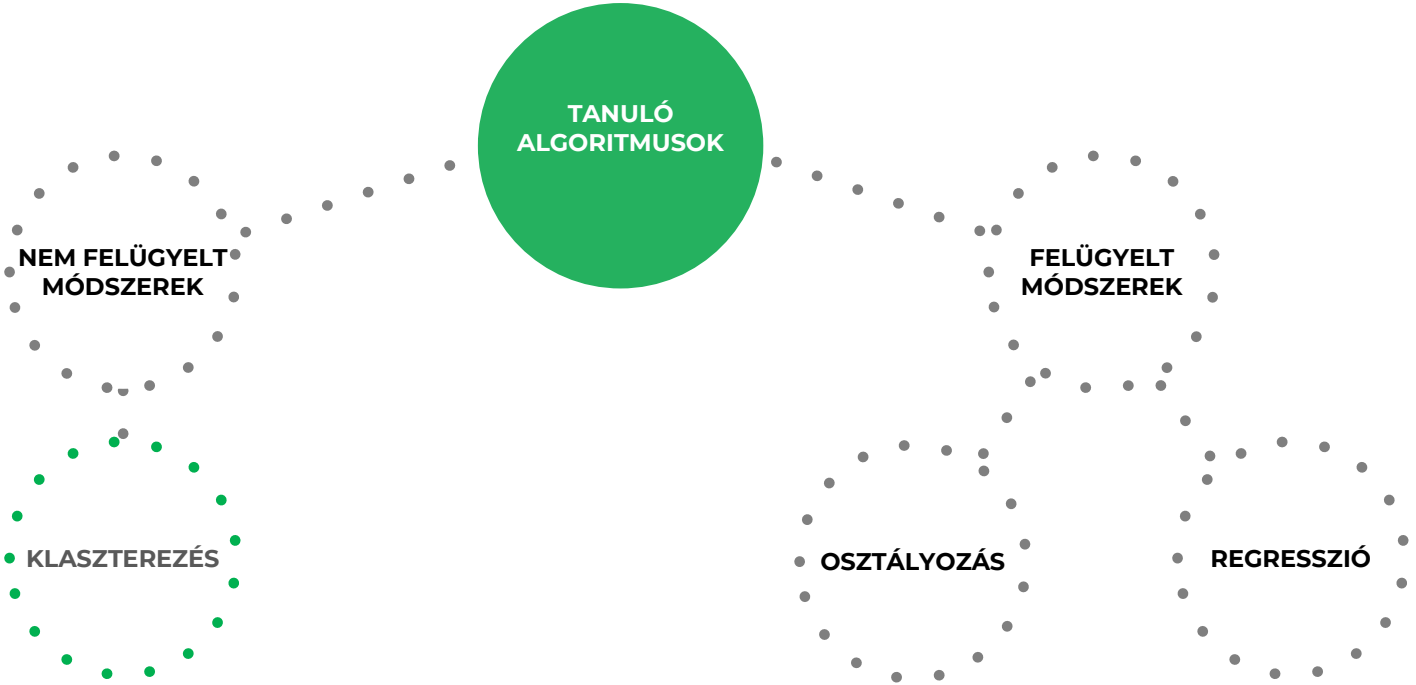
Vásárlási szokások

Klaszterezés

2023.10.09.

Huszti Dorottya
data scientist

MODELLEZÉSI ALAPOK



Feladat

Van egy tanító adathalmazom
... ahol ismerem a tulajdonságokat

Kérdés, hogy milyen csoportokba rakható úgy, hogy
... csoporton belül hasonlóak
.... csoportok ne hasonlítsanak egymásra

Klaszterezés
Csoportképzés

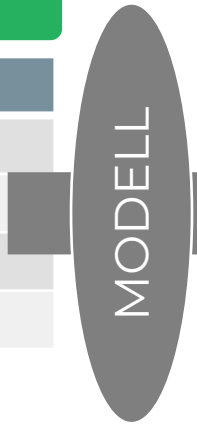
Kérdések

- **Hány csoportba soroljuk az adatokat?**
- **Hogyan mérjük a hasonlóságot?**
- **Honnan tudjuk, hogy mennyire jó a modell?**

ADATBÁNYÁSZATI ALAPPROBLÉMÁK

Tulajdonságleíró változók – input attribútumok

Életkor	Jövedelem	Lakásméret	Gyerekek	Autók száma
42	280 000	82	3	1
88	90 000	44	0	0
22	180 000	32	0	1
38	400 000	102	1	2



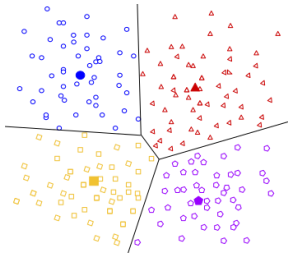
Csoport
A. klaszter
A. klaszter
B. klaszter
B. klaszter

