

2022-05-17

ML pred 11.

## Nenadzirano učenje

### Klasteriranje

Tvrdlo Klasteriranje - svaki podatok mora pripasti točno 1 klasteru

Meso Klasteriranje - svaki podatok ima vjerojatnost spajanja pripada

- metode za Klasteriranje: klasickijske metode (racemo na biloj) particijalne metode

Def Neka je  $X = \{x_i \in \mathbb{Q}^n : i=1, \dots, m\}$ ,  $m \geq 2$  sadar skup podataka.

Rastav skupa  $X$  na  $1 \leq s \leq m$  disjunktivnih skupova

$$\bar{u}_1, \dots, \bar{u}_s \text{ rastav } X \quad \bigcup_{j=1}^s \bar{u}_j = X, \quad \bar{u}_i \cap \bar{u}_j = \emptyset, \quad i \neq j, \quad |\bar{u}_j| \geq 1, \quad j=1, \dots, s$$

Zovemo  $s$ -particija skupa  $X$  i označavamo s

$$\Pi_s = \{\bar{u}_1, \dots, \bar{u}_s\}.$$

Elementi od  $\Pi_s$  zovemo klasteri, a skup svih particija skupax nastaljivih sa  $s$  klasterima označavamo

$$s P(X; s)$$

Pr

$$X = \{1, 2, 3\}$$

- želimo  $P(X; 2)$  - sve 2 particije

$$\{\{1\}, \{2, 3\}\} = \Pi_2^{(1)}$$

$$\{\{2\}, \{1, 3\}\} = \Pi_2^{(2)}$$

$$\{\{3\}, \{1, 2\}\} = \Pi_2^{(3)}$$

$$\Rightarrow P(X; 2) = \{\Pi_2^{(1)}, \Pi_2^{(2)}, \Pi_2^{(3)}\}$$

→ za  $X = \{1, 2, \dots, 10\}$   $2 \cdot 2$  je točko nacrtano

- broj  $|P(X; 2)|$  je povezan s brojem surjektija  
s m-članog skupu u 2-članu skup

$$|P(X; 2)| = \frac{1}{2!} \sum_{j=0}^2 (-1)^{2-j} \binom{2}{j} j^m \rightarrow \text{Stirlingov broj}$$

2. vrste

Pr.  $m=5$

$2=2$

$$|P(X; 2)| = 15$$

$m=50$

$2=2$

$$|P(X; 2)| = 10^{15}$$

$m=50$

$2=10$

$$|P(X; 10)| = 10^{44}$$

$\rightarrow$  za glasteriranje ne dolazi u oblik plesati sve partice

### Kriterijska funkcija

Def. Čiji  $d: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ , za svojstvo:

i)  $d(x, y) = 0 \Leftrightarrow x = y$

ii)  $x \mapsto d(x, y)$  je neprekidna + fiksni  $y$

iii)  $\lim_{\|x-y\| \rightarrow \infty} d(x, y) = \infty$ , + fiksni  $y$

zovemo braximetricka (distance-like) funkcija

- razliku od metričke - nije simetrična, nemas nejednakosti

Trotula

- svaka metrika je braximetricka

Pr. Braximetricke loge nije metrika

$$d: \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R},$$

$$d(x, y) = (x - y)^2 \rightarrow \text{jest simetrična, ali ne vrijedi nejednakost}$$

Trotula

$\rightarrow$  vrijede svojstva braximetricke

$\rightarrow$  nejednakost trokuta  $d(x, y) + d(y, z) \geq d(x, z)$  ne vrijedi

$$\text{za } x=1, y=2, z=4: 1^2 + 2^2 \geq 3^2$$

Primjeri 1)  $d_{L^2} : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}_+$ ,  $d_{L^2}(x, y) = \|x - y\|_2^2 \rightarrow$  Euklidska ali nije metrika

2)  $d_1(x, y) = \|x - y\|_1 = \sum_{i=1}^n |x_i - y_i| \rightarrow$  metrika  $\Rightarrow$  L1 metrika

3)  $d_p(x, y) = \|x - y\|_p^n, p \geq 1 \rightarrow$  metrika  $\Rightarrow$  Lp metrika

4)  $d_M(x, y) = (x - y)^T S^{-1} (x - y) \rightarrow$  Mahalanobis

$S$ - poz. semidefinitna Euklidska

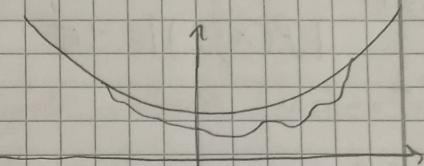
Na geom. predstavlja 2. svojstvo

$x \mapsto d(x, y)$ ,  $y$  - fiksan

$\rightarrow$  ii + iii svojstva

$x \mapsto d(x, y)$  uvijek postoji

globalni minimum



$\hookrightarrow$  nije vjerojatno da je ljevo

$\rightarrow$  nizovi idu u  $\mathbb{R}$  i ne prelaze

Def Neka je  $d : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}_+$  Euklidska f.

