi_2} \|c_2 - a\|.$$

U ovom slučaju vrijednost kriterijske funkcije $\mathcal{F}$ predstavlja sumu "rasipanja" točaka klastera π_1 do centroida c_1 i točaka klastera π_2 do centroida c_2.

Zadatak 6. *Na particije iz* Primjera 24 *primijenite LAD-princip.*

6 Grupiranje u k klastera na osnovi dva ili više obilježja

Neka je $\mathcal{A} = \{a^i = (a_1^{(i)}, \ldots, a_n^{(i)}) \in \mathbb{R}^n : i = 1, \ldots, m\}$ skup, koji treba grupirati u $1 \le k \le m$ nepraznih disjunktnih klastera.

Ako je zadana neka kvazimetrička funkcija $d\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$, onda svakom klasteru $\pi_j \in \Pi$ možemo pridružiti njegov centar $\mathbf{c}_j$ na sljedeći način

$$\mathbf{c}_j = c(\pi_j) := \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^n} \sum_{a\in\pi_j} d(\mathbf{x}, a), \quad j = 1, \ldots, k. \tag{30}$$

Na skupu svih particija $\mathcal{P}(\mathcal{A};k)$ skupa $\mathcal{A}$ sastavljenih od k klastera, potpuno analogno kao i ranije, definiramo kriterijsku funkciju cilja $\mathcal{F}:\mathcal{P}(\mathcal{A};k)\to\mathbb{R}_+$,

$$\mathcal{F}(\Pi)=\sum_{j=1}^{k}\sum_{a\in\pi_j} d(\mathbf{c}_j,a), \tag{31}$$

a d-optimalnu particiju $\Pi^\star$ tražimo rješavanjem optimizacijskog problema

$$\mathcal{F}(\Pi^\star)=\min_{\Pi\in\mathcal{P}(\mathcal{A};k)} F(\Pi). \tag{32}$$

Primijetite da na taj način optimalna particija $\Pi^\star$ ima svojstvo da je suma "rasipanja" (suma odstupanja) elemenata klastera oko svog centra minimalna. Na taj način nastojimo postići što bolju unutrašnju kompaktnost i vanjsku separiranost klastera.

6.1 Kriterij najmanjih kvadrata

Neka je $\mathcal{A}=\{a^i=(a_1^i,\dots,a_n^i)\in\mathbb{R}^n\colon i=1,\dots,m\}$. Centri $\mathbf{c}_1,\dots,\mathbf{c}_k$ klastera $\pi_1,\dots,\pi_k$ određeni su s

$$\mathbf{c}_j=\operatorname*{argmin}_{\mathbf{c}\in\mathbb{R}^n}\sum_{a\in\pi_j}\|\mathbf{c}-a\|_2^2=\tfrac{1}{|\pi_j|}\sum_{a\in\pi_j}a=\left(\tfrac{1}{|\pi_j|}\sum_{a\in\pi_j}a_1,\dots,\tfrac{1}{|\pi_j|}\sum_{a\in\pi_j}a_n\right),\quad j=1,\dots,k, \tag{33}$$

pri čemu $\sum\limits_{a\in\pi_j}a_1$ označava sumu prvih komponenti svih elemenata klastera π_j, a $\sum\limits_{a\in\pi_j}a_n$ označava sumu n-tih komponenti svih elemenata klastera π_j. Funkcija cilja (31) u ovom slučaju zadana je s

$$\mathcal{F}(\Pi)=\sum_{j=1}^{k}\sum_{a\in\pi_j}\|\mathbf{c}_j-a\|_2^2 \tag{34}$$

Primjer 26. *Zadan je skup* $\mathcal{A}=\{a^i=(x_i,y_i)\in\mathbb{R}^2: i=1,\dots,m\}$, *gdje je*

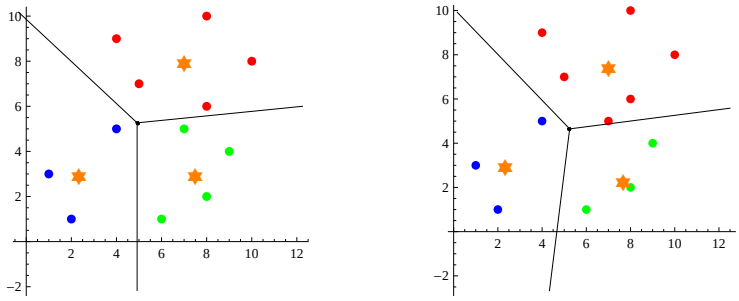
i	1	2	3	4	5	6	7	8	9	10	11	12
x_i	1	4	2	5	8	4	8	10	6	7	8	9
y_i	3	5	1	7	6	9	10	8	1	5	2	4

