Hierarchikus klaszterezés

dmlab

Klaszterezés

2024.09.16.

Huszti Dorottya
data scientist

Klaszterező algoritmusok

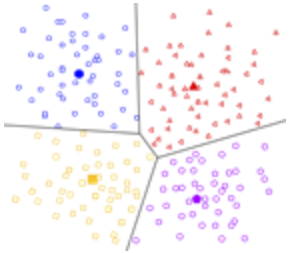
ALGORITMUSOK FAJTÁI

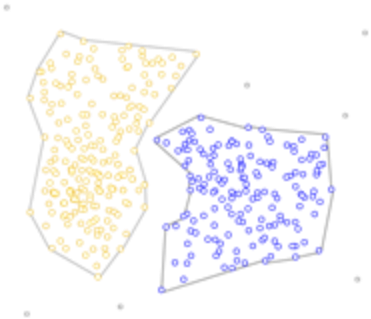
Távolság alapú

K-Means, K-Medoids

Sűrűség alapú

DBSCAN, Mean-shift clustering





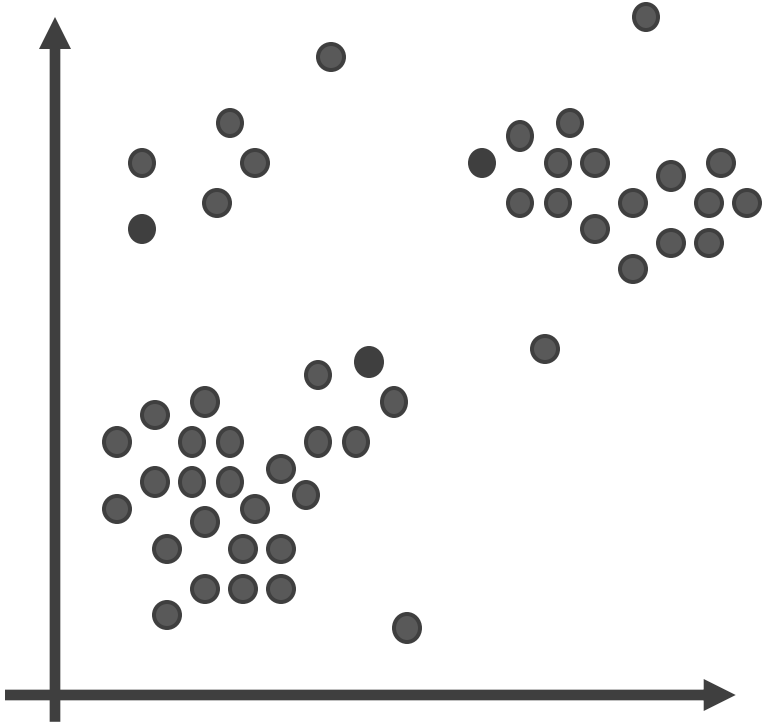
Partícionáló

Hierarchikus

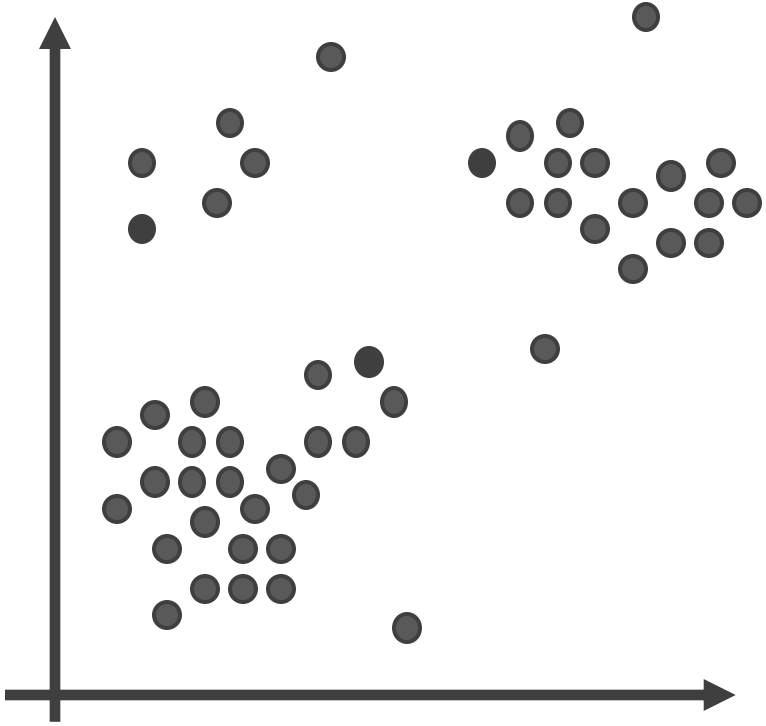
K-közép (K-means)

