

## Matematički praktikum

K. Sabo, R. Scitovski, I. Vazler

### 1 Problem mjerenja

Zadani su podaci mjerenja  $y_1, y_2, \dots, y_m \in \mathbb{R}$ .

$y = (y_1, y_2, \dots, y_m)^T$  – vektor podataka

Tražimo vektor  $a = (\lambda, \dots, \lambda)^T$  tako da  $d(y, a) = \|y - a\|$  bude minimalno.

#### 1.1 Najbolja $l_2$ aproksimacija mjerenja

$$d_2(y, a) = \|y - a\|_2 = \sqrt{\sum_{i=1}^m (y_i - \lambda)^2} =: G(\lambda) \rightarrow \min_{\lambda} \quad (1)$$

Buduci da vrijedi

$$\lambda^* = \operatorname{argmin}_{\lambda \in \mathbb{R}} G(\lambda) \quad \Leftrightarrow \quad \lambda^* = \operatorname{argmin}_{\lambda \in \mathbb{R}} G^2(\lambda),$$

onda u cilju traženja  $\lambda^* \in \mathbb{R}$ , koji je rješenje od (11), možemo minimizirati

$$\|y - a\|_2^2 = \sum_{i=1}^m (y_i - \lambda)^2 \rightarrow \min_{\alpha}.$$

(Metoda najmanjih kvadrata – Least Squares Method)

$$\lambda^* = \frac{1}{m} \sum_{i=1}^m y_i =: \bar{y} \quad (\text{aritmetička sredina}) \quad (2)$$

Dakle, vrijedi

$$\sum_{i=1}^m (y_i - \lambda)^2 \geq \sum_{i=1}^m (y_i - \bar{y})^2 \quad \forall \lambda \in \mathbb{R}. \quad (3)$$

**Zadatak 1.** *Nacrtati odgovarajuće slike podataka i minimizirajuće funkcije.*

## 1.2 Najbolja $l_1$ aproksimacija mjerenja

Minimizirajući sumu apsolutnih odstupanja (**metoda najmanjih apsolutnih odstupanja – Least Absolute Deviations Method**)

$$d_1(y, a) = \|y - a\|_1 = \sum_{i=1}^m |y_i - \lambda| \rightarrow \min_{\lambda}$$

dobivamo

$$\lambda^* = \text{med}(y) \quad (\text{medijan podataka}). \quad (4)$$

Rejesavanje ovog problema u literaturi se zove (**obični medijan problem**). Vrijedi

$$\sum_{i=1}^m |y_i - \lambda| \geq \sum_{i=1}^m |y_i - \text{med}(y)| \quad \forall \lambda \in \mathbb{R}. \quad (5)$$

**Zadatak 2.** *Nacrtati odgovarajuće slike podataka i minimizirajuće funkcije.*

## 1.3 Najbolja $l_\infty$ aproksimacija mjerenja

Minimizirajući maksimalno apsolutno odstupanje (**Minimax problem**)

$$d_\infty(y, a) = \|y - a\|_\infty = \max_{i=1, m} |y_i - \lambda| \rightarrow \min_{\lambda}$$

dobivamo

$$\lambda^* = \frac{y_{\min} + y_{\max}}{2}. \quad (6)$$

Dakle, vrijedi

$$\max_{i=1, m} |y_i - \lambda| \geq \max_{i=1, m} |y_i - \frac{y_{\min} + y_{\max}}{2}| \quad \forall \lambda \in \mathbb{R}. \quad (7)$$

**Zadatak 3.** *Nacrtati odgovarajuće slike podataka i minimizirajuće funkcije.*

**Zadatak 4.** *Zadani su podaci mjerenja  $y_1, \dots, y_m \in \mathbb{R}$ . Treba odrediti centre  $c_1, c_2 \in \mathbb{R}$  tako da bude*

$$f(c_1, c_2) = \sum_{i=1}^m \min\{d(y_i, c_1), d(y_i, c_2)\} \rightarrow \min_{c_1, c_2 \in \mathbb{R}}.$$

*U literaturi ovaj problem se nalazi pod imenom 2-medijan (2-centar) problem. Za slučajno izabrane podatke grafički ilustrirajte problem, a rješenje potražite nekim gotovim programom. Za različite razdaljinske funkcije  $d: \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}_+$  analizirajte svojstva minimizirajuće funkcije  $f$ . Definirajte  $k$ -medijan problem za  $k \leq m$ . Što je rješenje u slučaju  $k = m$ ?*

## 1.4 Najbolja $l_p$ , $p \geq 1$ aproksimacija mjerenja

$$d_p(y, a) = \|y - a\|_p = \left( \sum_{i=1}^m |y_i - \lambda|^p \right)^{\frac{1}{p}} \rightarrow \min_{\lambda}$$

## 1.5 Najbolja težinska aproksimacija mjerenja

Zadani su podaci  $(w_i, y_i)$ ,  $i = 1, \dots, m$ , gdje su  $w_i > 0$  težine podataka.

Težinske norme:

$$\|x\|_1 = \sum_{i=1}^m w_i |x_i|, \quad \|x\|_2 = \sqrt{\sum_{i=1}^m w_i x_i^2}, \quad \|x\|_\infty = \max_{i,m} w_i |x_i|, \quad \|x\|_p = \left( \sum_{i=1}^m w_i |x_i|^p \right)^{\frac{1}{p}}$$

**Za vježbe:** osjetljivost (robusnost) pojedinih normi.

### 1.5.1 Najbolja težinska $l_2$ aproksimacija mjerenja

Kako je

$$\operatorname{argmin}_{\lambda \in \mathbb{R}} \sqrt{\sum_{i=1}^m w_i (y_i - \lambda)^2} = \operatorname{argmin}_{\lambda \in \mathbb{R}} \sum_{i=1}^m w_i (y_i - \lambda)^2,$$

lako dobivamo

$$\lambda^* = \bar{y} = \frac{1}{W} \sum_{i=1}^m w_i y_i, \quad W = \sum_{i=1}^m w_i \quad (\text{težinska aritmetička sredina}). \quad (8)$$

Primijetimo također da, slično kao u (3), vrijedi

$$\sum_{i=1}^m w_i (y_i - \lambda)^2 \geq \sum_{i=1}^m w_i (y_i - \bar{y})^2 \quad \forall \lambda \in \mathbb{R}.$$

### 1.5.2 Najbolja težinska $l_1$ aproksimacija mjerenja

Minimizirajući sumu težinskih apsolutnih odstupanja

$$\sum_{i=1}^m w_i |y_i - \lambda| \rightarrow \min_{\lambda},$$

dobivamo težinski medijan podataka

$$\lambda^* = \operatorname{med}(w_i, y_i). \quad (9)$$

U literaturi ovo se obično naziva **težinski medijan problem**. Primijetimo također da, slično kao u (3), vrijedi

$$\sum_{i=1}^m w_i |y_i - \lambda| \geq \sum_{i=1}^m w_i |y_i - \operatorname{med}(w_i, y_i)| \quad \forall \lambda \in \mathbb{R}.$$

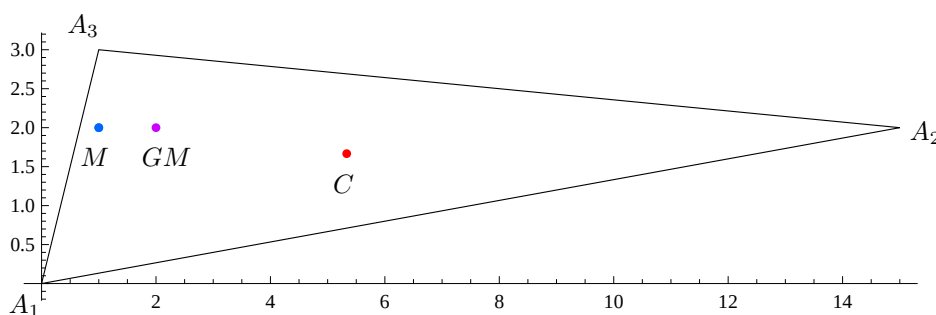
## 2 Fermat – Torricelli – Weberov problem