Promatramo dvoje particije

$$\Pi_1=\{\{a_1,a_2,a_3\},\{a_4,a_5,a_6,a_7,a_8\},\{a_9,a_{10},a_{11},a_{12}\}\quad\dots\quad \textit{Slika 6a}$$
$$\Pi_2=\{\{a_1,a_2,a_3\},\{a_4,a_5,a_6,a_7,a_8,a_{10}\},\{a_9,a_{11},a_{12}\}\}\quad\dots\quad \textit{Slika 6b}$$

Treba ustanoviti na kojoj particije funkcija cilja prima nižu vrijednost.

(a) Particija Π_1 (b) Particija Π_2



Slika 6: Usporedba dviju particija

Rješenje:

$$\Pi_1 : c_1 = (2.33, 3),\ c_2 = (7, 8),\ c_3 = (7.5, 3); \quad \mathcal{F} = \tfrac{38}{3} + 34 + 15 = 61.67;$$
$$\Pi_2 : c_1 = (2.33, 3),\ c_2 = (7, 7.5),\ c_3 = (7.67, 2.33); \quad \mathcal{F} = \tfrac{38}{3} + \tfrac{83}{2} + \tfrac{28}{3} = 63.5.$$

Dakle, niža vrijednost funkcije cilja $\mathcal{F}$ postiže se na particiji Π_2, pa nju po tom kriteriju smatramo optimalnijom.

6.1.1 Dualni LS-problem za podatke s n obilježja

Analogno, kao u jednodimenzionalnom slučaju može se pokazati da vrijedi

$$\sum_{i=1}^{m} \|a^i - \mathbf{c}\|_2^2 = \sum_{j=1}^{k} \sum_{a \in \pi_j} \|\mathbf{c}_j - a\|_2^2 + \sum_{j=1}^{k} m_j \|\mathbf{c}_j - \mathbf{c}\|_2^2, \tag{35}$$

gdje je $\mathbf{c} = \frac{1}{m} \sum_{i=1}^{m} a^i$ centar skupa $\mathcal{A}$, a $m_j = |\pi_j|$. Zato umjesto minimizacije funkcije $\mathcal{F}$ zadane s (34) optimalnu LS-particiju možemo tražiti također i maksimizacijom funkcije

$$\mathcal{G}(\Pi) = \sum_{j=1}^{k} m_j \|\mathbf{c}_j - \mathbf{c}\|_2^2. \tag{36}$$

Određenim prilagođavanjem (Dhillon et al., 2004) problem se svodi na poznate probleme i metode linearne algebre.

Primjer 27. *Centroid čitavog skupa $\mathcal{A}$ iz* Primjera 26 *je $c = (6, \frac{61}{12})$. Vrijednost funkcije cilja $\mathcal{G}$ na particiji Π_1 je $\mathcal{G}(\Pi_1) = 127.25$, a na particiji Π_2, $\mathcal{G}(\Pi_2) = 125.42$, što opet potvrđuje da je Π_1 bolja particija.*

6.2 Kriterij najmanjih apsolutnih odstupanja

Centri $\mathbf{c}_1, \ldots, \mathbf{c}_k$ klastera $\pi_1, \ldots, \pi_k$ određeni su s

$$\mathbf{c}_j = \operatorname*{argmin}_{\mathbf{c} \in \mathbb{R}^n} \sum_{a \in \pi_j} \|\mathbf{c} - a\|_1 = \left(\operatorname*{med}_{a \in \pi_j} a_1, \ldots, \operatorname*{med}_{a \in \pi_j} a_n \right) =: \operatorname{med}(\pi_j), \quad j = 1, \ldots, k, \tag{37}$$

pri čemu $\underset{a\in\pi_j}{\text{med}}\, a_1$ označava medijan prvih komponenti svih elemenata klastera π_j, a $\underset{a\in\pi_j}{\text{med}}\, a_n$ označava medijan n-tih komponenti svih elemenata klastera π_j. Funkcija cilja (31) u ovom slučaju zadana je s

$$\mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{a\in\pi_j} \|\mathbf{c}_j - a\|_1 \tag{38}$$