Klaszterező
algoritmusok

K-KÖZÉP K-MEANS

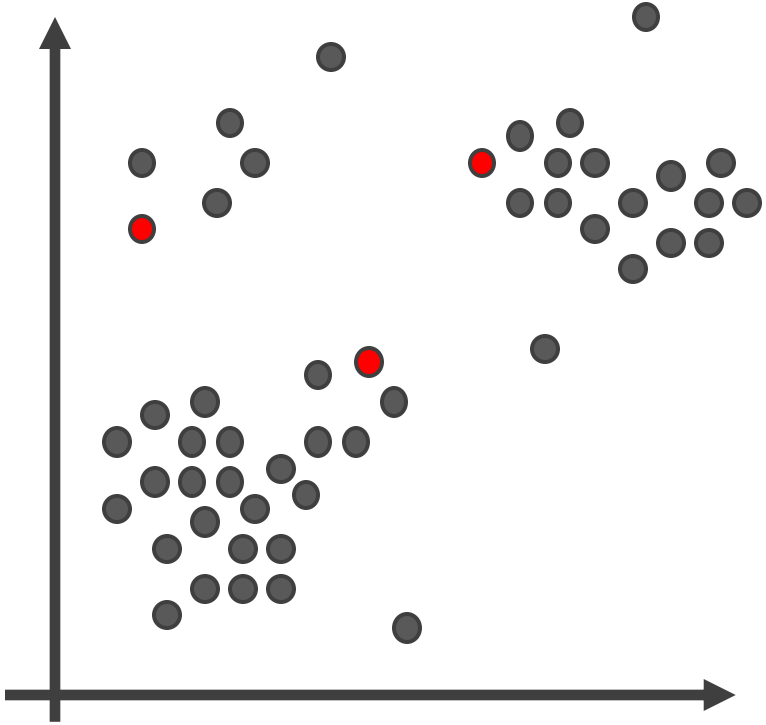


K-KÖZÉP K-MEANS



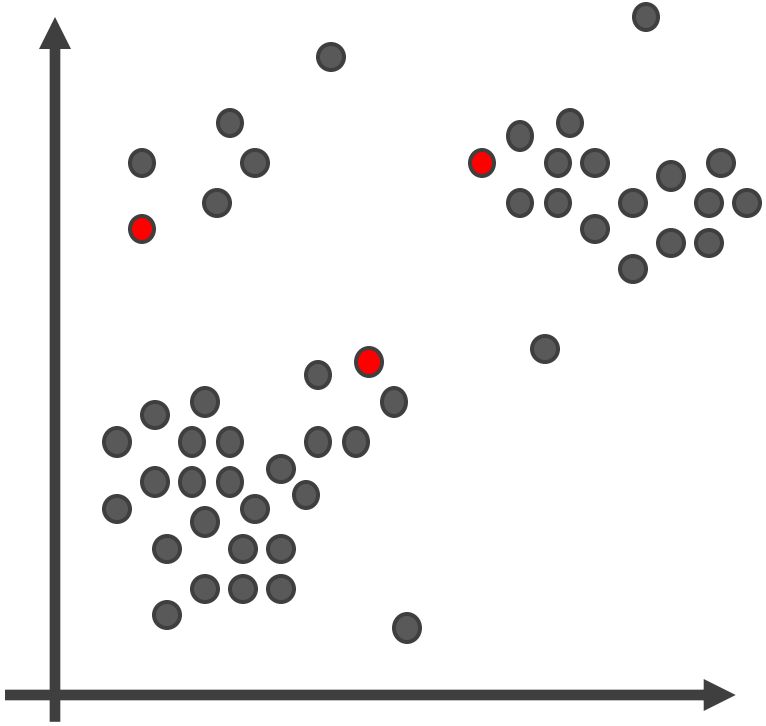
1. Klaszterközéppontok (centroid) véletlenszerű kiválasztása

K-KÖZÉP K-MEANS



1. Klaszterközéppontok (centroid) véletlenszerű kiválasztása

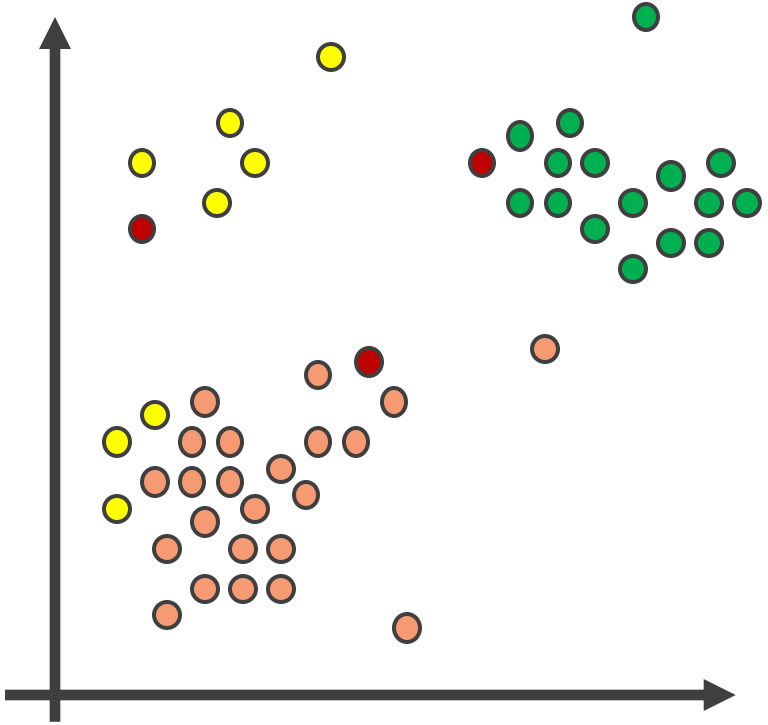
K-KÖZÉP K-MEANS



1. Klaszterközéppontok (centroid) véletlenszerű kiválasztása

2. Pontok hozzárendelése a legközelebbi centroidhoz

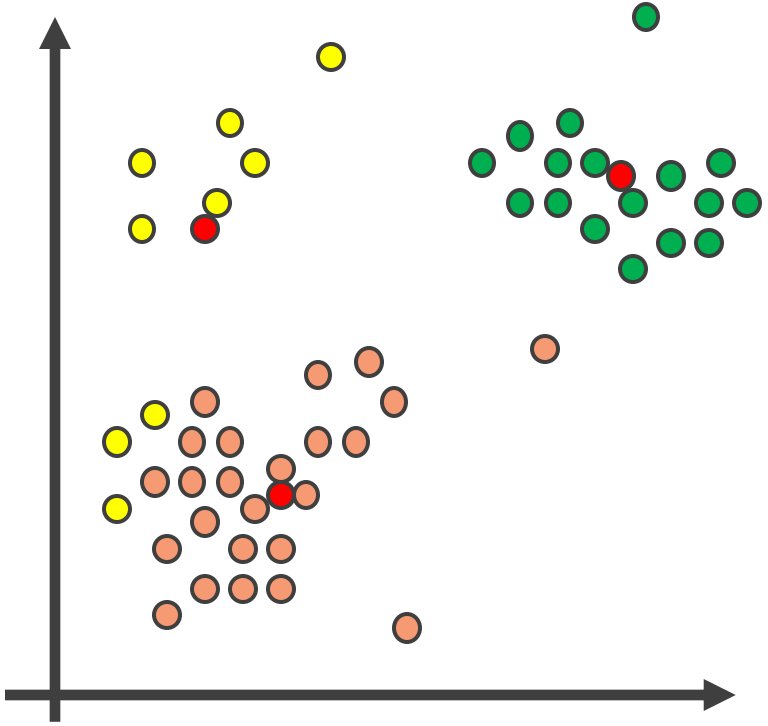
K-KÖZÉP K-MEANS



1. Klaszterközéppontok (centroid) véletlenszerű kiválasztása

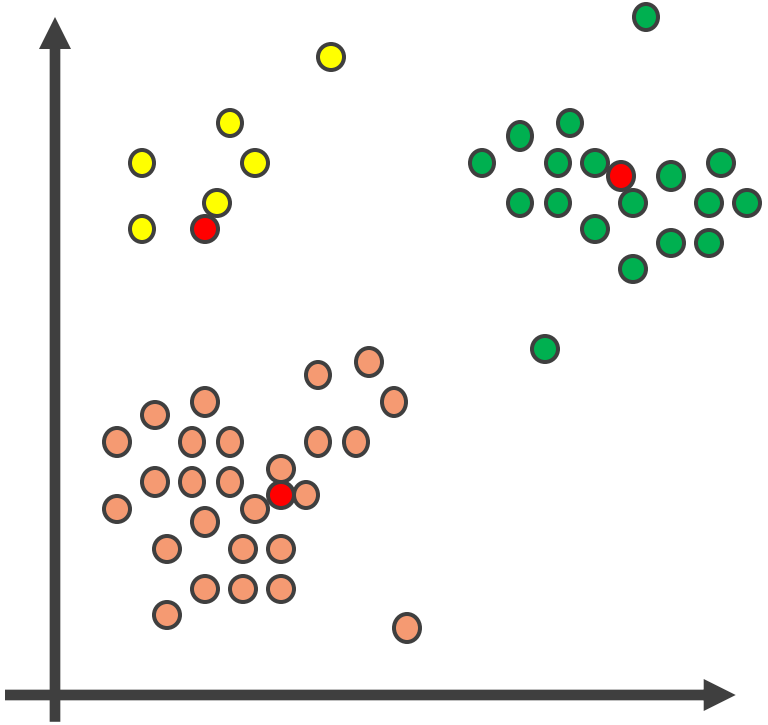
2. Pontok hozzárendelése a legközelebbi centroidhoz