### 2.1 Tri točke u ravnini

#### 2.1.1 Centroid

Neka su  $A_1, A_2, A_3 \in \mathbb{R}^2$  tri nekolinearne točke u ravnini. Točka  $C \in \mathbb{R}^2$ , za koju je suma kvadrata euklidskih udaljenosti do vrhova trokuta minimalna zove se **centroid** ili **Steinerova točka** (povezano s pojmom centra masa u fizici). Naka je

$$A_1 = (x_1, y_1), \quad A_2 = (x_2, y_2), \quad A_3 = (x_3, y_3).$$



Tada je centroid  $C = (x_c, y_c)$  točka za koju se postiže minimum

$$\sum_{i=1}^3 d_2^2(T, A_i) \rightarrow \min_{T \in \mathbb{R}^2}, \quad tj. \quad \sum_{i=1}^3 [(x - x_i)^2 + (y - y_i)^2] \rightarrow \min_{(x,y) \in \mathbb{R}^2}$$

$$\text{odakle dobivamo: } x_c = \frac{1}{3} \sum_{i=1}^3 x_i, \quad y_c = \frac{1}{3} \sum_{i=1}^3 y_i.$$

#### 2.1.2 Median

Točka  $M = (x_M, y_M)$  za koju se postiže minimum

$$\sum_{i=1}^3 d_1(T, A_i) \rightarrow \min_{T \in \mathbb{R}^2}, \quad tj. \quad \sum_{i=1}^3 (|x - x_i| + |y - y_i|) \rightarrow \min_{(x,y) \in \mathbb{R}^2}$$

naziva se **medijan** točaka  $A_1, A_2, A_3 \in \mathbb{R}^2$ . Lako dobivamo

$$M = (x_M, y_M), \quad x_M = \text{med } x_i, \quad y_M = \text{med } y_i.$$

#### 2.1.3 Geometrijski median

Točka  $GM = (x_{GM}, y_{GM})$  za koju se postiže minimum

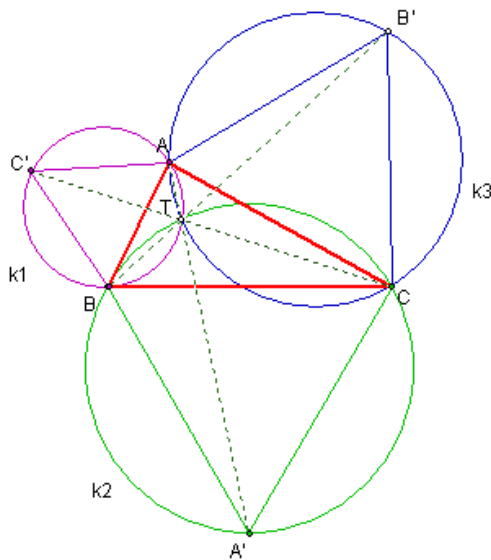
$$\sum_{i=1}^3 d_2(T, A_i) \rightarrow \min_{T \in \mathbb{R}^2}, \quad tj. \quad \sum_{i=1}^3 \sqrt{(x - x_i)^2 + (y - y_i)^2} \rightarrow \min_{(x,y) \in \mathbb{R}^2}$$

naziva se **geometrijski medijan** točkaka  $A_1, A_2, A_3 \in \mathbb{R}^2$  i ne može se eksplicitno izraziti.

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

Geometrijski medijan  $M = (x_{GM}, y_{GM})$  može se dobiti geometrijskim konstrukcijama na dva načina:

- kao sjecište Torricellijevih kružnica;
- kao sjecište Simpsonovih pravaca.



Slika 1: Sliku izradile: D. Jankov, S. Sušić

## 2.2 Problem $n$ točkaka u ravnini s odgovarajućim težinama

Zadan je skup točkaka  $T_i = (x_i, y_i)$ ,  $i = 1, \dots, m$ ,  $m \geq 2$  u ravnini s odgovarajućim težinama  $w_i > 0$ .

### 2.2.1 Centroid

Točku  $C \in \mathbb{R}^2$  takva da suma kvadrata težinskih euklidskih udaljenosti točkaka  $T_i$  do nje bude minimalna,

$$\sum_{i=1}^m w_i d_2^2(T, T_i) \rightarrow \min_{T \in \mathbb{R}^2}, \quad tj. \quad \sum_{i=1}^m w_i ((x - x_i)^2 + (y - y_i)^2) \rightarrow \min_{(x, y) \in \mathbb{R}^2} \quad (10)$$

zovemo **težinski centroid** točkaka  $T_i$ ,  $i = 1, \dots, m$  u ravnini. Fizikalno, točka  $C$  predstavlja centar masa, tj. **težište** diskretnog tijela sastavljeno od materijalnih točkaka  $T_i$  s težinama  $w_i$ . Primjenom Zadatka 7 lako se može vidjeti da je  $C = (\bar{x}, \bar{y})$ , gdje je

$$\bar{x} = \frac{1}{W} \sum_{i=1}^m w_i x_i, \quad \bar{y} = \frac{1}{W} \sum_{i=1}^m w_i y_i, \quad W = \sum_{i=1}^m w_i.$$

### 2.2.2 Median

Treba odrediti točku  $M \in \mathbb{R}^2$  tako da suma težinskih  $l_1$  udaljenosti točkaka  $T_i$  do točke  $M$  bude minimalna, tj.

$$M = \operatorname{argmin}_{T \in \mathbb{R}^2} \sum_{i=1}^m w_i d_1(T, T_i), \quad \text{tj.} \quad \sum_{i=1}^m w_i (|x - x_i| + |y - y_i|) \rightarrow \min_{(x,y) \in \mathbb{R}^2} \quad (11)$$

Točka  $M = (x_M, y_M)$  s navedenim svojstvo naziva se **težinski medijan** točkaka  $T_i$ ,  $i = 1, \dots, m$  u ravnini. Primjenom Zadatka 8 može se dobiti

$$x_M = \operatorname{med}_{i=1, \dots, m} (w_i, x_i), \quad y_M = \operatorname{med}_{i=1, \dots, m} (w_i, y_i)$$

### 2.2.3 Geometrijski median

Točku  $GM \in \mathbb{R}^2$  takvu da suma težinskih euklidskih udaljenosti točkaka  $T_i$  do nje bude minimalna,

$$GM = \operatorname{argmin}_{T \in \mathbb{R}^2} \sum_{i=1}^m w_i d_2(T, T_i), \quad \text{tj.} \quad F(x, y) = \sum_{i=1}^m w_i \sqrt{(x - x_i)^2 + (y - y_i)^2} \rightarrow \min_{(x,y) \in \mathbb{R}^2} \quad (12)$$

zovemo **težinski geometrijski median** točkaka  $T_i$ ,  $i = 1, \dots, m$  u ravnini i ne može se eksplicitno izraziti.