Primjer 28. *Zadan je skup* $\mathcal{A} = \{a^i = (x_i, y_i) \in \mathbb{R}^2 : i = 1, \ldots, m\}$, *gdje je*

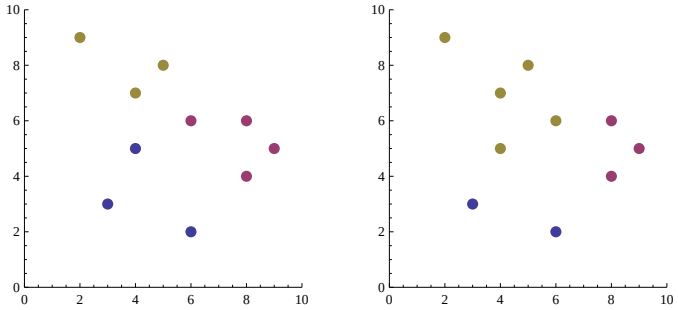
i	1	2	3	4	5	6	7	8	9	10
x_i	2	3	4	4	5	6	6	8	8	9
y_i	9	3	5	7	8	2	6	4	6	5

Promatramo dvije particije

$$\Pi_1 = \{\{a_2, a_3, a_6\}, \{a_7, a_8, a_9, a_{10}\}, \{a_1, a_4, a_5\}\} \quad \ldots \quad \textit{Slika 7a}$$
$$\Pi_2 = \{\{a_2, a_6\}, \{a_8, a_9, a_{10}\}, \{a_1, a_3, a_4, a_5, a_7\}\} \quad \ldots \quad \textit{Slika 7b}$$

Treba ustanoviti na kojoj particije funkcija cilja prima nižu vrijednost.

(a) Particija Π_1 (b) Particija Π_2



Slika 7: Usporedba dviju particija

Niže su izračunati LAD-centri pojedinih klastera u obje particije i vrijednost funkcije cilja na obje particije. Vidi se da je Π_1 "bolja" particija jer se na njoj postiže niža vrijednost funkcije cilja.

	c_1	c_2	c_3	$\mathcal{F}$
Π_1	$(4,3)$	$(8,6)$	$(4,8)$	$(1+2+3)+(2+2+0+2)+(3+1+1) = 17$
Π_2	$(4,2)$	$(8,5)$	$(4,7)$	$()+()+() =$

Primjedba 8. Kao što smo u Odjeljku 3.3 razmatrali problem grupiranja jednodimenzionalnih težinskih podataka, slično bi mogli postupiti i u slučaju grupiranja težinskih dvodimenzionalnih i višedimenzionalnih podataka.

7 Traženje lokalno optimalne particije podataka s više obilježja

7.1 Slučaj dva klastera

Potražit ćemo lokalno optimalnu particiju skupa $\mathcal{A} \subset \mathbb{R}^n$ s m elemena koja se sastoji od 2 klastera. k-means algoritam pokrenut ćemo zadavanjem početne particije $\Pi = \{\pi_1, \pi_2\}$.

Algoritam 5. `(Izbor početne particije)`

`Korak 1:` Inicijalizacija: učitati m i elemente skupa $\mathcal{A}$;
Izabrati početnu particiju: $\Pi = \{\pi_1, \pi_2\}$;

`Korak 2:` Priduživanje (assignment step)

$$c_1 = \operatorname*{argmin}_{x \in \mathbb{R}^n} \sum_{a \in \pi_1} d(x, a), \qquad c_2 = \operatorname*{argmin}_{x \in \mathbb{R}^n} \sum_{a \in \pi_2} d(x, a),$$
$$\mathcal{F}(\Pi) = \sum_{a \in \pi_1} d(c_1, a) + \sum_{a \in \pi_2} d(c_2, a);$$

`Korak 3:` Korekcija (update step)

$$\nu_1 = \{a \in \mathcal{A} : d(c_1, a) \leq d(c_2, a)\},$$
$$\nu_2 = \{a \in \mathcal{A} : d(c_2, a) < d(c_1, a)\};$$