K-KÖZÉP K-MEANS



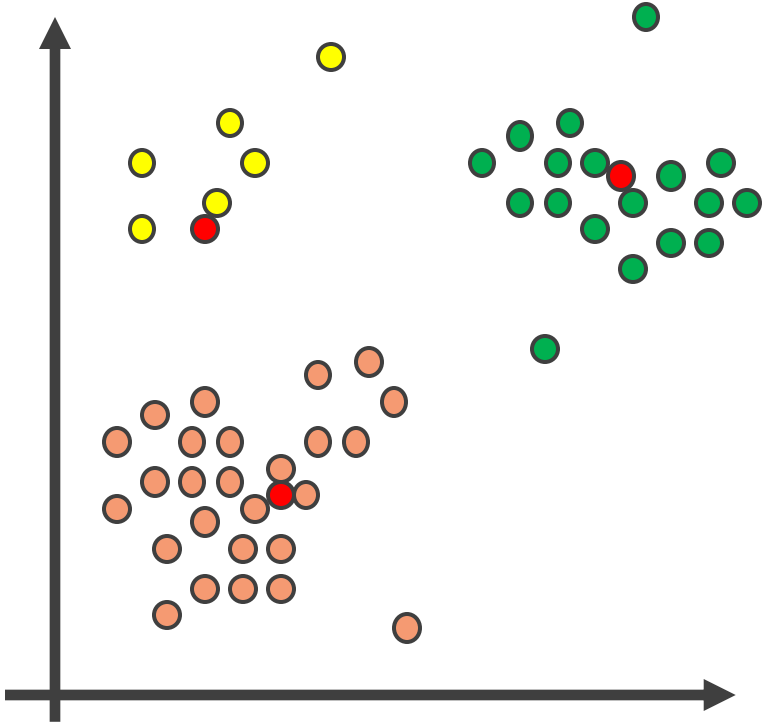
1. Klaszterközéppontok (centroid) véletlenszerű kiválasztása
2. Pontok hozzárendelése a legközelebbi centroidhoz
3. Centroid újraszámítás (klaszterek súlypontja)

K-KÖZÉP K-MEANS



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

K-KÖZÉP K-MEANS



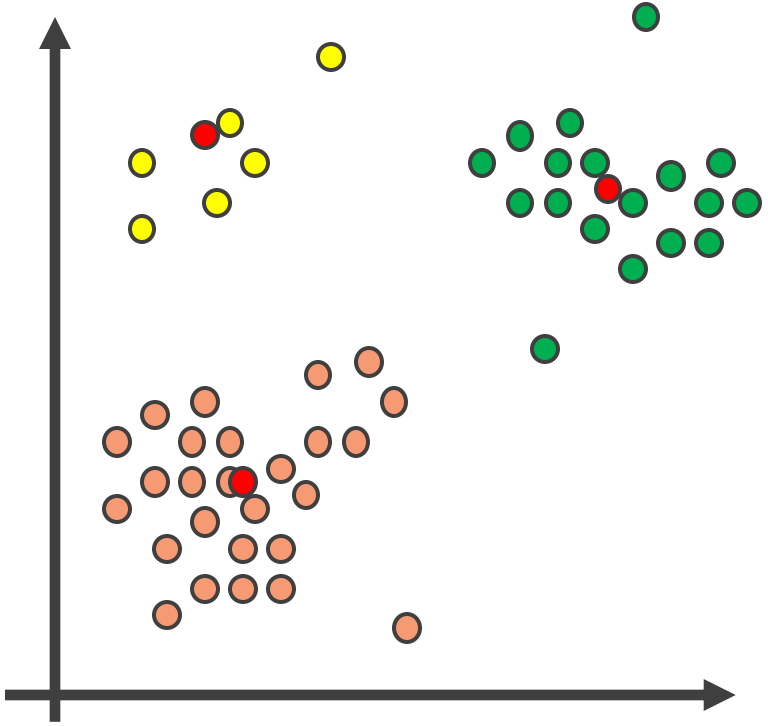
1. Klaszterközéppontok (centroid) véletlenszerű kiválasztása

2. Pontok hozzárendelése a legközelebbi centroidhoz

3. Centroid újraszámítás

2-3. Pontok ismétlése, míg nem változik

K-KÖZÉP K-MEANS



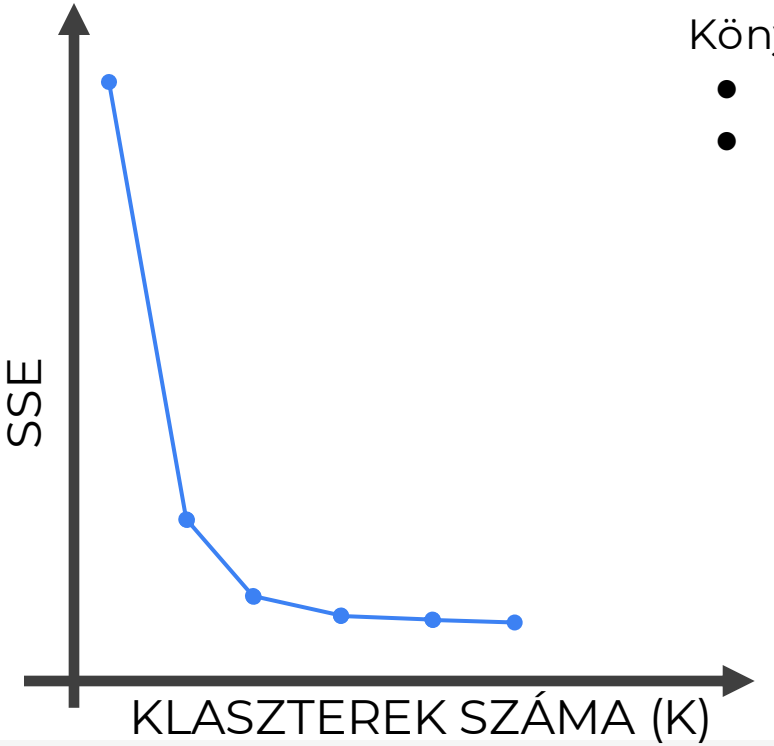
1. Klaszterközéppontok (centroid) véletlenszerű kiválasztása
2. Pontok hozzárendelése a legközelebbi centroidhoz
3. Centroid újraszámítás

2-3. ismétlése, míg nem változik

K-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

K MEGHATÁROZÁSA

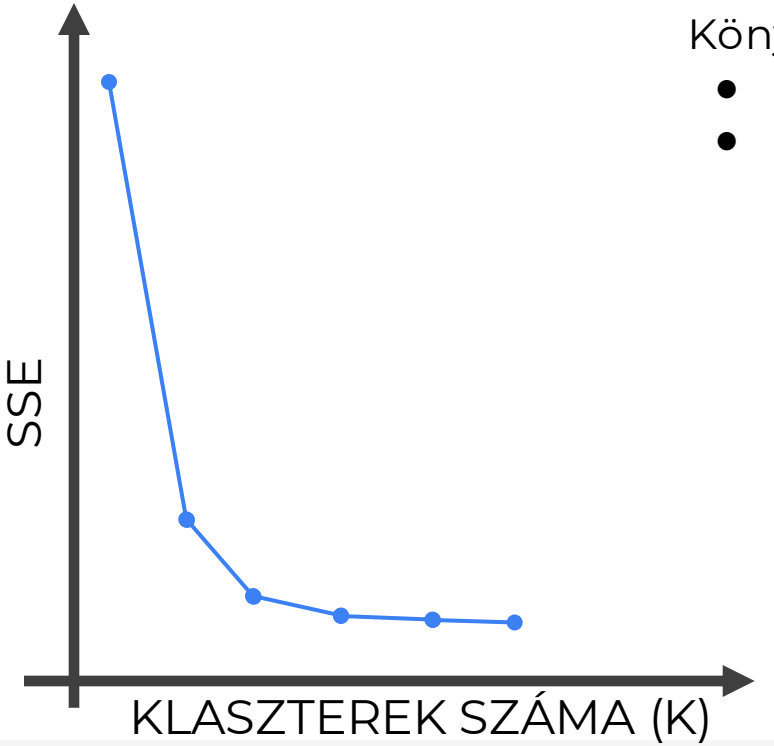


Könyökpont keresés

- 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)