### 2.2.4 Weiszfeldov algoritam za traženje geometrijskog mediana

Stacionarne točke funkcionala  $F$  iz (12) određene su s  $\frac{\partial F}{\partial x} = \frac{\partial F}{\partial y} = 0$ , iz čega dobivamo sustav

$$\sum_{i=1}^m w_i \frac{x - x_i}{\sqrt{(x - x_i)^2 + (y - y_i)^2}} = 0, \quad \sum_{i=1}^m w_i \frac{y - y_i}{\sqrt{(x - x_i)^2 + (y - y_i)^2}} = 0, \quad (13)$$

koji možemo zapisati u obliku

$$x = \varphi(x, y), \quad y = \psi(x, y), \quad (14)$$

$$\varphi(x, y) = \frac{\sum_{i=1}^m w_i \frac{x_i}{\sqrt{(x - x_i)^2 + (y - y_i)^2}}}{\sum_{i=1}^m w_i \frac{1}{\sqrt{(x - x_i)^2 + (y - y_i)^2}}}, \quad \psi(x, y) = \frac{\sum_{i=1}^m w_i \frac{y_i}{\sqrt{(x - x_i)^2 + (y - y_i)^2}}}{\sum_{i=1}^m w_i \frac{1}{\sqrt{(x - x_i)^2 + (y - y_i)^2}}}. \quad (15)$$

Izaberemo početnu aproksimaciju  $(x_0, y_0) \in \mathbb{R}^2$ . Sljedeći iterativni proces, koji se zasniva na metodi jednostavnih iteracija, naziva se **Weiszfeldov algoritam**.

$$x^{k+1} = \varphi(x^k, y^k), \quad y^{k+1} = \psi(x^k, y^k), \quad k = 0, 1, \dots, \quad (16)$$

što možemo pisati

$$x^{k+1} = \varphi(x^k, y^k) = \sum_{i=1}^m \omega_i(x^k, y^k) x_i, \quad y^{k+1} = \psi(x^k, y^k) = \sum_{i=1}^m \omega_i(x^k, y^k) y_i$$

$$\omega_i(x, y) = \frac{\frac{w_i}{\sqrt{(x-x_i)^2 + (y-y_i)^2}}}{\sum_{s=1}^m w_s \frac{1}{\sqrt{(x-x_s)^2 + (y-y_s)^2}}}, \quad i = 1, \dots, m.$$

Lako se može vidjeti da za proizvoljne  $x, y \in \mathbb{R}$  vrijedi

$$0 < \omega_i(x, y) < 1, \quad \sum_{i=1}^m \omega_i(x, y) = 1.$$

## 2.3 Problem $m$ točaka u $\mathbb{R}^n$ s odgovarajućim težinama

Neka je zadan skup vektora  $\mathcal{A} = \{a^i = (a_1^i, \dots, a_n^i)^T \in \mathbb{R}^n : i = 1, \dots, m\} \subset \mathbb{R}^n$  s odgovarajućim težinama  $w_i > 0$ .

### 2.3.1 Centroid

Zbog konveksnosti funkcije  $F_2$  iz Zadatka 5 postoji vektor  $c^* \in \mathbb{R}^n$ , takav da je

$$c^* = \operatorname{argmin}_{x \in \mathbb{R}^n} F_2(x), \quad F_2(x) = \sum_{i=1}^m w_i \|x - a^i\|_2^2. \quad (17)$$

Vektor  $c^*$  zovemo **težinski centroid** skupa vektora  $\mathcal{A} \subset \mathbb{R}^n$ , a dobije se iz

$$\frac{\partial F_2}{\partial x} = 2 \sum_{i=1}^m w_i (x - a^i) = 0.$$

$$c^* = \frac{1}{W} \sum_{i=1}^m w_i a^i, \quad W = \sum_{i=1}^m w_i. \quad (18)$$

### 2.3.2 Median

Zbog konveksnosti funkcije  $F_1$  iz Zadatka 5 postoji vektor  $c^* \in \mathbb{R}^n$ , takav da je

$$c^* = \operatorname{argmin}_{x \in \mathbb{R}^n} F_1(x), \quad F_1(x) = \sum_{i=1}^m w_i \|x - a^i\|_1. \quad (19)$$

Vektor  $c^*$  zovemo **težinski median** skupa vektora  $\mathcal{A} \subset \mathbb{R}^n$ . Vrijedi

$$c^* = \left( \operatorname{med}_{j=1,\dots,m} (w_j, a_1^j), \dots, \operatorname{med}_{j=1,\dots,m} (w_j, a_n^j) \right)^T. \quad (20)$$

Dokaz formule (20) provest ćemo na osnovi t. 1.5.2 i Zadatka 8

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \sum_{i=1}^m w_i \|x - a^i\|_1 &= \min_{x \in \mathbb{R}^n} \sum_{i=1}^m w_i \left( \sum_{k=1}^n |x_k - a_k^i| \right) \\ &= \min_{x \in \mathbb{R}^n} \sum_{k=1}^n \left( \sum_{i=1}^m w_i |x_k - a_k^i| \right) = \sum_{k=1}^n \min_{x_k \in \mathbb{R}} \sum_{i=1}^m w_i |x_k - a_k^i| \\ &= \sum_{k=1}^n \sum_{i=1}^m w_i \left| \operatorname{med}_{j=1,\dots,m} (w_j, a_k^j) - a_k^i \right| = \sum_{i=1}^m \sum_{k=1}^n w_i \left| \operatorname{med}_{j=1,\dots,m} (w_j, a_k^j) - a_k^i \right| \\ &= \sum_{i=1}^m w_i \|c^* - a^i\|_1. \end{aligned}$$

### 2.3.3 Geometrijski median

Zbog konveksnosti funkcije  $F$  iz Zadatka 5 postoji vektor  $c^* \in \mathbb{R}^n$ , takav da je

$$c^* = \operatorname{argmin}_{x \in \mathbb{R}^n} F(x), \quad F(x) = \sum_{i=1}^m w_i \|x - a^i\|_2. \quad (21)$$

Vektor  $c^*$  zovemo **težinski geometrijski median** skupa vektora  $\mathcal{A} \subset \mathbb{R}^n$  i ne može se eksplicitno izraziti.

### 2.3.4 Weiszfeldov algoritam za traženje geometrijskog mediana

Stacionarne točke funkcionala  $F$  iz (21) poopćenjem formula (13) određene su s

$$\operatorname{grad} F(x) = \sum_{i=1}^m w_i \|x - a^i\|^{-1} (x - a^i) = 0, \quad (22)$$

što možemo zapisati u obliku

$$x = \frac{\sum_{i=1}^m w_i \|x - a^i\|^{-1} a^i}{\sum_{j=1}^m w_j \|x - a^j\|^{-1}},$$

odnosno

$$x = \sum_{i=1}^m \omega_i(x) a^i, \quad \omega_i(x) = \frac{w_i \|x - a^i\|^{-1}}{\sum_{j=1}^m w_j \|x - a^j\|^{-1}} \quad (23)$$

Očigledno za svaki  $x \in \mathbb{R}^n$  vrijedi

$$0 < \omega_i(x) < 1, \quad \sum_{i=1}^m \omega_i(x) = 1. \quad (24)$$

Izaberemo početnu aproksimaciju  $x_0 \in \mathbb{R}^n$ . Sljedeći iterativni proces, koji se zasniva na metodi jednostavnih iteracija, naziva se **Weiszfeldov algoritam**.

$$x^{k+1} = \sum_{i=1}^m \omega_i(x^k) a^i, \quad k = 0, 1, \dots \quad (25)$$

## 2.4 Zadaci

**Zadatak 5.** Pokažite da su funkcija,

$$\begin{aligned} (i) \quad F_2: \mathbb{R}^n &\rightarrow \mathbb{R}, & F_2(x) &= \sum_{i=1}^m w_i \|x - a_i\|_2^2, \\ (ii) \quad F_1: \mathbb{R}^n &\rightarrow \mathbb{R}, & F_1(x) &= \sum_{i=1}^m w_i \|x - a_i\|_1, \\ (iii) \quad F: \mathbb{R}^n &\rightarrow \mathbb{R}, & F(x) &= \sum_{i=1}^m w_i \|x - a_i\|_2, \end{aligned}$$

konveksne. Jesu li sve ove funkcije strogo konveksne?