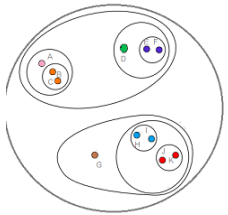
Klaszterező algoritmusok

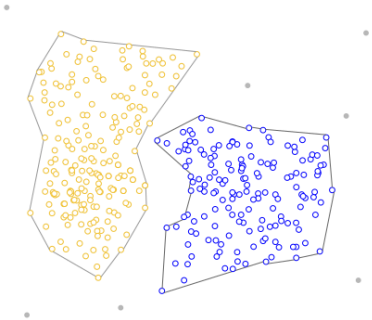
ALGORITMUSOK FAJTÁI

Távolság alapú
K-Means, K-Medoids

Sűrűség alapú
DBSCAN, Mean-shift clustering







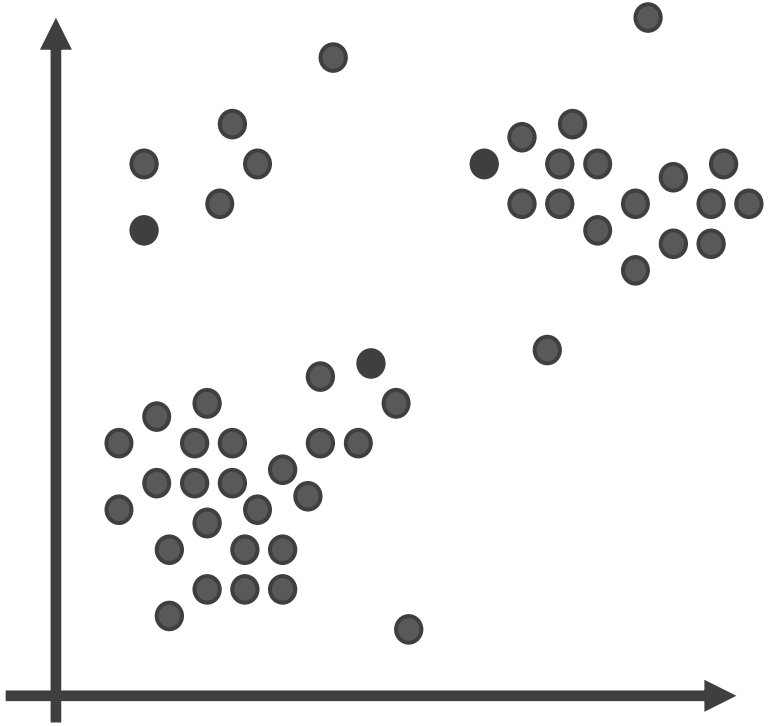
Partícionáló

Hierarchikus

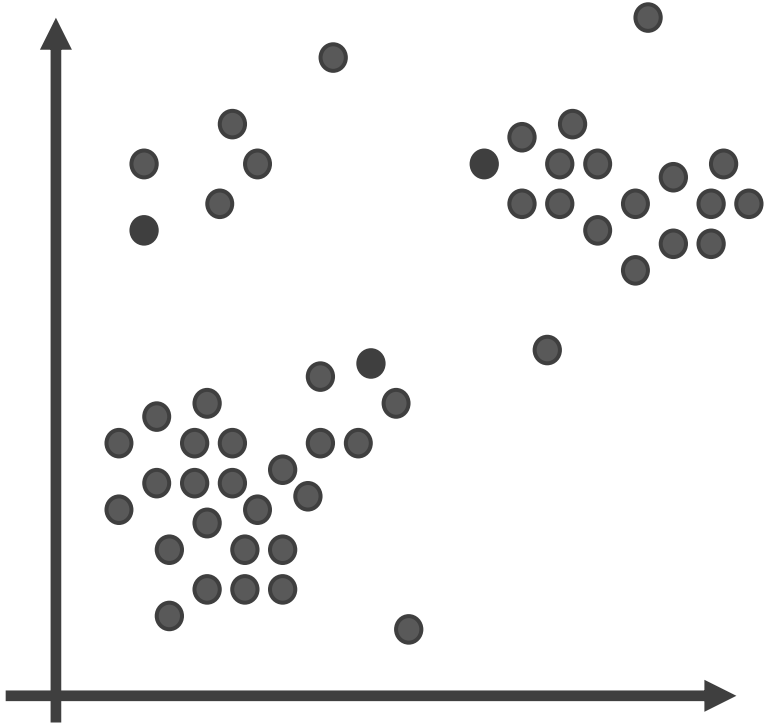
K-közép (K-means)