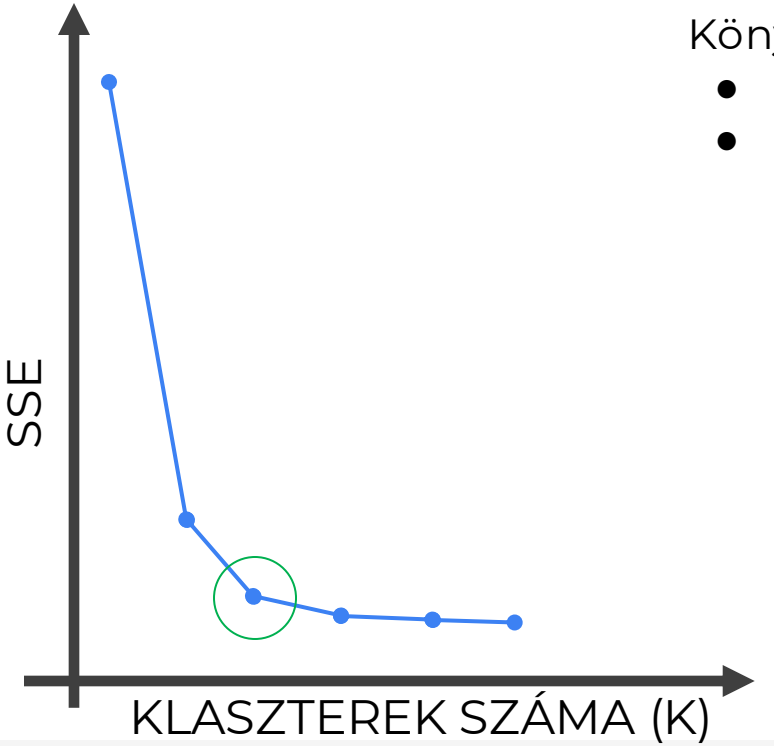
Klaszter centroid

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

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

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

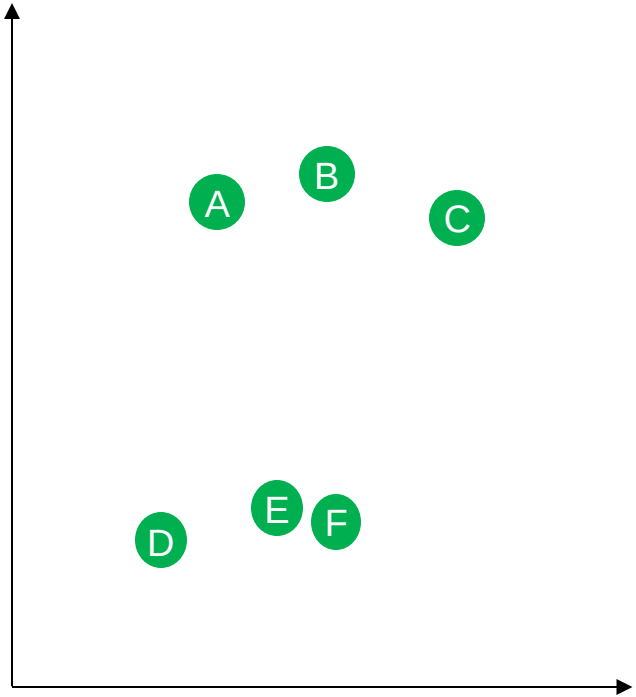
Klaszterek száma

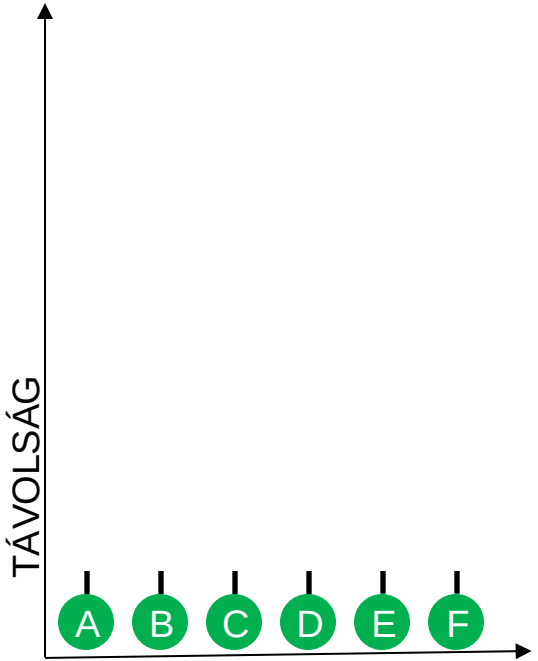
Klaszter i. eleme

Hierarchikus klaszterezés