`Korak 4:` Ispitivanje optimalnosti: Ako je $\nu_1 \neq \pi_1$ ili $\nu_2 \neq \pi_2$, staviti

$$\pi_1 = \nu_1 \quad \text{i} \quad \pi_2 = \nu_2,$$

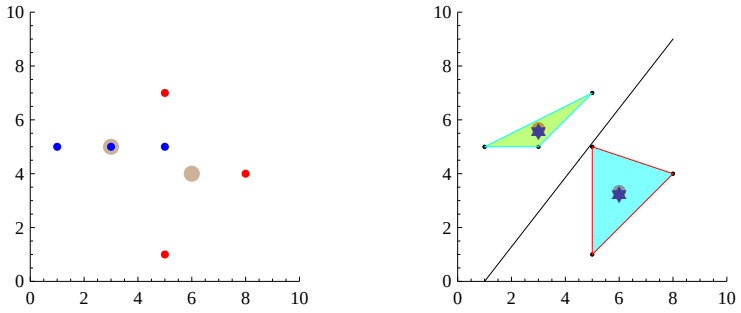
i prijeći na `Korak 2`; Inače `STOP`.

Primjedba 9. Iterativni proces traje tako dugo dok se particije ne počnu ponavljati (u tom slučaju i vrijednost funkcije cilja prestane opadati). Primijetite da se u `Korak 3` novi klasteri ν_1, ν_2 geometrijski mogu odrediti tako da odredimo simetralu spojnice centroida. Tada svi elementi s iste strane simetrale kao i centar c_1 pripadaju novom klasteru ν_1, a svi elementi s druge strane simetrale pripadaju novom klasteru ν_2. Ako se neki element pojavi baš na simetrali, sukladno `Koraku 3` svrstat ćemo ga u prvi klaster

Primjer 29. *Zadan je skup* $\mathcal{A} = \{a^i = (x_i, y_i) \in \mathbb{R}^2 : i = 1, \ldots, 6\}$

i	1	2	3	4	5	6
x_i	1	3	5	5	5	8
y_i	5	5	5	1	7	4

(a) Početna iteracija (b) Rješenje



Slika 8: k-means iterativni proces

Centroid cijelog skupa A je $c = (\frac{9}{2}, \frac{9}{2})$.
Uz početnu particiju $\Pi = \{\pi_1, \pi_2\}$ s klasterima

$$\pi_1 = \{(1,5),(3,5),(5,5)\}, \qquad \pi_2 = \{(5,7),(5,1),(8,4)\},$$

nakon određivanja centroida

$$c_1 = (3,5), \quad c_2 = (6,4),$$

dobivamo vrijednost funkcije cilja $\mathcal{F}_1 = 32$ (i $\mathcal{G}_1 = 15$). (vidi Slika 8a).

Primjenom Koraka 3 dobivamo novu particiju (Slika 8b) $\{\nu_1, \nu_2\}$ s klasterima

$$\nu_1 = \{(1,5),(3,5),(5,7)\}, \qquad \nu_2 = \{(5,5),(5,1),(8,4)\}.$$

Nakon određivanja njihovih centroida

$$c_1 = (3, \tfrac{17}{3}), \quad c_2 = (6, \tfrac{10}{3}),$$

dobivamo novu vrijednost funkcije cilja $\mathcal{F}_2 = \frac{32}{3} + \frac{44}{3} = \frac{76}{3} \approx 25.3$ (i $\mathcal{G}_2 = \frac{65}{3} = 21.67$). Pokazuje se da se klasteri particije i centroidi više neće mijenjati, što znači da smo dobili lokalno optimalnu particiju. Sve možemo prikazati u *Tablici 20*).

It.	π_1	π_2	c_1	c_2	$\mathcal{F}(\Pi)$	$\mathcal{G}(\Pi)$
1	$\{(1,5),(3,5),(5,5)\}$	$\{(5,7),(5,1),(8,4)\}$	(3,5)	(6,4)	32	15
2	$\{(1,5),(3,5),(5,7)\}$	$\{(5,5),(5,1),(8,4)\}$	$(3,\frac{17}{3})$	$(6,\frac{10}{3})$	$\frac{76}{3} \approx 25.3$	$\frac{65}{3} \approx 21.67$
3	$\{(1,5),(3,5),(5,7)\}$	$\{(5,5),(5,1),(8,4)\}$	$(3,\frac{17}{3})$	$(6,\frac{10}{3})$	$\frac{76}{3} \approx 25.3$	$\frac{65}{3} \approx 21.67$