**Zadatak 6.** Definirajte pojmove analogne centroidu i geometrijskom medijanu, tako da umjesto minimizacije "sume udaljenosti" minimizirate "maksimalna udaljenost". Ovaj problem u literaturi se obično naziva **problem centra**. Izradite odgovarajuće ilustracije.

**Zadatak 7.** Ako su  $\varphi_i \in C^2(R)$ ,  $i = 1, \dots, n$  konveksne dvostruko neprekidno diferencijabilne funkcije za koje je

$$\hat{c}_i = \operatorname{argmin}_{x \in \mathbb{R}} \varphi_i(x), \quad i = 1, \dots, n,$$

pokažite da je tada i funkcija  $f: \mathbb{R}^n \rightarrow \mathbb{R}$ ,

$$f(x_1, \dots, x_n) = \varphi_1(x_1) + \dots + \varphi_n(x_n),$$

konveksna dvostruko neprekidno diferencijabilne funkcija za koju vrijedi

$$\operatorname{argmin}_{x^n \in \mathbb{R}^n} f(x) = (\hat{c}_1, \dots, \hat{c}_n)^T.$$

Vrijedi li obrat? Primjerice, ako je  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  konveksna dvostruko neprekidno diferencijabilne funkcija i ako je  $f(x_1, x_2) = \varphi_1(x_1) + \varphi_2(x_2)$ , moraju li tada i funkcije  $\varphi_1, \varphi_2$  biti konveksna dvostruko neprekidne diferencijabilne funkcije. Izradite odgovarajuće primjere.

**Zadatak 8.** Ako su  $\varphi_i \in C(R)$ ,  $i = 1, \dots, n$  konveksne nediferencijabilne funkcije za koje je

$$\hat{c}_i = \operatorname{argmin}_{x \in \mathbb{R}} \varphi_i(x), \quad i = 1, \dots, n,$$

pokažite da je tada i funkcija  $f: \mathbb{R}^n \rightarrow \mathbb{R}$ ,

$$f(x_1, \dots, x_n) = \varphi_1(x_1) + \dots + \varphi_n(x_n),$$

konveksna funkcija za koju vrijedi

$$\operatorname{argmin}_{x^n \in \mathbb{R}^n} f(x) = (\hat{c}_1, \dots, \hat{c}_n)^T.$$

Vrijedi li obrat? Primjerice, ako je  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  konveksna (pa onda i neprekidna) nediferencijabilna funkcija i ako je  $f(x_1, x_2) = \varphi_1(x_1) + \varphi_2(x_2)$ , moraju li tada i funkcije  $\varphi_1, \varphi_2$  biti konveksne (pa onda i neprekidne) funkcije. Izradite odgovarajuće primjere.

**Zadatak 9.** Na osnovi t. 1.5.1 i Zadatka 7, pokažite da se minimum funkcionala  $F_2$ , zadanog sa (17) postiže u točki  $c^*$ , zadanoj s (18).

**Zadatak 10.** Zadan je skup realnih brojeva  $y_1 \leq y_2 \leq \dots \leq y_m$  s odgovarajućim težinama  $w_i > 0$ , takvim da je  $\sum_{i=1}^m w_i = 1$ . Pokažite da tada vrijedi

$$y_1 \leq \sum_{i=1}^m w_i y_i \leq y_m.$$

## 3 Klasteri: grupiranje (segmentiranje) podataka u skupine

### 3.1 Klasteri u $\mathbb{R}$ : Least Square princip

Neka je  $\mathcal{A} = \{a_1, \dots, a_m\} \subset \mathbb{R}$  skup realnih brojeva. *Partitivni skup* skupa  $\mathcal{A}$  je familija svih njegovih podskupova. Uobičajena oznaka je  $\mathcal{P}(\mathcal{A})$ .

Rastav skupa  $\mathcal{A}$  na disjunktne podskupove  $\pi_1, \pi_2$ , takve da vrijedi

$$\pi_1 \cup \pi_2 = \mathcal{A}, \quad \pi_1 \cap \pi_2 = \emptyset, \quad m_1 = |\pi_1| \geq 1, \quad m_2 = |\pi_2| \geq 1. \quad (26)$$

označit ćemo s  $\Pi(\mathcal{A}) = \{\pi_1, \pi_2\}$  i zvati *particija skupa*  $\mathcal{A}$ . Skupove  $\pi_1, \pi_2$  zvat ćemo *klasteri*. Skup svih particija skupa  $\mathcal{A}$  sastavljenih od 2 klastera koje zadovoljavaju (26) označit ćemo s  $\mathcal{P}(\mathcal{A}, k)$ . Broj svih ovakvih particija je

$$N(m, 2) = \frac{1}{2}(2^m - 2). \quad (27)$$

*Primjedba 1.* Ovaj problem spada u grupu tzv. NP-teških problema. Primjerice ako je  $m = 50$ , onda je sukladno (34) moguće načiniti  $\frac{1}{2}(2^{50} - 2) \approx 563\,000\,000\,000\,000$  različitih particija.

Za podskupove  $\pi_1, \pi_2$  odredimo centroeide  $c_1, c_2$  tako da suma kvadrata udaljenosti svih  $a_i \in \pi_1$  do  $c_1$ , odnosno suma kvadrata udaljenosti svih  $a_j \in \pi_2$  do  $c_2$ , bude minimalna

$$\begin{aligned} \min_{x \in \mathbb{R}} \sum_{a_i \in \pi_1} (a_i - x)^2 &\Rightarrow c_1 = \frac{1}{m_1} \sum_{a_i \in \pi_1} a_i, \\ \min_{x \in \mathbb{R}} \sum_{a_i \in \pi_2} (a_i - x)^2 &\Rightarrow c_2 = \frac{1}{m_2} \sum_{a_i \in \pi_2} a_i. \end{aligned}$$

Mjeru kvalitete particije  $\Pi = \{\pi_1, \pi_2\}$  s centroidima  $c_1, c_2$  definirat ćemo funkcijom  $F: \mathcal{P}(\mathcal{A}, 2) \rightarrow \mathbb{R}_+$

$$F(\Pi) = \sum_{a_i \in \pi_1} (a_i - c_1)^2 + \sum_{a_i \in \pi_2} (a_i - c_2)^2, \quad (28)$$

koji predstavlja zbroj sume kvadrata udaljenosti svih točaka klastera  $\pi_1$  do centroida  $c_1$  i sume kvadrata udaljenosti svih točaka klastera  $\pi_2$  do centroida  $c_2$ .