Klaszterező
algoritmusok

K-KÖZÉP K-MEANS

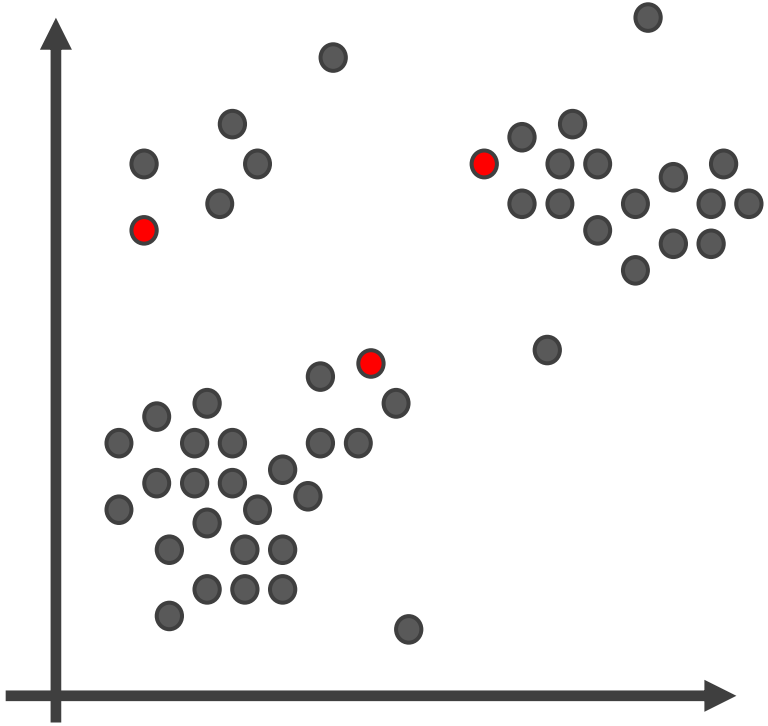


K-KÖZÉP K-MEANS



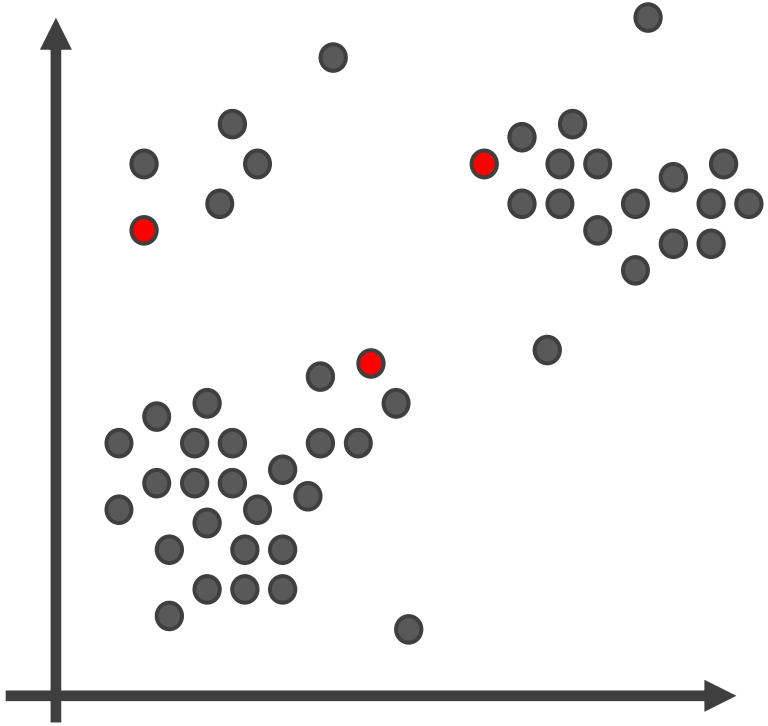
1. Klaszterközéppontok (centroid) véletlenszerű kiválasztása