Tablica 20: Traženje optimalne particije skupa $\mathcal{A}$ na osnovi LS-kriterija

Primjer 30. *Skup $\mathcal{A}$ iz* Primjera 29 *grupirat ćemo primjenom LAD-metričke funkcije počevši od iste početne particije. Tijek iterativnog postupka može se pratiti u* Tablici 21

It.	π_1	π_2	c_1	c_2	$\mathcal{F}(\Pi)$
1	$\{(1,5),(3,5),(5,5)\}$	$\{(5,7),(5,1),(8,4)\}$	(3,5)	(5,4)	4+9=13
2	$\{(1,5),(3,5)\}$	$\{(5,5),(5,1),(5,7),(8,4)\}$	$(2,5)$	$(5,4.5)$	$2+10=12$
3	$\{(1,5),(3,5)\}$	$\{(5,5),(5,1),(5,7),(8,4)\}$	$(2,5)$	$(5,4.5)$	$2+10=12$

Tablica 21: Traženje optimalne particije skupa $\mathcal{A}$ na osnovi LAD-kriterija

k-means algoritam također se može pokrenuti i izborom početnih centara. U tom smislu niže je navedena varijacija Algoritma 5.

Algoritam 6. `(Izbor početnih centara)`

`Korak 1:` Inicijalizacija: učitati m i elemente skupa $\mathcal{A}$;
Izabrati početne centre: $\zeta_1 \neq \zeta_2$;

`Korak 2:` Korekcija (update step)

$$\pi_1 = \{a \in \mathcal{A} : d(\zeta_1, a) \leq d(\zeta_2, a)\},$$
$$\pi_2 = \{a \in \mathcal{A} : d(\zeta_2, a) < d(\zeta_1, a)\};$$

`Korak 3:` Priduživanje (assignment step)

$$c_1 = \operatorname*{argmin}_{x \in \mathbb{R}^n} \sum_{a \in \pi_1} d(x, a), \qquad c_2 = \operatorname*{argmin}_{x \in \mathbb{R}^n} \sum_{a \in \pi_2} d(x, a),$$

$$\mathcal{F}(\Pi) = \sum_{a \in \pi_1} d(c_1, a) + \sum_{a \in \pi_2} d(c_2, a);$$

`Korak 4:` Ispitivanje optimalnosti: Ako je $c_1 \neq \zeta_1$ ili $c_2 \neq \zeta_2$, staviti

$$\zeta_1 = c_1 \quad \text{i} \quad \zeta_2 = c_2,$$

i prijeći na `Korak 2`; Inače `STOP`.

Primjer 31. *Zadan je skup* $\mathcal{A} = \{a^i = (x_i, y_i) \in \mathbb{R}^2 : i = 1, \ldots, 8\}$ *koji treba grupirati u dva klastera.*

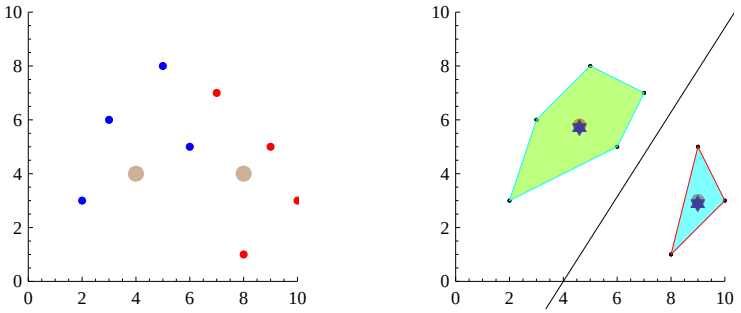
i	1	2	3	4	5	6	7	8
x_i	2	3	5	6	7	8	9	10
y_i	3	6	8	5	7	1	5	3

Grupiranje ćemo provesti primjenom Algoritma 6 pri čemu za početne centre uzmimo $\zeta_1 = (4,4)$, $\zeta_2 = (8,4)$. Principom minimalnih udaljenosti dobivamo početnu particiju

$$\Pi = \{\pi_1, \pi_2\}, \qquad \pi_1 = \{(2,3),(3,6),(5,8),(6,5)\}, \quad \pi_2 = \{(7,7),(8,1),(9,5),(10,3)\}.$$