Nakon što smo odredili sve particije  $\Pi$  skupa  $\mathcal{A}$ , koje zadovoljavaju (26) i izračunali vrijednosti odgovarajuće funkcije cilja  $F(\Pi)$ , među njima potražiti ćemo onu particiju za koju funkcija  $F$  prima najmanju vrijednost. Tada kažemo da smo pronašli najbolju (**optimalnu**) particiju skupa  $\mathcal{A}$  u smislu kriterija (28)

**Primjer 1.** Treba pronaći sve particije skupa  $\mathcal{A} = \{1, 3, 4, 8\}$ , koje zadovoljavaju (26), odrediti pripadne centroeide i vrijednosti funkcije cilja  $F$ .

| $\pi_1$    | $\pi_2$       | $c_1$ | $c_2$  | $F(\Pi)$       |       | $G(\Pi)$      |       |
|------------|---------------|-------|--------|----------------|-------|---------------|-------|
| $\{1\}$    | $\{3, 4, 8\}$ | 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     |

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

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

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

$$\min_{c_1, c_2 \in \mathbb{R}} \Phi(c_1, c_2), \quad \Phi(c_1, c_2) = \sum_{i=1}^m \min\{d(c_1, a_i), d(c_2, a_i)\}, \quad (29)$$

gdje je  $\Phi: \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}_+$ . Općenito, ovaj funkcional nije ni diferencijabilan, ni konveksan i može imati više lokalnih minimuma. Optimizacijski problem (29) u literaturi se može naći pod nazivom ***k*-median problem** i najčešće se rješava raznim metaheurističkim metodama (Domínguez and Muñoz, 2005) ili primjenom cjelobrojnog programiranja (Pan and Chen, 2007; Schöbel and Scholz, 2010; Rodríguez-Chia et al., 2010). Pregled radova iz ovog područja do 2006. godine može se vidjeti u (Reese, 2006).

### 3.1.1 Standardni k-means algorithm

**Algoritam 1.** **Step 1:** Izaberi početnu particiju klastera  $\Pi = \{\pi_1, \pi_2\}$ , izračunaj centre  $c_1, c_2$  i početnu vrijednost funkcije cilja  $F_0 = F(\Pi)$ ;

**Step 2:** Formiraj novu particiju  $\mathcal{V} = \{\nu_1, \nu_2\}$  s novim klasterima definiranim s

$$\begin{aligned}\nu_1 &= \{a_i \in \mathcal{A} : d(a_i, c_1) \leq d(a_i, c_2)\}, \\ \nu_2 &= \{a_i \in \mathcal{A} : d(a_i, c_2) < d(a_i, c_1)\},\end{aligned}$$

njihove centre  $\zeta_1, \zeta_2$  i vrijednost funkcije cilja  $F_1 = F(\mathcal{V})$ ;

**Step 3:** Ako je  $F_1 < F_0$ , stavi  $c_1 = \zeta_1$ ;  $c_2 = \zeta_2$ ;  $F_0 = F_1$  i prijedi na **Step 2**;  
U protivnom STOP.

*Primjedba 2.* Primijetite da se u **Step 2** novi klasteri  $\nu_1, \nu_2$  geometrijski mogu odrediti tako da odredimo polovište  $\bar{c} = \frac{c_1 + c_2}{2}$  spojnice centroida. Tada svi elementi lijevo od polovišta  $\bar{c}$  pripadaju novom klasteru  $\nu_1$ , a svi elementi desno od polovišta  $\bar{c}$  pripadaju novom klasteru  $\nu_2$ .

**Primjer 2.**  $\mathcal{A} = \{1, 2, 6, 7, 9\}$ ,  $k = 2$ ,  $\Pi^0 = \{\{1, 2, 6\}, \{7, 9\}\}$ .

| R.br. | Klasteri                 | Centroidi                    | Funkcija cilja               |
|-------|--------------------------|------------------------------|------------------------------|
| 1     | $\{1, 2, 6\}$ $\{7, 9\}$ | 3    8                       | 16                           |
| 2     | $\{1, 2\}$ $\{6, 7, 9\}$ | $\frac{3}{2}$ $\frac{22}{3}$ | $\frac{31}{6} \approx 5.167$ |

**Primjer 3.**  $\mathcal{A} = \{0, 2, 3\}$ ,  $k = 2$ ,  $\Pi^0 = \{\{0, 2\}, \{3\}\}$ .

| R.br. | Klasteri           | Centroidi | Funkcija cilja |
|-------|--------------------|-----------|----------------|
| 1     | $\{0, 2\}$ $\{3\}$ | 1    3    | 2              |
| 1     | $\{0, 2\}$ $\{3\}$ | 1    3    | 2              |

Ovo je primjer koji pokazuje da standardni *k*-means algoritam ne daje optimalno rješenje. Bolja particija je  $\Pi_1 = \{\{0\}, \{2, 3\}\}$  jer je  $F(\Pi_1) = 0.25$ .

### 3.1.2 Princip maksimalne kvadratne udaljenosti centara klastera

Neka su  $c_1, c_2$  centri klastera  $\pi_1, \pi_2$ . Tada za svaki  $x \in \mathbb{R}$  vrijedi

$$\sum_{a_i \in \pi_j} (a_i - x)^2 = \sum_{a_i \in \pi_j} (a_i - c_j)^2 + m_j(c_j - x)^2, \quad j = 1, 2. \quad (30)$$

$$\begin{aligned} \sum_{a_i \in \pi_1} (a_i - x)^2 &= \sum_{a_i \in \pi_1} ((a_i - c_1) + (c_1 - x))^2 \\ &= \sum_{a_i \in \pi_1} (a_i - c_1)^2 + m_1(c_1 - x)^2, \\ \text{jer je} \quad \sum_{a_i \in \pi_1} (a_i - c_1)(c_1 - x) &= (c_1 - x) \sum_{a_i \in \pi_1} (a_i - c_1) = 0. \end{aligned}$$

Ako u (30) umjesto  $x$  stavimo  $c = \frac{1}{m} \sum_{i=1}^m a_i$  i zbrojimo obje jednakosti, dobivamo

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

odnosno

$$\begin{aligned} \sum_{i=1}^m (a_i - c)^2 &= F(\Pi) + G(\Pi), \\ G(\Pi) &= m_1(c_1 - c)^2 + m_2(c_2 - c)^2. \end{aligned} \quad (31)$$

*Primjedba 3.* Ako su  $\varphi, \psi \in C^2(\mathbb{R})$ , takve funkcije za koje vrijedi  $\varphi(x) + \psi(x) = \text{const}$ , onda

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

**Teorem 1.** *Uz prethodne oznake vrijedi:*

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

To znači da u cilju pronalaženja optimalne LS particije, umjesto minimizacije funkcije (28) možemo maksimizirati funkciju (vidi Primjer 1)

$$G(\Pi) = m_1(c_1 - c)^2 + m_2(c_2 - c)^2 \quad (32)$$

### 3.1.3 Problem $k$ klastera u $\mathbb{R}$

Neka je  $\mathcal{A} = \{a_1, \dots, a_m\} \subset \mathbb{R}$  skup realnih brojeva. *Partitivni skup* skupa  $\mathcal{A}$  je familija svih njegovih podskupova. Uobičajena oznaka je  $\mathcal{P}(\mathcal{A})$ .

Rastav skupa  $\mathcal{A}$  na disjunktne podskupove  $\pi_1, \dots, \pi_k$ ,  $1 \leq k \leq m$ , takve da vrijedi

$$\bigcup_{i=1}^k \pi_i = \mathcal{A}, \quad \pi_i \cap \pi_j = \emptyset, \quad i \neq j, \quad m_j := |\pi_j| \geq 1, \quad j = 1, \dots, k, \quad (33)$$

označit ćemo s  $\Pi(\mathcal{A}) = \{\pi_1, \dots, \pi_k\}$  i zvat ćemo *particijom skupa  $\mathcal{A}$* . Skupove  $\pi_1, \dots, \pi_k$  zvat ćemo *klasteri*. Skup svih particija skupa  $\mathcal{A}$  sastavljenih od  $k$  klastera koje zadovoljavaju (33) označit ćemo s  $\mathcal{P}(\mathcal{A}, k)$ .