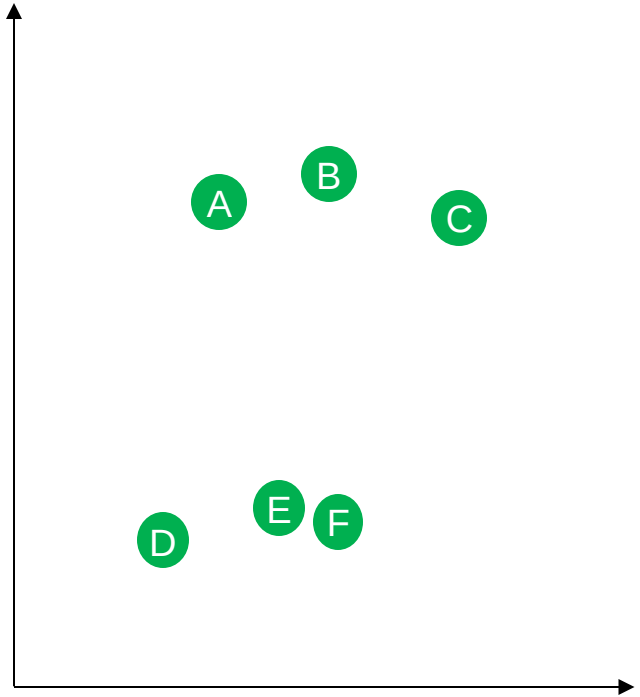
Klaszterező
algoritmusok

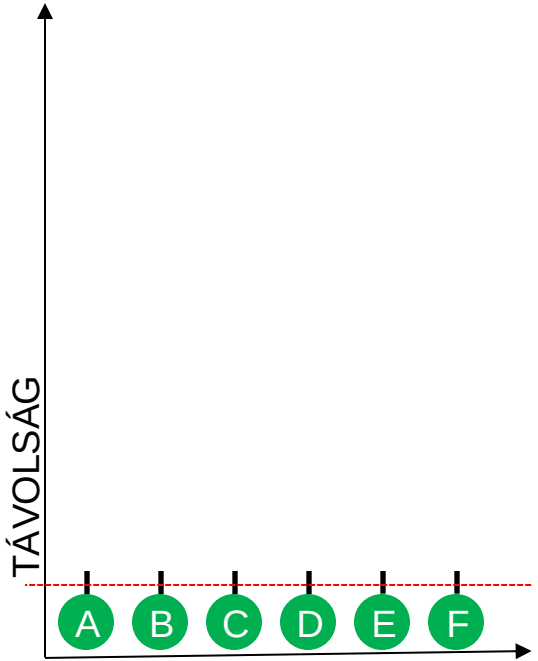
HIERARCHIKUS KLASZTEREZÉS



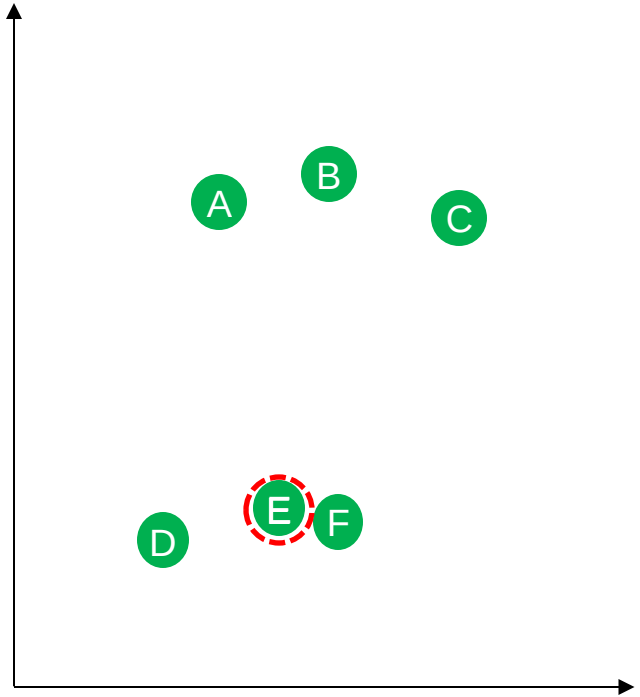


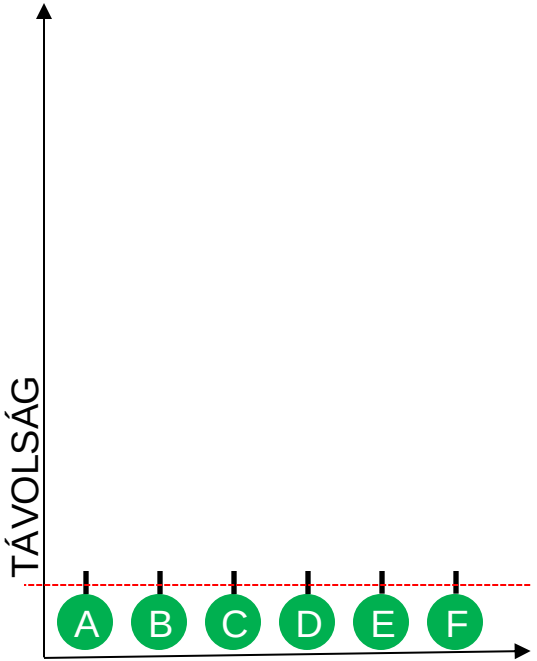
HIERARCHIKUS KLASZTEREZÉS



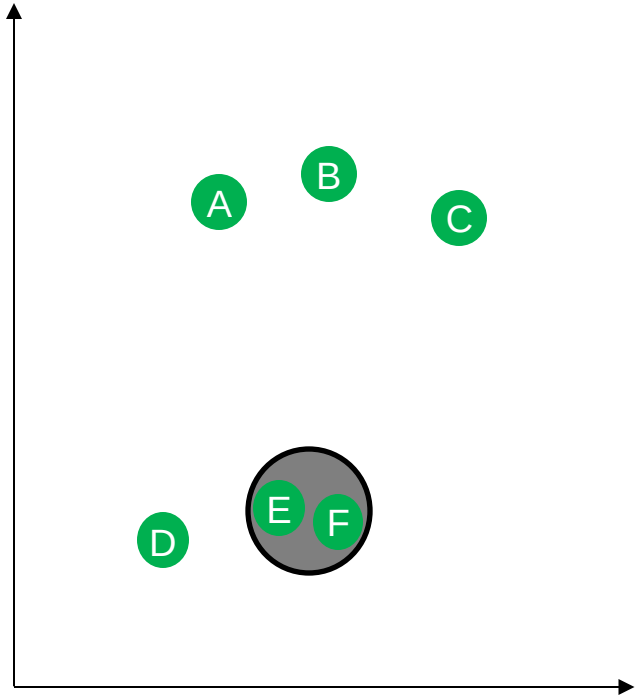


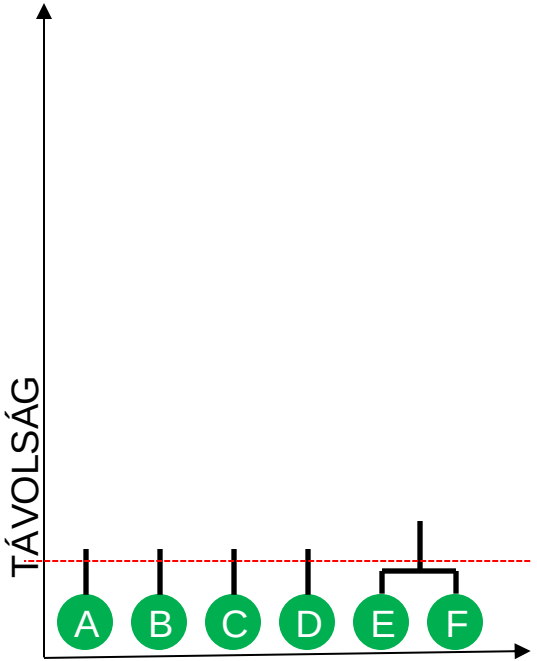
HIERARCHIKUS KLASZTEREZÉS