Daljnji tijek iterativnog postupka sukladno Algoritmu 6 vidljiv je u *Tablici 22.*

(a) Početna iteracija (b) Rješenje



Slika 9: k-means iterativni proces

It.	π_1	π_2	c_1	c_2	$\mathcal{F}(\Pi)$	$\mathcal{G}(\Pi)$
1	$\{(2,3),(3,6),(5,8),(6,5)\}$	$\{(7,7),(8,1),(9,5),(10,3)\}$	(4,5.5)	(8.5,4)	48	45
2	$\{(2,3),(3,6),(5,8),(6,5),(7,7)\}$	$\{(8,1),(9,5),(10,3)\}$	(4.6,5.8)	(9,3)	42	51
3	$\{(2,3),(3,6),(5,8),(6,5),(7,7)\}$	$\{(8,1),(9,5),(10,3)\}$	(4.6,5.8)	(9,3)	42	51

Tablica 22: Traženje optimalne particije skupa $\mathcal{A}$ na osnovi LS-kriterija

Zadatak 7. *Odredite LAD-optimalnu particiju skupa $\mathcal{A}$ iz* Primjera 31 *sastavljenu od dva klastera primijenivši Algoritam 6.*

7.2 Slučaj k klastera

Potražit ćemo lokalno optimalnu particiju skupa $\mathcal{A} \subset \mathbb{R}^n$ s m elemenata koja se sastoji od $1 \leq k \leq m$ klastera. Analogno *Algoritmu 3* k-means algoritam pokrenut ćemo zadavanjem početne particije $\Pi = \{\pi_1, \ldots, \pi_k\}$.

Algoritam 7. `(Izbor početne particije)`

`Korak 1:` Inicijalizacija: učitati m, k i elemente skupa $\mathcal{A}$;
Izabrati početnu particiju: $\Pi = \{\pi_1, \ldots, \pi_k\}$;

`Korak 2:` Priduživanje (assignment step)

$$c_j = \operatorname*{argmin}_{x \in \mathbb{R}^n} \sum_{a \in \pi_j} d(x, a), \quad j = 1, \ldots, k,$$

$$\mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{a \in \pi_j} d(c_j, a);$$

`Korak 3:` Korekcija (update step)

$$\nu_j = \{a \in \mathcal{A} : d(c_j, a) \leq d(c_s, a),\ \forall s = 1, \ldots, k\}, \quad j = 1, \ldots, k$$

`Korak 4:` Ispitivanje optimalnosti: Ako je $\nu_j = \pi_j$ za svaki $j = 1, \ldots, k$, `STOP`; U protivnom staviti

$$\pi_j = \nu_j \quad j = 1, \ldots, k,$$

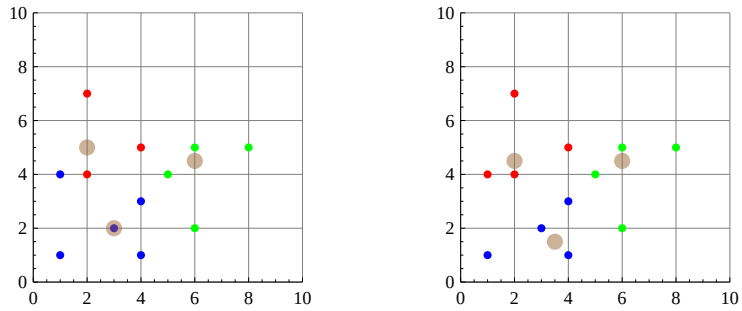
i prijeći na `Korak 2`.

Primjedba 10. Iterativni proces traje tako dugo dok se particije ne počnu ponavljati (u tom slučaju i vrijednost funkcije cilja prestane opadati). U `Korak 3` novi klasteri ν_j geometrijski se određuju pomoću tzv. *Voronoijevog dijagrama* određenog centroidima (vidi primjerice Gan et al. (2007); Kogan et al. (2007); Scitovski and Sabo (2014)).