**Zadatak 11.** *Općenito, broj elemenata partitivnog skupa, skupa  $\mathcal{A}$  od  $m$  elemenata jednak je  $2^m$ . Pokažite da je broj svih particija skupa  $\mathcal{A}$ , koji zadovoljavaju (33) jednak*

$$|\mathcal{P}(\mathcal{A}, k)| = \frac{1}{k!} \sum_{j=1}^k (-1)^{k-j} \binom{k}{j} j^m. \quad (34)$$

*Izračunajte broj svih particija skupa  $\mathcal{A}$  određenih s (33) 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           |
| $m = 50$       | $10^{15}$       | $10^{23}$       | $10^{33}$       | $10^{40}$       | $10^{43}$   |
| $m = 10^3$     | $10^{300}$      | $10^{476}$      | $10^{697}$      | $10^{898}$      | $10^{993}$  |
| $m = 10^6$     | $10^{301\,029}$ | $10^{477\,120}$ | $10^{698\,968}$ | $10^{903\,085}$ | $10^{10^6}$ |

Na skupu svih particija  $\mathcal{P}(\mathcal{A}, k)$  skupa  $\mathcal{A}$  sastavljenih od  $k$  klastera definiramo kriterijsku funkciju cilja

$$F : \mathcal{P}(\mathcal{A}, k) \rightarrow \mathbb{R}_+, \quad F(\Pi) = \sum_{j=1}^k f(\pi_j), \quad \Pi = \{\pi_1, \dots, \pi_k\}. \quad (35)$$

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

$$F(\Pi^{(o)}) = \min_{\Pi \in \mathcal{P}(\mathcal{A}, k)} F(\Pi), \quad (36)$$

gdje je funkcija cilja  $F$  definirana s (35), pri čemu je  $f$  definirana s

$$c_j = \operatorname{argmin}_{\alpha} \sum_{a_i \in \pi_j} (a_i - \alpha)^2 = \frac{1}{|\pi_j|} \sum_{a_i \in \pi_j} a_i, \quad f(\pi_j) = \sum_{a_i \in \pi_j} (a_i - c_j)^2. \quad (37)$$

---

<sup>1</sup>Varianzkriterium, Least Squares

## Algoritam 2. Standardni $k$ -means algoritam

**Step 1:** Učitati  $m, k$ , skup  $\mathcal{A} = \{a_1, \dots, a_m\}$  i izabrati početnu pariciju  $\Pi = \{\pi_1, \dots, \pi_k\}$   
Izračunati centre  $c_1, \dots, c_k$  klastera  $\pi_1, \dots, \pi_k$  i početnu vrijednost funkcije cilja  
 $F_0 = F(\Pi)$ ;

**Step 2:** Formirati novu particiju  $\mathcal{V} = \{\nu_1, \dots, \nu_k\}$  s novim klasterima definiranim s

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

njihove centroide  $\zeta_1, \dots, \zeta_k$  i novu vrijednost funkcije cilja  $F_1 = F(\mathcal{V})$ ;

**Step 3:** Ako je  $F_1 < F_0$ , stavi  $c_j = \zeta_j$ ;  $j = 1, \dots, k$ ;  $F_0 = F_1$  i priredi na **Step 2**;  
U protivnom STOP.

**Primjer 4.** *Primijenite standardni  $k$ -means algorithm uz Least Squares kriterij,  $\mathcal{A} = \{-7, -5, -4, 4, 5, 7\}$ ,  $k = 3$*

| R.br. | Klasteri         |             |               | Centroidi       |   |                | Funkcija cilja             |
|-------|------------------|-------------|---------------|-----------------|---|----------------|----------------------------|
| 1     | $\{-7, -5\}$     | $\{-4, 4\}$ | $\{5, 7\}$    | -6              | 0 | 6              | 36                         |
| 2     | $\{-7, -5, -4\}$ | $\emptyset$ | $\{4, 5, 7\}$ | $-\frac{16}{3}$ |   | $\frac{16}{3}$ | $\frac{188}{9} \approx 20$ |

Ovaj primjer pokazuje da standardni  $k$ -means algoritam može smanjiti broj klastera. Primijenite  $l_1$ -kriterij u ovom slučaju.

**Zadatak 12.** *Zadan je skup  $\mathcal{A} = \{a_i \in \mathbb{R} : i = 1, \dots, m\}$ . Definirajte princip maksimalnih kvadratne udaljenosti centara klastera u cilju dobivanja optimalne particije  $\Pi^o$  s klasterima  $= \pi_1^o, \dots, \pi_k^o$ , koji zadovoljavaju (33).*

**Primjer 5.** *Vektor kome svaka komponenta predstavlja broj bodova nekog studenta postignutom na kolokviju treba razdijeliti u 4 klastera (ocjena 5, ocjena 4, ocjena 3 i ocjena 2).*

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

Primjenom *Mathematica*-naredbe `FindClusters[ $\mathcal{A}$ , 4]` dobivamo

$\pi_1(5) : \{418, 385, 357, 351, 347, 333, 330, 323, 321, 318\}$   
 $\pi_2(4) : \{308, 308, 302, 302, 300, 294, 293, 290, 287, 286, 286, 277, 271, 271, 269, 268, 266\}$   
 $\pi_3(3) : \{260, 253, 250, 248, 237, 237, 233, 232, 230, 228, 227, 226, 217, 214\}$   
 $\pi_4(2) : \{211, 208, 207, 187, 186, 181, 179, 175, 165, 161\}$

### 3.2 Klasteri u $\mathbb{R}$ : Least Absolute Deviations princip

Neka je  $\mathcal{A} = \{a_1, \dots, a_m\} \subset \mathbb{R}$  skup realnih brojeva. Rastav skupa  $\mathcal{A}$  na disjunktne podskupove  $\pi_1, \dots, \pi_k$ ,  $1 \leq k \leq m$ , takve da vrijedi (33)

$$\bigcup_{i=1}^k \pi_i = \mathcal{A}, \quad \pi_i \cap \pi_j = \emptyset, \quad i \neq j, \quad m_j := |\pi_j| \geq 1, \quad j = 1, \dots, k,$$

označit ćemo s  $\Pi(\mathcal{A}) = \{\pi_1, \dots, \pi_k\}$  i zvat ćemo *particiju skupa*  $\mathcal{A}$ . Skupove  $\pi_1, \dots, \pi_k$  zvat ćemo *klasteri*. Skup svih particija skupa  $\mathcal{A}$  sastavljenih od  $k$  klastera koje zadovoljavaju (33) označit ćemo s  $\mathcal{P}(\mathcal{A}, k)$ .

**Definicija 2.** Kažemo da je particija  $\Pi^{(o)}$  optimalna u smislu  $l_1$  norme<sup>2</sup> onda ako vrijedi (36)

$$F(\Pi^{(o)}) = \min_{\Pi \in \mathcal{P}(\mathcal{A}, k)} F(\Pi),$$

gdje je funkcija cilja  $F$  definirana s (35), pri čemu je  $f$  definirana s

$$c_j = \operatorname{argmin}_{\alpha} \sum_{a_i \in \pi_j} |a_i - \alpha| = \operatorname{med}_{a_i \in \pi_j} a_i, \quad f(\pi_j) = \sum_{a_i \in \pi_j} |a_i - \operatorname{med}_{a_i \in \pi_j} a_i|. \quad (38)$$

**Zadatak 13.** Pronađite sve particije skupa  $\mathcal{A} = \{1, 3, 4, 8\}$ , koje zadovoljavaju (33) za  $k = 2$ , odrediti pripadne centre i vrijednosti funkcije cilja  $F$  u smislu  $l_1$  norme.