Najbolji reprezentant skupine  $\{x_i \in \mathbb{R}^n : i = 1, \dots, m\}$

je vektor

$$\mu_j^* \in \arg \min \sum_{i=1}^{m_j} d(\mu, x_i), \quad \mu \in \mathbb{R}^n$$

$\rightarrow$  Šešimo točke koja je po Euklidsici najmanje udaljena

$$F(\mu) = \sum_{i=1}^{m_j} d(\mu, x_i) \rightarrow \min_{\mu}$$

$\rightarrow \mu_j^*$  je element jer nije jedinstven

Primjedba Zbroj svojstva ii i iii Euklidske f., reprezentant je slobo definiran

Primerjet:  $\bar{w}_j = \{x_i \in \mathbb{R}^n : i=1, \dots, m_j\}$

$$1) d_{LS}(x, y) = \|x - y\|_2^2$$

$$F(\mu) = \sum_{i=1}^{m_j} \|\mu - x_i\|_2^2 \rightarrow \min_{\mu}$$

$$\nabla F(\mu) = 2 \sum_{i=1}^{m_j} (\mu - x_i) = 0$$

$$\Rightarrow m_j \mu = \sum_{i=1}^{m_j} x_i$$

$$\mu^* = \frac{1}{m_j} \sum_{i=1}^{m_j} x_i$$

centroal (triste)

nodalala

$$F(\mu_1, \mu_2) = \sum_{i=1}^m (u_i - \mu_1)^2 + (v_i - \mu_2)^2$$

$$x_i = (u_i, v_i)$$

$$\nabla F(\mu_1, \mu_2) = \begin{bmatrix} -2 \sum_{i=1}^m (u_i - \mu_1) \\ -2 \sum_{i=1}^m (v_i - \mu_2) \end{bmatrix}$$

$$= 2 \begin{bmatrix} \sum (u_i - \mu_1) \\ \sum (v_i - \mu_2) \end{bmatrix}$$

$$= 2 \begin{bmatrix} \mu_1 - u \\ \mu_2 - v \end{bmatrix}$$

$$2) d_e(x, y) = \|x - y\|_1$$

$$F(\mu) = \sum_{i=1}^{m_j} \|\mu - x_i\|_1 \rightarrow \min_{\mu}$$

$$\mu^* = \text{medijan } (\bar{w}_j)$$

$\rightarrow$  se aseamna cu medijan

medijan se urima pe elemente

po si doborina  
reprezentare

$$F(\mu) = \sum_{i=1}^{m_j} |\mu - x_i| \rightarrow \min_{\mu}$$

$\rightarrow$  moza se polarizati de  
si min nu medijanu.

$$\bar{x} \in \bar{w}_j = \{1, 3, 7\}$$

$$\text{medijan } (\bar{w}_j) = 3$$

$$\bar{w}_j = \{1, 3, 5, 7\}$$

$$\text{medijan } (\bar{w}_j) \in [3, 5]$$

Kriterijus bje:

$$\mathcal{F}: \mathcal{P}(X; \Omega) \rightarrow \mathbb{R}_+$$

$$\mathcal{F}(\Pi) = \sum_{j=1}^k \sum_{x_i \in \bar{w}_j} d(\mu_j, x_i), \quad \mu_j - \text{reprezentant klastera } \bar{w}_j$$

Def Zs particijs  $\Pi^* \in \mathcal{P}(X; \Omega)$  sačemo da je globalna  
optimalna ako je  $\mathcal{F}(\Pi^*) = \min_{\Pi \in \mathcal{P}(X; \Omega)} \mathcal{F}(\Pi)$

Za particiju  $\Pi_8 = \{\bar{c}_1, \dots, \bar{c}_8\}$  mamo određeni centri  $\mu_1, \dots, \mu_8$

$$\bar{c}_{ij} = \{x_i : d(\mu_j, x_i) \leq d(\mu_s, x_i), s \neq j, s=1, \dots, 8\}, i=1, \dots, 2$$

↳ Sastavljamo blaster  $\rightarrow$  princip minimalnih udaljenosti

$\rightarrow$  algoritam: od particije pravimo centra

( $k$ -means) od centara pravimo blaster i time particije  
repeat

$$F: \mathbb{R}^{n_2} \rightarrow \mathbb{R}_+$$

↳ Sastavljeni su centri

$$F(\mu_1, \dots, \mu_8) = \sum_{i=1}^m \min_{j=1, \dots, 8} d(\mu_j, x_i)$$

rijek diferencijabilna  
kontinuirana

$\rightarrow$  umjetna minimizacija  $F$  možemo  $F(\mu_1, \dots, \mu_8) \rightarrow \min_{\mu_1, \dots, \mu_8}$

Def Za reprezentante  $\mu_1^*, \dots, \mu_8^*$  sazemo da su globalno optimalni ako je

$$F(\mu_1^*, \dots, \mu_8^*) = \min_{\mu_1, \dots, \mu_8} F(\mu_1, \dots, \mu_8)$$

$$F(\mu_1^*, \dots, \mu_8^*) = F(\Pi^*)$$

Tri pristupa za određivanje  $\Pi^*$ :

- globalna minimizacija od  $F \rightarrow$  tešak problem

- lokalna minimizacija ( $k$ -means algoritam)

- implementacione metode (global  $k$ -means)

$\rightarrow$  možete čitati  $k$ -means

- problem je što postoji  $\mathcal{S}$ : točka globalnog minimuma

- još više lokalnih  $\rightarrow$  nujno sada će dovesti do globalnog minimuma

## 2-means algoritam

1. Odrediti početne reprezentante (assignment step)

$$\mu_j, j=1, \dots, 2$$

2. Po principu minimalnih udaljenosti određimo klasterne

$$\bar{w}_j = \{x_i \in X : d(\mu_j, x_i) \leq d(\mu_s, x_i), \forall s, s \neq j, s=1, \dots, 2\}$$

3. Definiramo nove reprezentante (update step)

$$\mu_j \in \arg \min_{\mu} \sum_{x_i \in \bar{w}_j} d(\mu, x_i), j=1, \dots, 2$$

4. Ponavljaj 2 i 3 dok se reprezentanti ili participo ne podudaraju

- može doći do pada broja klastera ako se neli podatci odabire bazu jedini element klastera

→ slučaj se može odbaciti - može se polarizati da će se tako što minimizirati bolje sa vecim brojem klastera

- 2-means može klasterinaci lin. separabilne podatke