K-KÖZÉP K-MEANS



1. Klaszterközéppontok (centroid) véletlenszerű kiválasztása

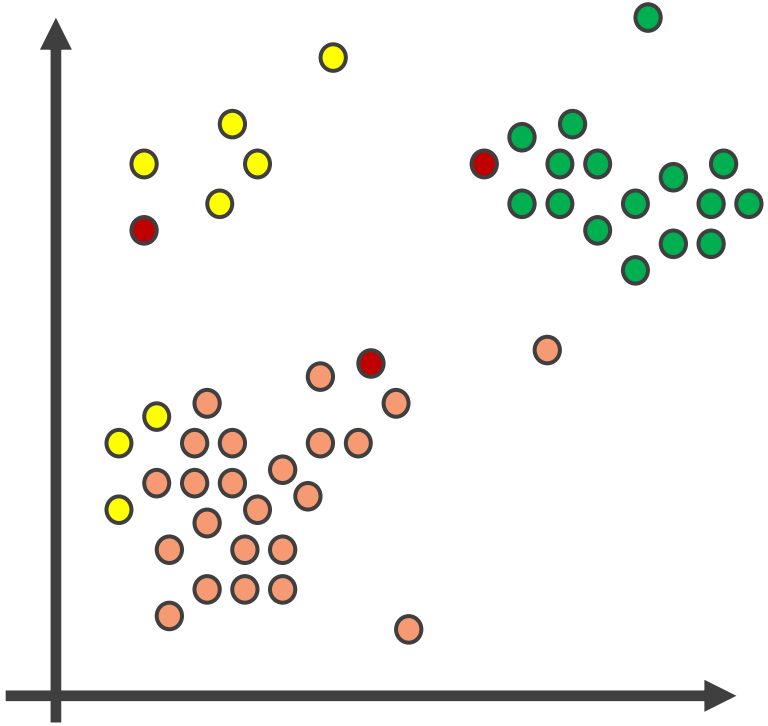
K-KÖZÉP K-MEANS



1. Klaszterközéppontok (centroid) véletlenszerű kiválasztása

2. Pontok hozzárendelése a legközelebbi centroidhoz

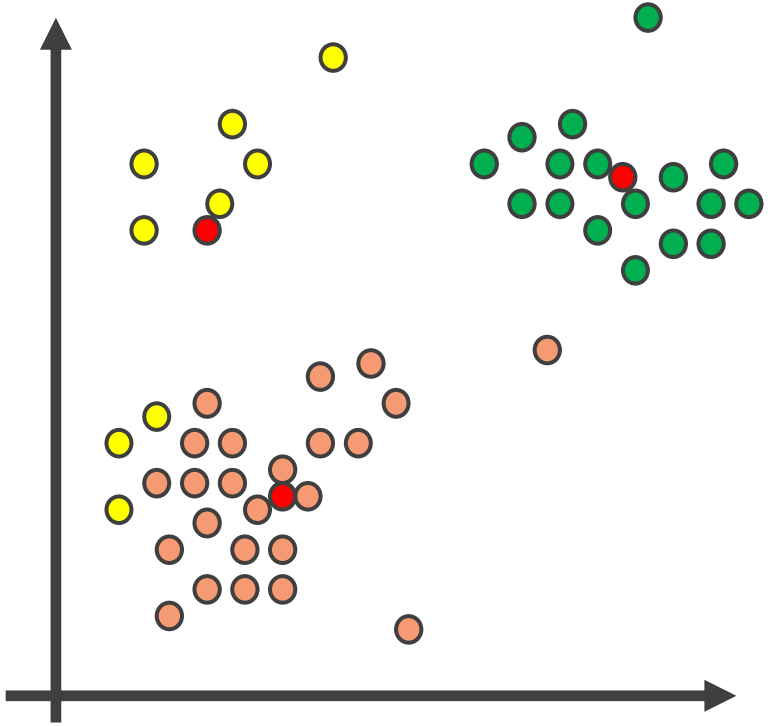
K-KÖZÉP K-MEANS



1. Klaszterközéppontok (centroid) véletlenszerű kiválasztása

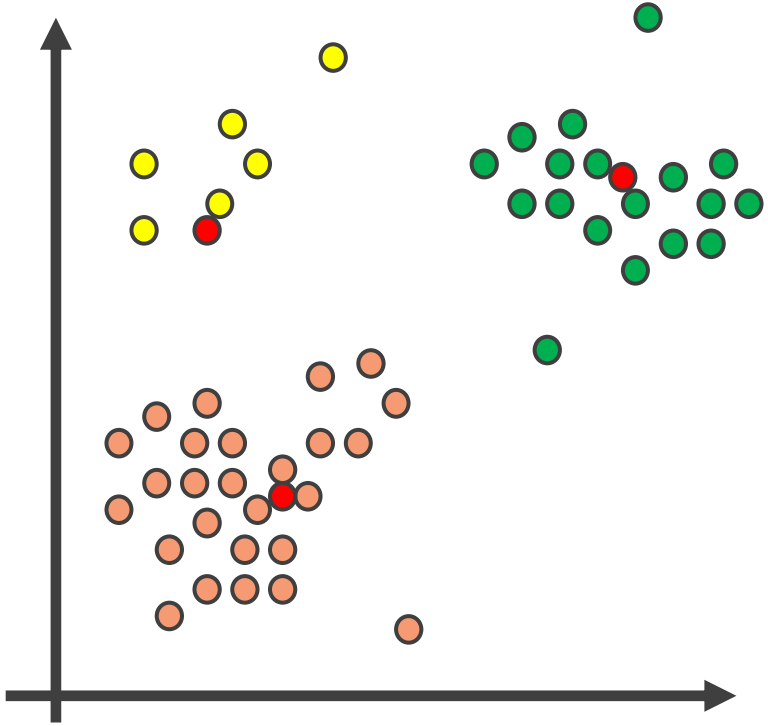
2. Pontok hozzárendelése a legközelebbi centroidhoz

K-KÖZÉP K-MEANS



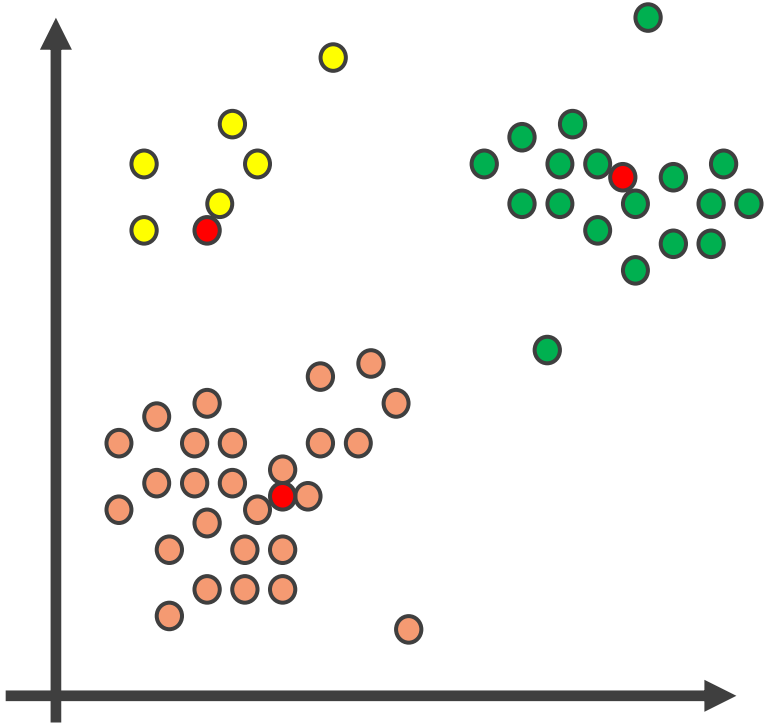
1. Klaszterközéppontok (centroid) véletlenszerű kiválasztása
2. Pontok hozzárendelése a legközelebbi centroidhoz
3. Centroid újraszámítás (klaszterek súlypontja)

K-KÖZÉP K-MEANS



1. Klaszterközéppontok (centroid) véletlenszerű kiválasztása
2. Pontok hozzárendelése a legközelebbi centroidhoz
3. Centroid újraszámítás (klaszterek súlypontja)