**Zadatak 14.** Na podatke iz Zadatka 13 primijenite standardni  $k$ -means algoritam za slučaj kada je kriterijska funkcija cilja zadana s (36), pri čemu je funkcija  $f$  definirana s (38). Početnu particiju  $\Pi^0$  izaberite proizvoljno. Može li se i u ovom slučaju dogoditi da  $k$ -means algoritam ne daje optimalno rješenje?

**Zadatak 15.** Neka je  $\mathcal{A} = \{a_1, \dots, a_m\} \subset \mathbb{R}$ . Koristeći LAD-kriterij optimalnosti, skup  $\mathcal{A}$  treba rastaviti na disjunktne podskupove  $\pi_1, \pi_2$ , takve da vrijedi

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

Može li se to uraditi analogno kao u t.3.4.1, korištenjem principa maksimalne  $l_1$  udaljenosti centara klastera?

### 3.3 Klasteri u $\mathbb{R}^n$ : Least Squares princip

Neka je  $\mathcal{A} = \{a^i = (a_1^{(i)}, \dots, a_n^{(i)})^T \in \mathbb{R}^n : i = 1, \dots, m\}$  skup vektora iz vektorskog prostora  $\mathbb{R}^n$ . Rastav skupa  $\mathcal{A}$  na disjunktne podskupove  $\pi_1, \dots, \pi_k$ ,  $1 \leq k \leq m$ , takve da vrijedi

$$\bigcup_{i=1}^k \pi_i = \mathcal{A}, \quad \pi_i \cap \pi_j = \emptyset, \quad i \neq j, \quad m_j := |\pi_j| \geq 1, \quad j = 1, \dots, k, \quad (39)$$

---

<sup>2</sup>Least Absolute Deviations – najmanja apsolutna odstupanja.

označit ćemo s  $\Pi(\mathcal{A}) = \{\pi_1, \dots, \pi_k\}$  i zvat ćemo *particijom skupa*  $\mathcal{A}$ . Skupove  $\pi_1, \dots, \pi_k$  zvat ćemo *klasteri*. Skup svih particija skupa  $\mathcal{A}$  sastavljenih od  $k$  klastera koje zadovoljavaju (33) označit ćemo s  $\mathcal{P}(\mathcal{A}, k)$ .

Na skupu svih particija  $\mathcal{P}(\mathcal{A}, k)$  skupa  $\mathcal{A}$  sastavljenih od  $k$  klastera potpuno analogno definiramo kriterijsku funkciju cilja i optimalnu particiju.

**Definicija 3.** Kažemo da je particija  $\Pi^{(o)} = \{\pi_1^o, \dots, \pi_k^o\}$  optimalna u smislu kvadrata  $l_2$  norme ili u smislu najmanjih kvadrata onda ako vrijedi

$$F(\Pi^{(o)}) = \min_{\Pi \in \mathcal{P}(\mathcal{A}, k)} F(\Pi), \quad (40)$$

gdje je funkcija cilja  $F: \mathcal{P}(\mathcal{A}, k) \rightarrow \mathbb{R}_+$  definirana s

$$F(\Pi) = \sum_{j=1}^k f(\pi_j), \quad \Pi = \{\pi_1, \dots, \pi_k\} \quad (41)$$

$$c_j = \operatorname{argmin}_{\alpha \in \mathbb{R}^n} \sum_{a^i \in \pi_j} \|a^i - \alpha\|_2^2 = \frac{1}{|\pi_j|} \sum_{a^i \in \pi_j} a^i, \quad f(\pi_j) = \sum_{a^i \in \pi_j} \|a^i - c_j\|_2^2. \quad (42)$$

### 3.4 Slučaj $n = k = 2$ : Least Squares princip

Neka je  $\mathcal{A} = \{a^i = (x_i, y_i)^T \in \mathbb{R}^2 : i = 1, \dots, m\}$  skup točaka (vektora) iz  $\mathbb{R}^2$ . Rastav skupa  $\mathcal{A}$  na disjunktne neprazne podskupove  $\pi_1, \pi_2$ , takve da vrijedi (39) zovemo *particijom skupa*  $\mathcal{A}$  i označavamo s  $\Pi(\mathcal{A}) = \{\pi_1, \pi_2\}$ . Skupove  $\pi_1, \pi_2$  zvat ćemo *klasteri u ravni*. Skup svih particija skupa  $\mathcal{A}$  sastavljenih od dva klastera koje zadovoljavaju (39) označit ćemo s  $\mathcal{P}(\mathcal{A}, 2)$ .

Broj svih dvočlanih particija skupa  $\mathcal{A}$  zadanih s (39) također je zadan s (34).

U slučaju LS kriterija centriodi  $c_1, c_2 \in \mathbb{R}^2$  klastera  $\pi_1, \pi_2$  zadani su formulama

$$c_1 = \frac{1}{|\pi_1|} \sum_{a^i \in \pi_1} a^i, \quad c_2 = \frac{1}{|\pi_2|} \sum_{a^i \in \pi_2} a^i,$$

(aritmetičke sredine svih točaka  $a^i \in \pi_j$  koje pripadaju klasteru  $\pi_j$ ,  $j = 1, 2$ ).

**Zadatak 16.** Čemu je jednako

$$(a) \min_{x \in \mathbb{R}^2} \sum_{a^i \in \pi_1} d_1(a^i, x) \quad (b) \min_{x \in \mathbb{R}^2} \sum_{a^i \in \pi_1} d_\infty(a^i, x) ?$$

i za koji  $x \in \mathbb{R}^2$  se postiže ?

Kriterijska funkciju cilja  $F: \mathcal{P}(\mathcal{A}, 2) \rightarrow \mathbb{R}_+$  zadana je s

$$F(\Pi) = \sum_{a^i \in \pi_1} \|a^i - c_1\|^2 + \sum_{a^i \in \pi_2} \|a^i - c_2\|^2, \quad \Pi = \{\pi_1, \pi_2\}.$$

**Primjer 6.** Koristeći princip najmanjih kvadrata, za skup točaka (vektora) u ravnini

$$\mathcal{A} = \{a_1 = (-1, -1), a_2 = (1, -1), a_3 = (1, 1), a_4 = (-1, 1)\},$$

treba odrediti sve dvočlane particije, koje zadovoljavaju (33), odgovarajuće centroide i vrijednosti funkcije cilja.

$$\begin{aligned} \Pi_1 &= \{\{a_1\}, \{a_2, a_3, a_4\}\}, & c_1 &= a_1, & c_2 &= \left(\frac{1}{3}, \frac{1}{3}\right), & F(\Pi_1) &= 0 + \frac{16}{3} \approx 5.3 \\ \Pi_2 &= \{\{a_2\}, \{a_1, a_3, a_4\}\}, & c_1 &= a_2, & c_2 &= \left(-\frac{1}{3}, \frac{1}{3}\right), & F(\Pi_2) &= 0 + \frac{16}{3} \approx 5.3 \\ \Pi_3 &= \{\{a_3\}, \{a_1, a_2, a_4\}\}, & c_1 &= a_3, & c_2 &= \left(-\frac{1}{3}, -\frac{1}{3}\right), & F(\Pi_3) &= 0 + \frac{16}{3} \approx 5.3 \\ \Pi_4 &= \{\{a_4\}, \{a_1, a_2, a_3\}\}, & c_1 &= a_4, & c_2 &= \left(\frac{1}{3}, -\frac{1}{3}\right), & F(\Pi_4) &= 0 + \frac{16}{3} \approx 5.3 \\ \Pi_5 &= \{\{a_1, a_2\}, \{a_3, a_4\}\}, & c_1 &= (0, -1), & c_2 &= (0, 1), & F(\Pi_5) &= 2 + 2 = 4 \\ \Pi_6 &= \{\{a_1, a_3\}, \{a_2, a_4\}\}, & c_1 &= (0, 0), & c_2 &= (0, 0), & F(\Pi_6) &= 4 + 4 = 8 \\ \Pi_7 &= \{\{a_1, a_4\}, \{a_2, a_3\}\}, & c_1 &= (-1, 0), & c_2 &= (1, 0), & F(\Pi_7) &= 2 + 2 = 4 \end{aligned}$$

Dakle, postoje dvije optimalne particije:  $\Pi_5, \Pi_7$

**Zadatak 17.** Odredite sve dvočlane particije, koje zadovoljavaju (33), odgovarajuće centre i vrijednosti funkcije cilja za točke iz Primjera 6 uz primjenu  $l_1$ , odnosno  $l_\infty$  norme.