Primjer 32. *Zadan je skup* $\mathcal{A} = \{a^i = (x_i, y_i) \in \mathbb{R}^2 : i = 1, \ldots, 12\}$ *i početna particija* Π *s klasterima* $\pi_1 = \{a^1, a^2, a^3, a^4, a^5\}$, $\pi_2 = \{a^6, a^7, a^8\}$, $\pi_3 = \{a^9, a^{10}, a^{11}, a^{12}\}$. *Primjenom Algoritma 7 uz LAD metričku funkciju treba odrediti lokalno optimalnu particiju.*

Π	π_1					π_2			π_3			
i	1	2	3	4	5	6	7	8	9	10	11	12
x_i	1	1	3	4	4	2	2	4	5	6	6	8
y_i	1	4	2	1	3	4	7	5	4	2	5	5

(a) Početna iteracija (b) Rješenje



Slika 10: k-means iterativni proces

Centri početne particije su: $c_1 = (3, 2)$, $c_2 = (2, 5)$, $c_3 = (6, 4.5)$ (*Slika 10a*), a funkcija cilja $\mathcal{F}_1 = 23$.

Principom minimalnih udaljenosti uz l_1-metričku funkciju dobivamo nove klastere $\nu_1 = \{a^1, a^3, a^4, a^5\}$, $\nu_2 = \{a^2, a^6, a^7, a^8\}$, $\nu_3 = \{a^9, a^{10}, a^{11}, a^{12}\}$. Nakon toga ponovo izračunamo nove centre $c_1 = (3.5, 1.5)$, $c_2 = (2, 4.5)$, $c_3 = (6, 4.5)$ (*Slika 10b*) i novu vrijednost funkcije cilja $\mathcal{F}_2 = 21$.

Algoritam se također može pokrenuti izborom početnih centara. U tom smislu niže je navedena varijacija Algoritma 7

Algoritam 8. `(Izbor početnih centara)`

`Korak 1:` Inicijalizacija: učitati m, k i elemente skupa $\mathcal{A}$;
Izabrati k različitih početnih centara: $\zeta_1, \cdots, \zeta_k$;

`Korak 2:` Korekcija (update step)

$$\pi_j = \{a \in \mathcal{A} : d(\zeta_j, a) \leq d(\zeta_s, a),\ \forall s = 1, \ldots, k\}, \quad j = 1, \ldots, k$$

`Korak 3:` Pridruživanje (assignment step)

$$c_j = \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \sum_{a \in \pi_j} d(x, a), \quad j = 1, \ldots, k,$$

$$\mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{a \in \pi_j} d(c_j, a);$$

`Korak 4:` Ispitivanje optimalnosti: ako je $c_j = \zeta_j$ za svaki $j = 1, \ldots, k$, `STOP`;
U protivnom staviti

$$\zeta_j = c_j, \quad j = 1, \ldots, k,$$

i prijeći na `Korak 2`.

Primjer 33. *Zadan je skup* $\mathcal{A} = \{a^i = (x_i, y_i) \in \mathbb{R}^2 : i = 1, \ldots, 7\}$. *Primjenom Algoritma 8 uz LS kvazimetričku funkciju i početne centre* $\zeta_1 = (2, 8)$, $\zeta_2 = (4, 5)$, $\zeta_3 = (8, 7)$ *treba odrediti lokalno optimalnu particiju s tri klastera.*

i	1	2	3	4	5	6	7
x_i	1	1	2	4	7	8	9
y_i	3	9	9	6	7	6	8

Podaci i početni centri vidljivi su na *Slici 11a*. Principom minimalnih udaljenosti dobivamo particiju s tri klastera

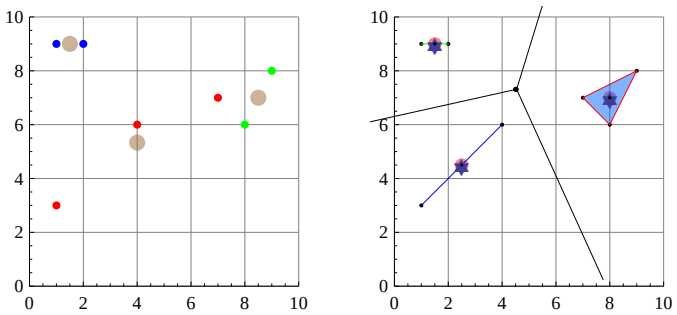
$$\Pi = \{\pi_1, \pi_2, \pi_3\}, \qquad \pi_1 = \{(1, 9), (2, 9)\}, \quad \pi_2 = \{(1, 3), (4, 6)\}, \quad \pi_3 = \{(7, 7), (8, 6), (9, 8)\}.$$

s novim centroidima (*Slika 11b*) i funkcijom cilja

$$c_1 = (1.5, 9), \quad c_2 = (2.5, 4.5), \quad c_3 = (8, 7), \qquad \mathcal{F} = 13.5, \quad \mathcal{G} = 83.07$$

Provjerite da je ovo lokalno optimalna particija.