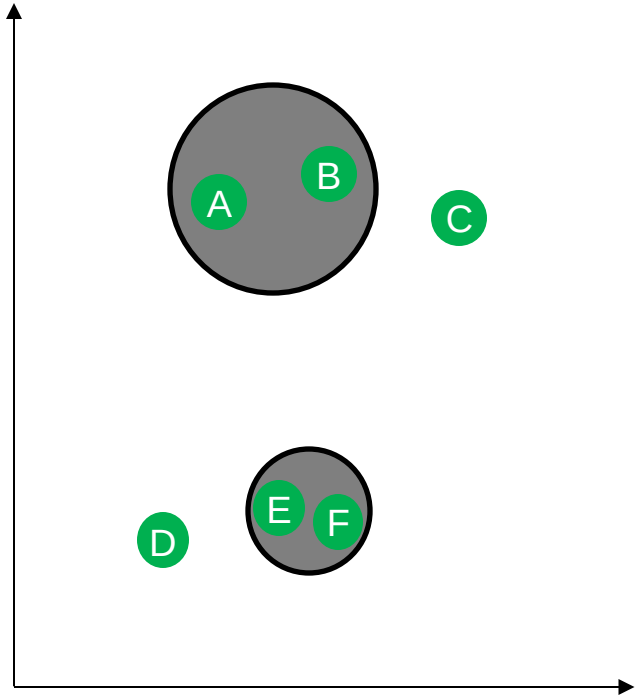


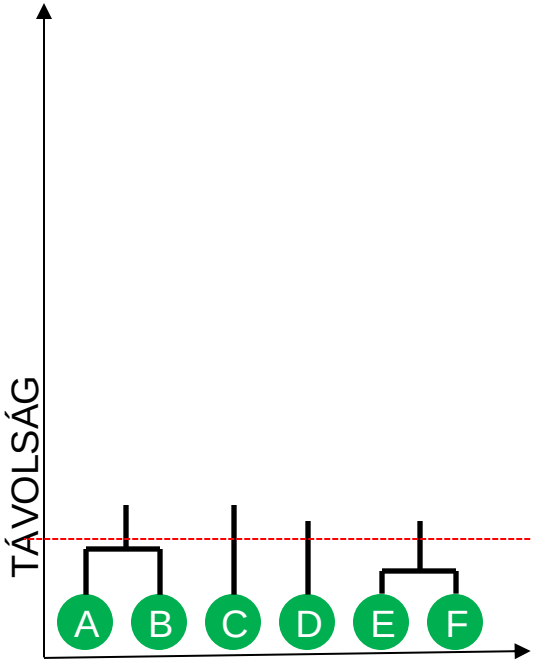
HIERARCHIKUS KLASZTEREZÉS



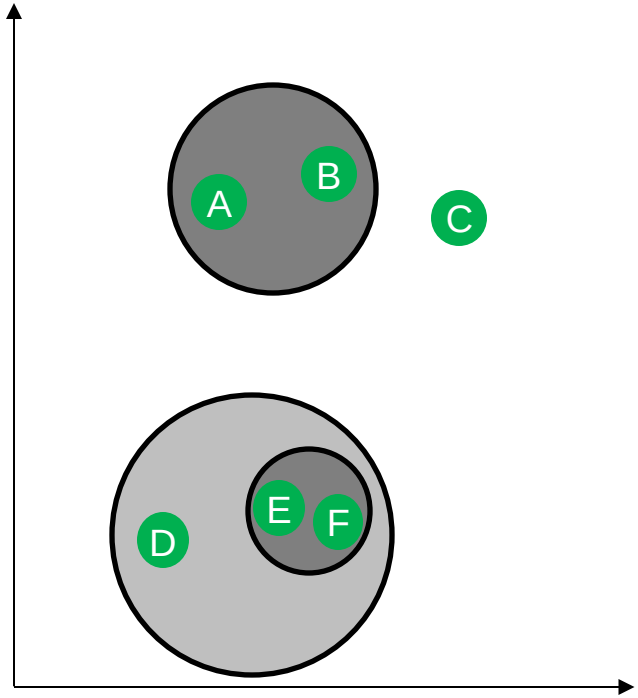


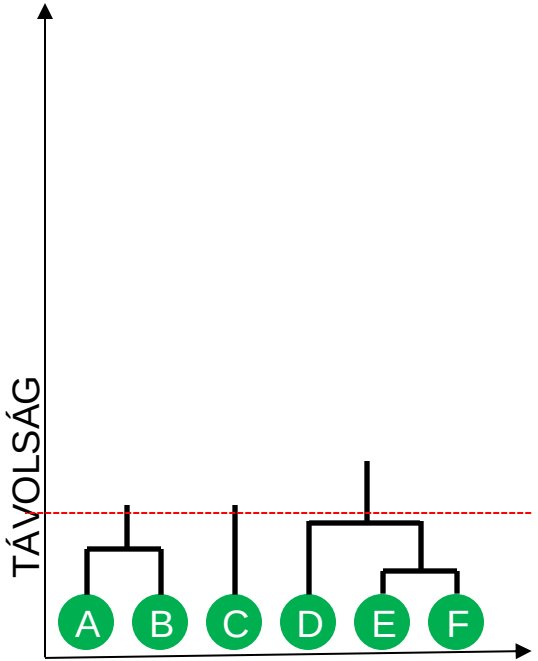
HIERARCHIKUS KLASZTEREZÉS



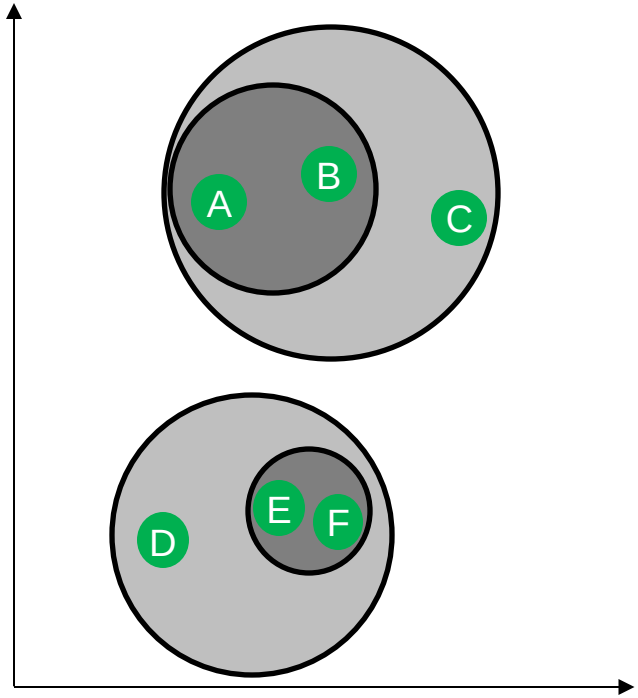


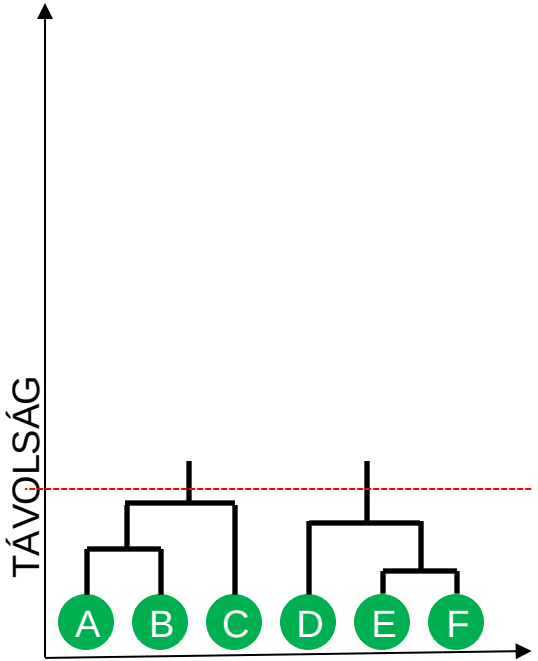
HIERARCHIKUS KLASZTEREZÉS