K-KÖZÉP K-MEANS



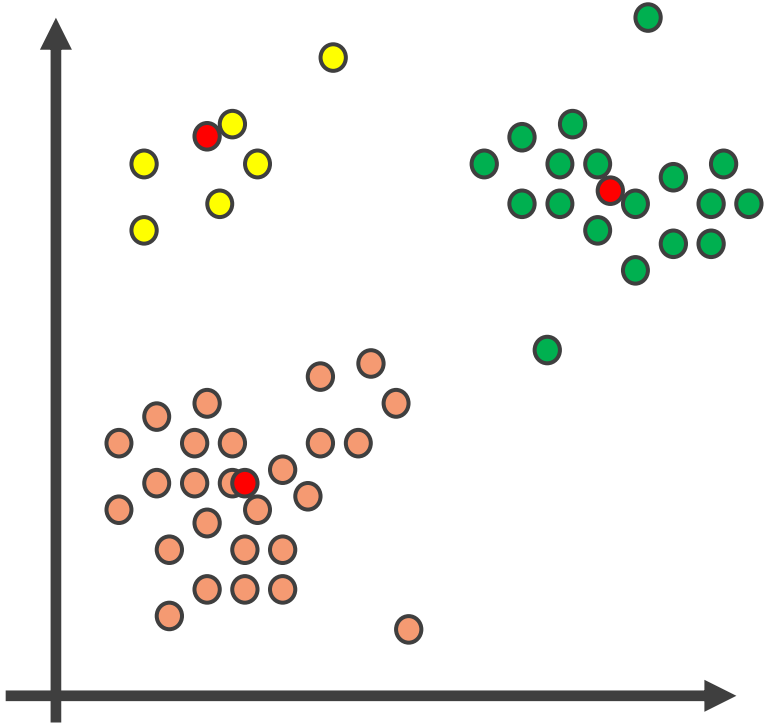
1. Klaszterközéppontok (centroid) véletlenszerű kiválasztása

2. Pontok hozzárendelése a legközelebbi centroidhoz

3. Centroid újraszámítás

2-3. Pontok ismétlése, míg nem változik

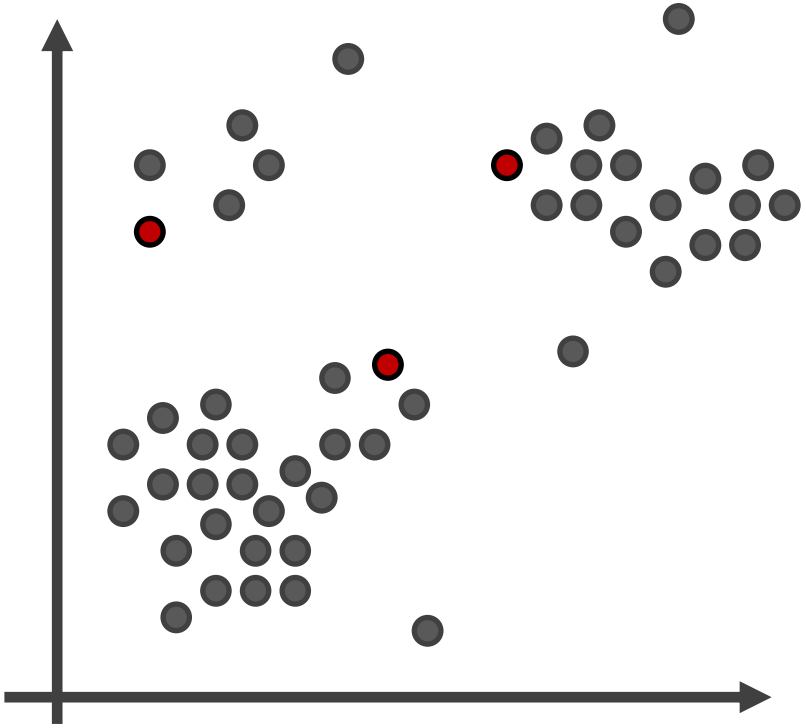
K-KÖZÉP K-MEANS



1. Klaszterközéppontok (centroid) véletlenszerű kiválasztása
2. Pontok hozzárendelése a legközelebbi centroidhoz
3. Centroid újraszámítás