(a) Početna iteracija (b) Rješenje



Slika 11: k-means iterativni proces

8 Indeksi

Literatura

A. Ben-Israel, C. Iyigun, *Probabilistic D-clustering*, Journal of Classification **25**(2008), 5–26

D. L. Boyd, L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, 2004.

T. Calinski, J. Harabasz, *A dendrite method for cluster analysis*, Communications in Statistics, **3**(1974), 1–27

D. Davies, D. Bouldin, *A cluster separation measure*, IEEE Transactions on Pattern Analysis and Machine Intelligence, **2**(1979), 224–227

I. S. Dhillon, Y. Guan, B. Kulis, *Kernel k-means, spectral clustering and normalized cuts*, Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), August 22–25, 2004, Seattle, Washington, USA, 551–556, 2004

Z. Drezner, *Facility Location: A Survey of Aplications and Methods*, Springer-Verlag, Berlin, 2004.

B. S. Everitt, S. Landau, M. Leese, *Cluster analysis*, Wiley, London, 2001.

C. A. Floudas, C. E. Gounaris, *A review of recent advances in global optimization*, J. Glob. Optim. **45**(2009), 3-–38

G. Gan, C Ma, J. Wu, *Data Clustering: Theory, Algorithms, and Applications*, SIAM, Philadelphia, 2007.

M. Hudec, M. Vujošević, *A fuzzy system for municipalities classification*, CEJOR **18**(2010), 171–180

C. Iyigun, A. Ben-Israel, *A generalized Weiszfeld method for the multi-facility location problem*, Operations Research Letters **38**(2010), 207–214

C. Iyigun, *Probabilistic Distance Clustering*, Dissertation, Graduate School – New Brunswick, Rutgers, 2007

J. Kogan, *Introduction to Clustering Large and High-Dimensional Data*, Cambridge University Press, 2007.

J. Kogan, C. Nicholas, M. Wiacek, *Hybrid Clustering of large high dimensional data*, In M. Castellanos and M. W. Berry (Eds.), Proceedings of the Workshop on Text Mining, SIAM, 2007.

J. Kogan, M. Teboulle, *Scaling clustering algorithms with Bregman distances*. In: M. W. Berry and M. Castellanos (Eds.), Proceedings of the Workshop on Text Mining at the Sixth SIAM International Conference on Data Mining, 2006.

J. Kogan, C. Nicholas, M. Wiacek, *Hybrid clustering with divergences*. In: M. W. Berry and M. Castellanos (Eds.), Survey of Text Mining: Clustering, Classification, and Retrieval, Second Edition, Springer, 2007.

Leisch, F., 2006. A toolbox for k-centroids cluster analysis. Computational Statistics & Data Analysis 51, 526–544.

K. Sabo, R. Scitovski, *The best least absolute deviations line – properties and two efficient methods*, ANZIAM Journal **50**(2008), 185–198

K. Sabo, R. Scitovski, I. Vazler, *Grupiranje podataka> klasteri*, Osječki matematički list **10**(2010), 149–178

K. Sabo, R. Scitovski, I. Vazler, M. Zekić-Sušac, *Mathematical models of natural gas consumption*, Energy Conversion and Management **52**(2011), 1721-1727

K. Sabo, R. Scitovski, I. Vazler, *One-dimensional center-based l_1-clustering method*, Optimization Letters **7**(2013), 5-22

A. Schöbel, *Locating Lines and Hyperplanes: Theory and Algorithms*, Springer Verlag, Berlin, 1999.

R. Scitovski, K. Sabo, *Analysis of the k-means algorithm in the case of data points occurring on the border of two or more clusters*, Knowledge-Based Systems **57**(2014), 1-7

H. Späth, *Cluster-Formation und Analyse*, R. Oldenburg Verlag, München, 1983.

M. Teboulle, *A unified continuous optimization framework for center-based clustering methods*, Journal of Machine Learning Research **8**(2007), 65–102

D. Veljan, *Kombinatorna i diskretna matematika*, Algoritam, Zagreb, 2001.