**Zadatak 18.** Koliko ima tročlanih particija osnovnog skupa iz Primjera 6 koje zadovoljavaju (33)? Ispišite ih, izračunajte odgovarajuće centre i vrijednosti funkcije cilja uz primjenu  $l_2$ ,  $l_1$  i  $l_\infty$  norme.

**Zadatak 19.** Na podatke iz Primjera 6 i početnu particiju  $\Pi_1$  primijenite standardni  $k$ -means algorithm.

**Zadatak 20.** Napišite odgovarajući  $k$ -means algorithm ( $1 \leq k \leq m$ ).

### 3.4.1 Princip maksimalne kvadratne udaljenosti centara klastera

Neka je  $\mathcal{A} = \{a^i = (x_i, y_i)^T \in \mathbb{R}^2 : i = 1, \dots, m\}$  skup točaka (vektora) iz  $\mathbb{R}^2$  i neka su  $c_1, c_2$  centri klastera  $\pi_1, \pi_2$ . Tada za svaki  $x \in \mathbb{R}^2$  vrijedi

$$\sum_{a^i \in \pi_j} \|a^i - x\|^2 = \sum_{a^i \in \pi_j} \|a^i - c_j\|^2 + m_j \|c_j - x\|^2, \quad j = 1, 2. \quad (43)$$

$$\begin{aligned} \sum_{a^i \in \pi_1} \|a^i - x\|^2 &= \sum_{a^i \in \pi_1} \|(a^i - c_1) + (c_1 - x)\|^2 \\ &= \sum_{a^i \in \pi_1} \|a^i - c_1\|^2 + m_1 \|c_1 - x\|^2, \end{aligned}$$

jer je

$$\sum_{a^i \in \pi_1} (c_1 - x)^T (a^i - c_1) = (c_1 - x)^T \sum_{a^i \in \pi_1} (a^i - c_1) = 0.$$

Ako u (43) umjesto  $x$  stavimo  $c = \frac{1}{m} \sum_{i=1}^m a^i$  i zbrojimo obje jednakosti, dobivamo

$$\sum_{a^i \in \pi_1} \|a^i - c\|^2 + \sum_{a^i \in \pi_2} \|a^i - c\|^2 = \left( \sum_{a^i \in \pi_1} \|a^i - c_1\|^2 + \sum_{a^i \in \pi_2} \|a^i - c_2\|^2 \right) + m_1 \|c_1 - c\|^2 + m_2 \|c_2 - c\|^2$$

odnosno

$$\begin{aligned} \sum_{i=1}^m \|a^i - c\|^2 &= F(\Pi) + G(\Pi), \\ G(\Pi) &= m_1 \|c_1 - c\|^2 + m_2 \|c_2 - c\|^2. \end{aligned} \tag{44}$$

**Teorem 2.** *Uz prethodne oznake vrijedi:*

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

To znači da u cilju pronalaženja optimalne LS particije, umjesto minimizacije funkcije (28) možemo maksimizirati funkciju (vidi Primjer 1)

$$G(\Pi) = m_1 \|c_1 - c\|^2 + m_2 \|c_2 - c\|^2 \tag{45}$$

**Zadatak 21.** *Za podatke iz Primjera 6 odredite LS-optimalnu particiju prema principu maksimalne kvadratne udaljenosti centara klastera.*

**Zadatak 22.** *Neka je  $\mathcal{A} = \{a^i = (a_1^{(i)}, \dots, a_n^{(i)})^T \in \mathbb{R}^n : i = 1, \dots, m\}$  skup vektora iz vektorskog prostora  $\mathbb{R}^n$ . Definirajte LS-optimalnu particiju  $\Pi^o = \{\pi_1^o, \dots, \pi_k^o\}$ ,  $1 \leq k \leq m$  prema principu maksimalne kvadratne udaljenosti centara klastera.*

## Literatura

- A. Ben-Israel, C. Iyigun, *Probabilistic D-clustering*, Journal of Classification **25**(2007) DOI: 10.1007/s00357-007-0021-y
- D. L. Boley, *Principal direction divisive partitioning*, Data Mining and Knowledge Discovery 2(1998), 325–344
- D. L. Boyd, L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, 2004.
- 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
- I. S. Dhillon, S. Mallela, R. Kumar, *A divisive information-theoretic feature clustering algorithm for text classification*, Journal of Machine Learning Research **3**(2003), 1265–1287.
- 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

- E. Domínguez, J. Muñoz, *Applying bio-inspired techniques to the  $p$ -median problem*, IWANN 2005; Computational Intelligence Bioinspired Syst., 8th Int. Workshop Artificial Neural Networks, Lecture Notes in Computer Science, Springer-Verlag, Berlin, 2005, pp. 67 – 74
- G. Gan, C Ma, J. Wu, *Data Clustering: Theory, Algorithms, and Applications*, SIAM, Philadelphia, 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.
- C. Iyigun, A. Ben-Israel, *A generalized Weiszfeld method for the multi-facility location problem*, Operations Research Letters **38**(2010) 207–214
- D. Littau, D. L. Boley, *Clustering very large data sets with PDDP*. In J. Kogan, C. Nicholas, M. Teboulle (eds), *Grouping Multidimensional Data: Recent Advances in Clustering*, 99–126, Springer-Verlag, New York, 2006.
- S. Pan, J., S. Chen, *Two unconstrained optimization approaches for the Euclidean  $k$ -centrum location problem*, Applied Mathematics and Computation **189**(2007) 1368–1383
- J. Reese, *Solution methods for the  $p$ -median problem: an annotated bibliography*, Published online in Wiley InterScience, Wiley, 2006.
- A. M. Rodríguez-Chia, I. Espejo, Z. Drezner, *On solving the planar  $k$ -centrum problem with Euclidean distances*, EJOR, to appear
- K. Sabo, R. Scitovski, *The best least absolute deviations line – properties and two efficient methods*, ANZIAM Journal **50**(2008), 185–198
- A. Schöbel, D. Scholz, *The big cube small cube solution method for multidimensional facility location problems*, Computers & Operations Research **37**(2010), 115–122
- H. Späth, *Cluster-Formation und Analyse*, R. Oldenburg Verlag, München, 1983.
- Z. Su, J. Kogan, C. Nicholas, *Constrained clustering with  $k$ -means type algorithms*, In M.W. Berry, J. Kogan (eds), *Text Mining Applications and Theory*, 81–103, Willey, Chichester, 2010.
- M. Teboulle, *A unified continuous optimization framework for center-based clustering methods*, Journal of Machine Learning Research **8**(2007), 65–102