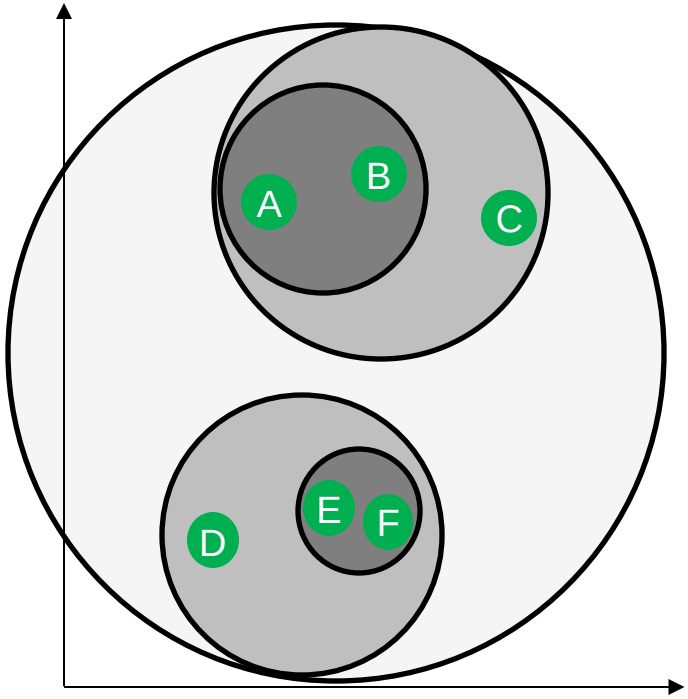


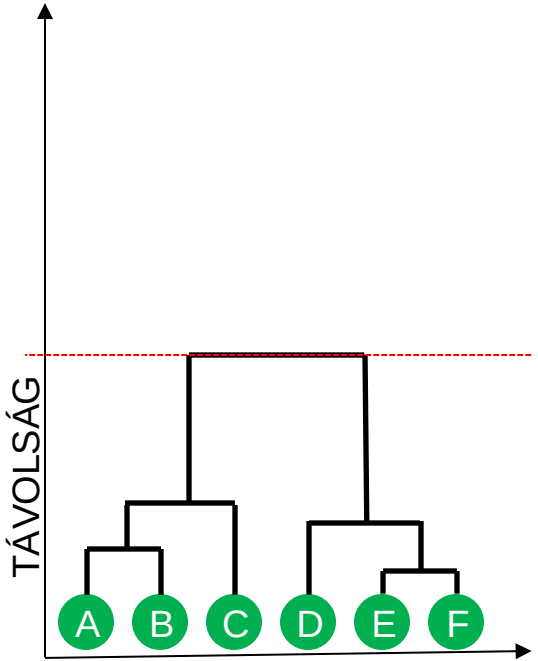
HIERARCHIKUS KLASZTEREZÉS



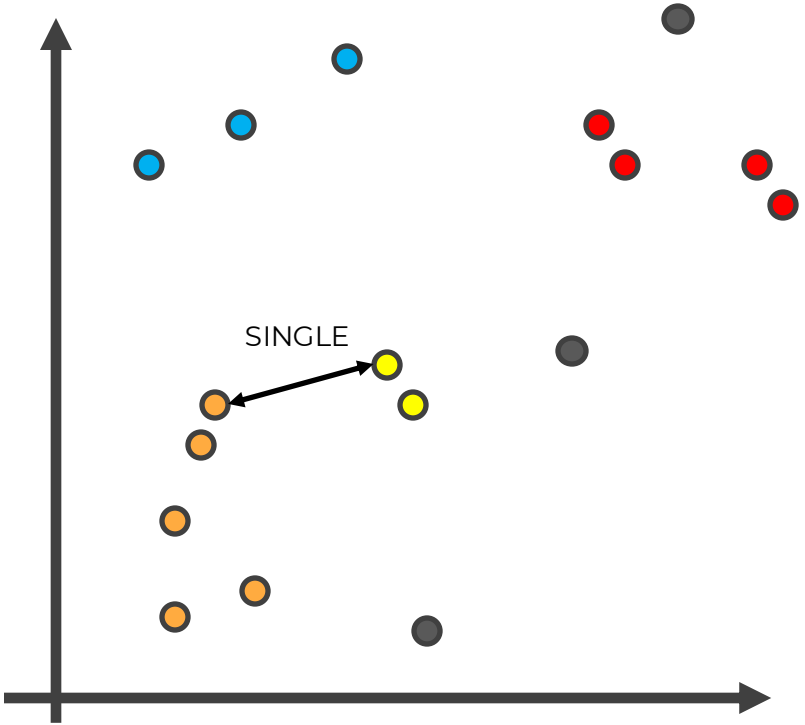


HIERARCHIKUS KLASZTEREZÉS





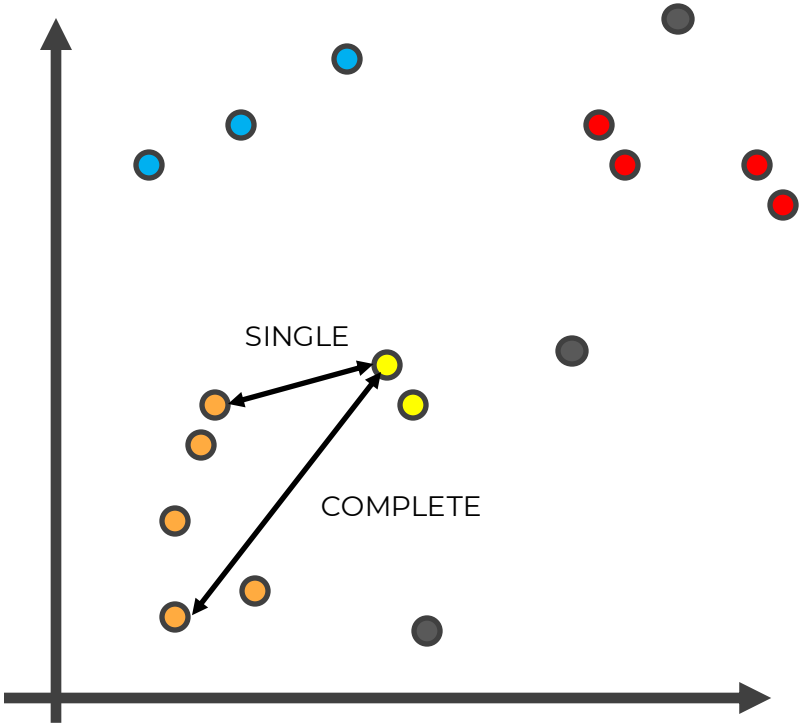
HIERARCHIKUS KLASZTEREZÉS



KÉT KLASZTER TÁVOLSÁGÁT TÖBB MÓDON IS SZÁMOLHATJUK:

1. **SINGLE** – LEGKÖZELEBBI PONTOK TÁVOLSÁGA (UBORKASZERŰ KLASZTEREK)