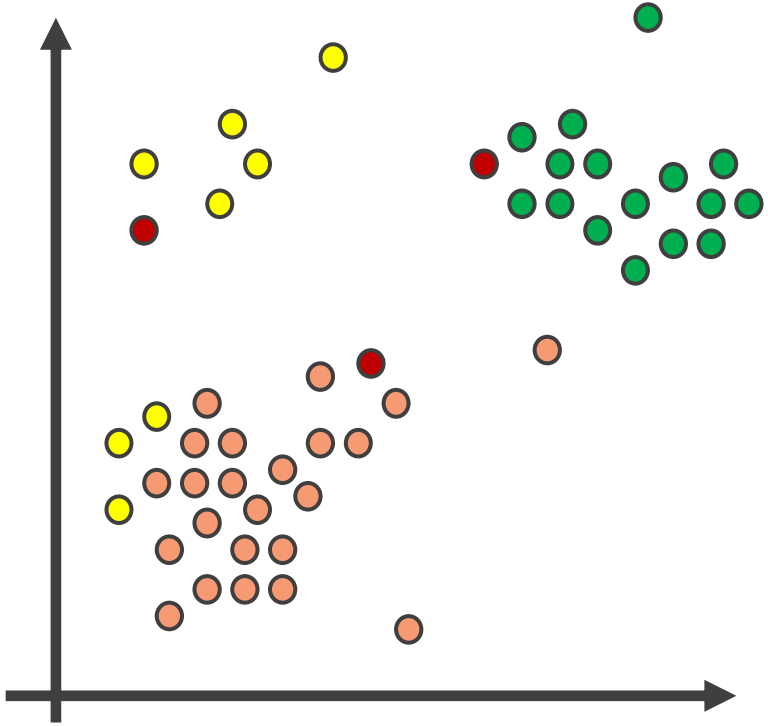
2-3. ismétlése, míg nem változik

K-MEDOID



1. Klaszter középpontok (centroid) véletlenszerű kiválasztása

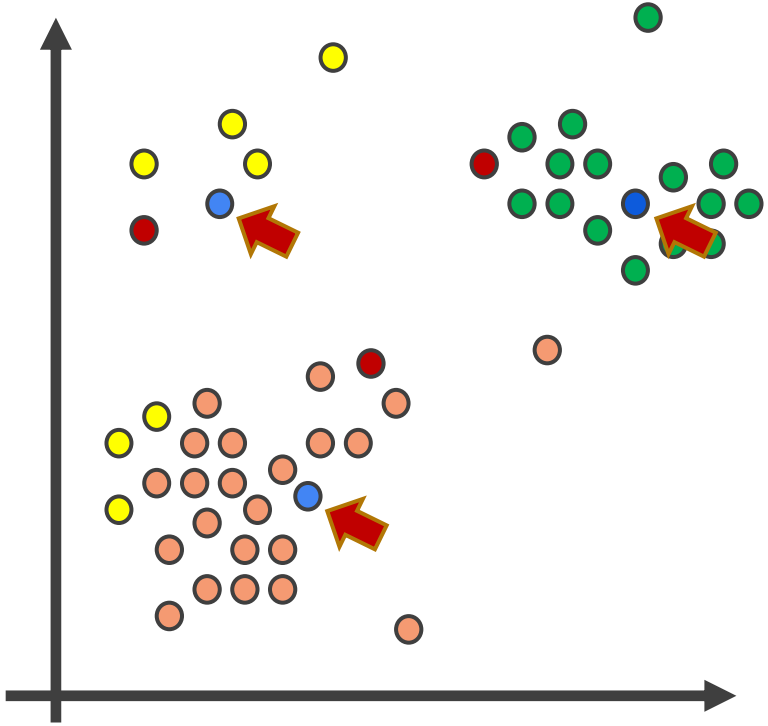
K-MEDOID



1. Klaszter középpontok (centroid) véletlenszerű kiválasztása

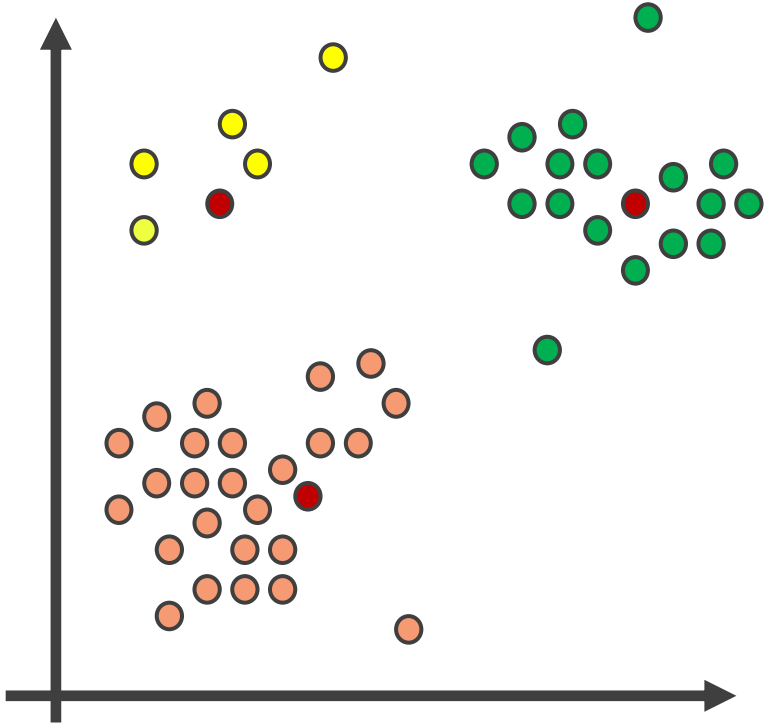
2. Pontok hozzárendelése a legközelebbi centroidhoz

K-MEDOID



1. Klaszter középpontok (centroid) véletlenszerű kiválasztása
2. Pontok hozzárendelése a legközelebbi centroidhoz
3. Centroid újraszámítás: Átlagosan legközelebb lévő pont

K-MEDOID

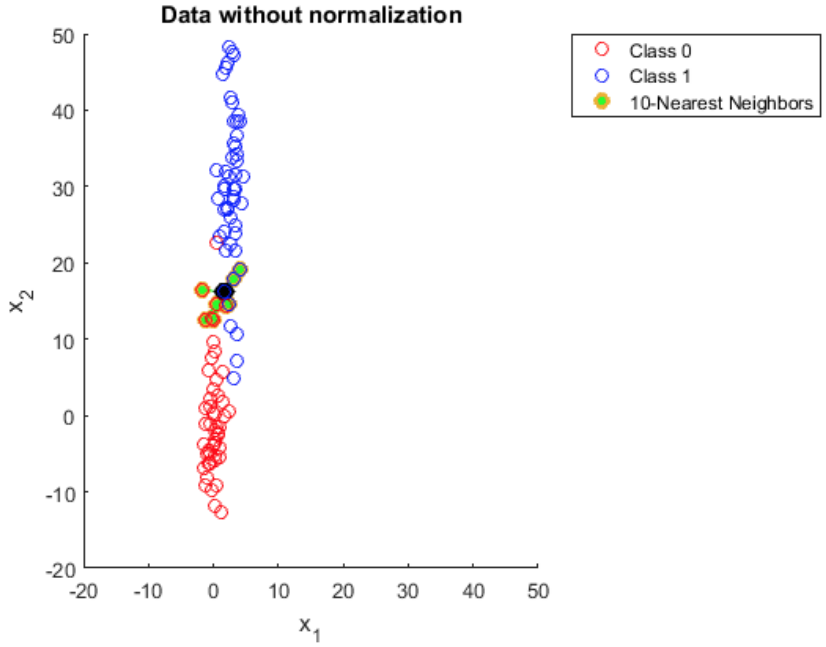


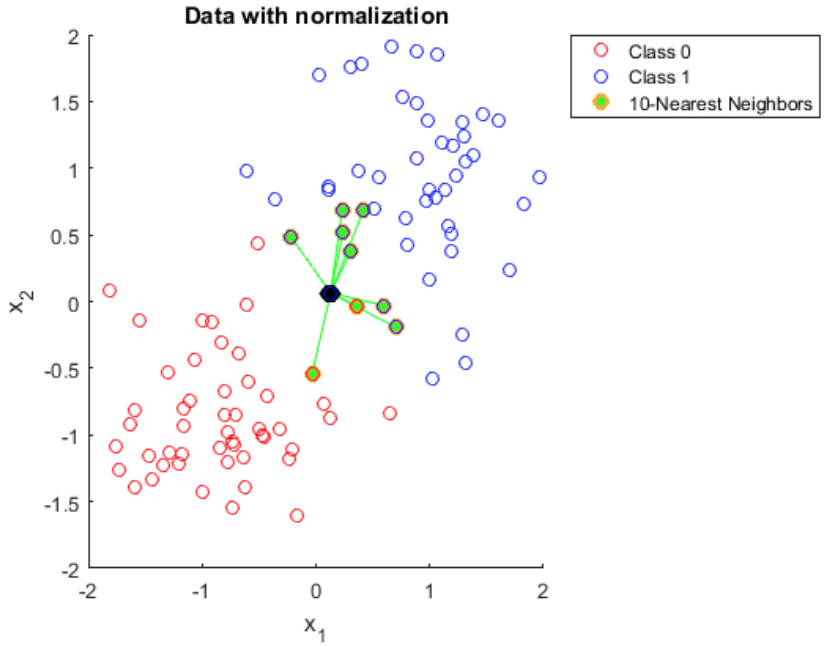
1. Klaszter középpontok (centroid) véletlenszerű kiválasztása
2. Pontok hozzárendelése a legközelebbi centroidhoz
3. Centroid újraszámítás: Átlagosan legközelebb lévő pont
4. Iteráció, míg nem konvergál 2. ponttól

K-MEANS KIHÍVÁSAI

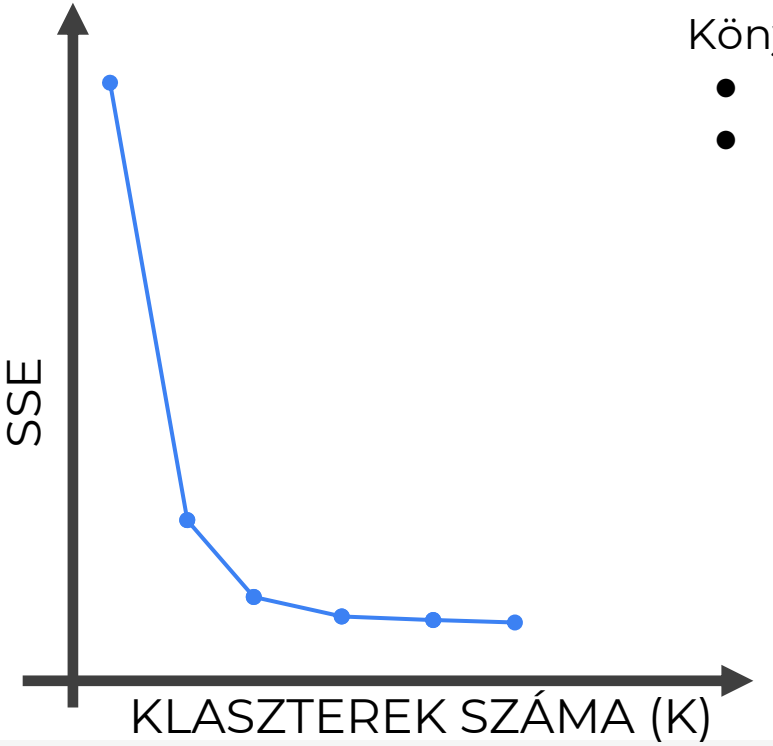
1. **Hiányzó értékek**
   Nem tudunk távolságot számolni
2. **Kiugró értékek**
   Torzítják a centroidok pozícióját
3. **Eltérő skálák**
   Nagyobb skálájú változók dominálnak
4. **Kategorikus változók**
   Nem tudunk távolságot számolni
5. **Bemenet meghatározása**
   Irreleváns változók zajt visznek a rendszerbe
6. **K meghatározása**
   Releváns és nem triviális csoportok megtalálása

Eltérő skálák





K MEGHATÁROZÁSA

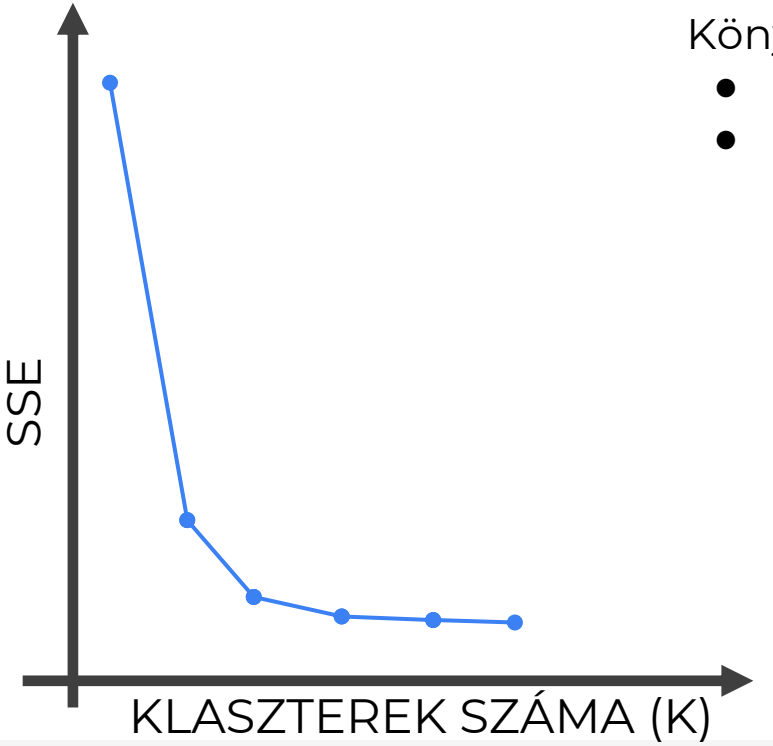


Könyökpont keresés

- Amíg az SSE dinamikusan csökken
- SSE: (sum of squared errors) Klaszteren belüli négyzetes távolság a klaszterközépponttól

$$SSE = \sum_{j=1}^{k} \sum_{i=1}^{n} (xi^{(j)} - c_j)^2$$

K MEGHATÁROZÁSA



Könyökpont keresés

- Amíg az SSE dinamikusan csökken
- SSE: (sum of squared errors) Klaszteren belüli négyzetes távolság a klaszterközépponttól

Elemek száma
(az adott klaszterben)

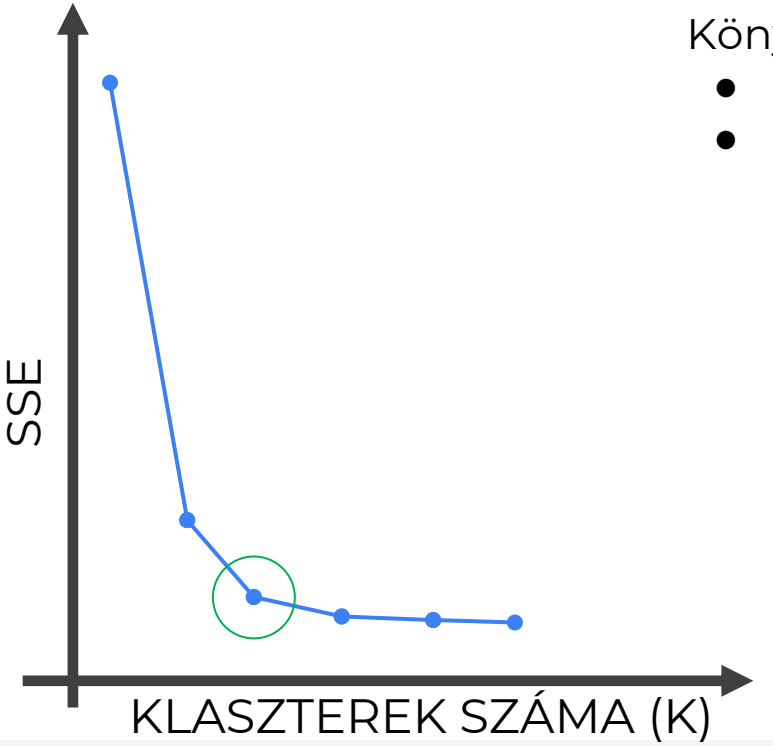
Klaszter centroid

$$SSE = \sum_{j=1}^{k} \sum_{i=1}^{n} (xi^{(j)} - c_j)^2$$

Klaszterek száma

Klaszter i. eleme

K MEGHATÁROZÁSA



Könyökpont keresés

- Amíg az SSE dinamikusan csökken
- SSE: (sum of squared errors) Klaszteren belüli négyzetes távolság a klaszterközépponttól

Elemek száma
(az adott klaszterben)

Klaszter centroid

$$SSE = \sum_{j=1}^{k} \sum_{i=1}^{n} (xi^{(j)} - c_j)^2$$

Klaszterek száma

Klaszter i. eleme

K-MEANS K-MEDOID

Előnyök

1. Könnyen implementálható
2. Gyors

Hátrányok

1. Előre meg kell határoznunk a klaszterek számát
2. Kezdeti centroidokra nagyon érzékeny
3. Gömb alakú klasztereket talál
4. Hasonló méretű és sűrűségű klaszterekre működik jól
5. Klasztereket hipersíkokkal (több D sík) vágja el egymástól

Kezdjünk bele!

Irány a notebook

URL: https://bme.dmlab.hu/tree?
JELSZÓ: BME_D@t@23

HA KÉRDÉSED VAN, BÁTRAN KERESS MINKET!

HUSZTI DOROTTYA
DATA SCIENTIST
huszti.dorottya@dmlab.hu