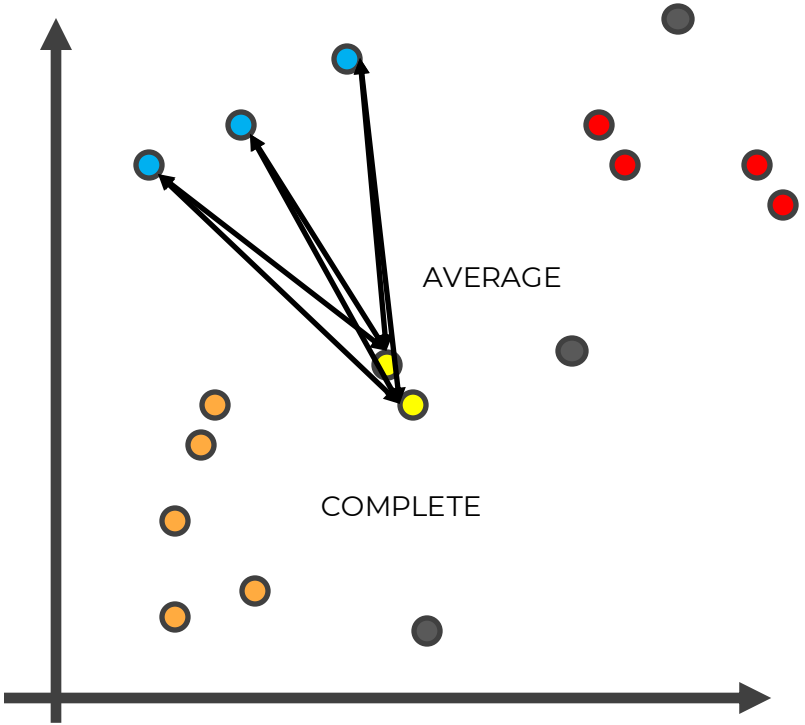
HIERARCHIKUS KLASZTEREZÉS



KÉT KLASZTER TÁVOLSÁGÁT TÖBB MÓDON IS SZÁMOLHATJUK:

1. **SINGLE** – LEGKÖZELEBBI PONTOK TÁVOLSÁGA (UBORKASZERŰ KLASZTEREK)
2. **COMPLETE** – LEGTÁVOLABBI PONTOK TÁVOLSÁGA (SZFÉRIKUS KLASZTEREK, KIUGRÓ ÉRTÉKEKRE ÉRZÉKENY)

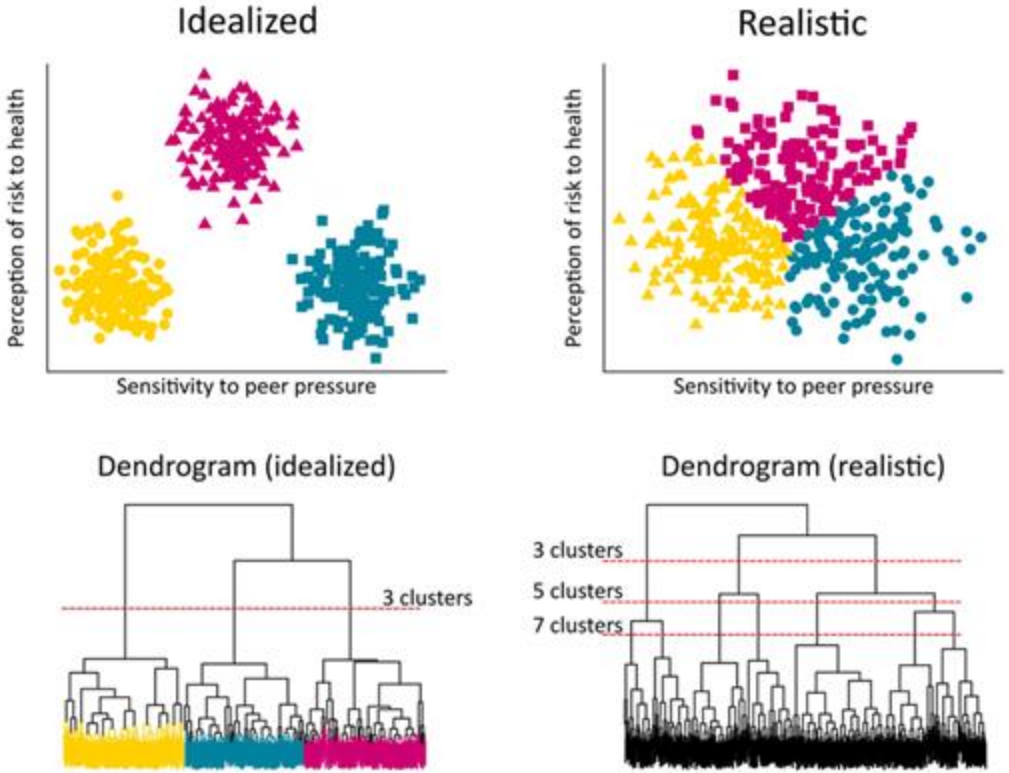
HIERARCHIKUS KLASZTEREZÉS



KÉT KLASZTER TÁVOLSÁGÁT TÖBB MÓDON IS SZÁMOLHATJUK:

1. **SINGLE** – LEGKÖZELEBBI PONTOK TÁVOLSÁGA (UBORKASZERŰ KLASZTEREK)
2. **COMPLETE** – LEGTÁVOLABBI PONTOK TÁVOLSÁGA (SZFÉRIKUS KLASZTEREK, KIUGRÓ ÉRTÉKEKRE ÉRZÉKENY)
3. **AVERAGE** – KÉT KLASZTER ÖSSZES ELEMÉN VETT ÁTLAGOS TÁVOLSÁGA

HIERARCHIKUS KLASZTEREZÉS

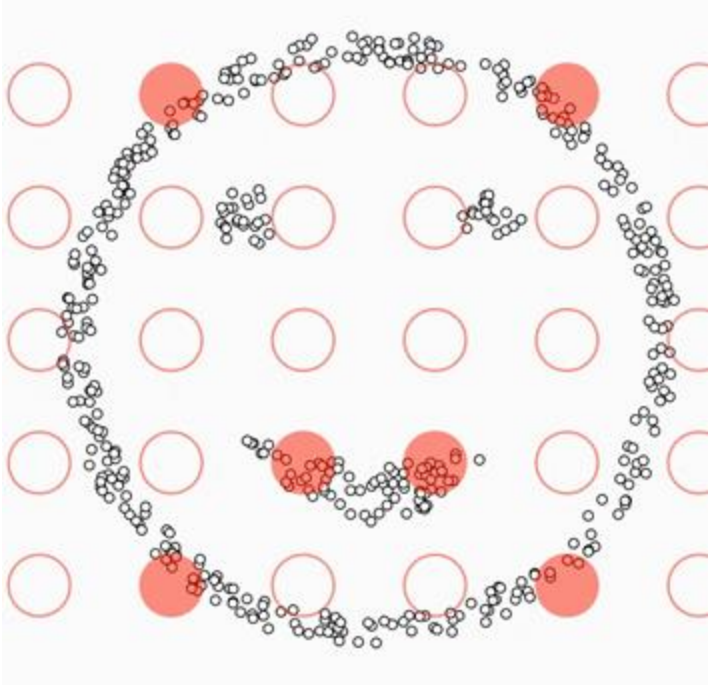


DBSCAN

Density-based spatial clustering of applications with noise

Klaszterező
algoritmusok

DBSCAN



1. Válasszunk ki véletlenszerűen egy pontot, ahol még nem jártunk,
    a. ha van kellő számú pont ennek környezetében, akkor kezdjük a klaszterezést,
    b. Ha nincs, zajnak minősítjük

DBSCAN

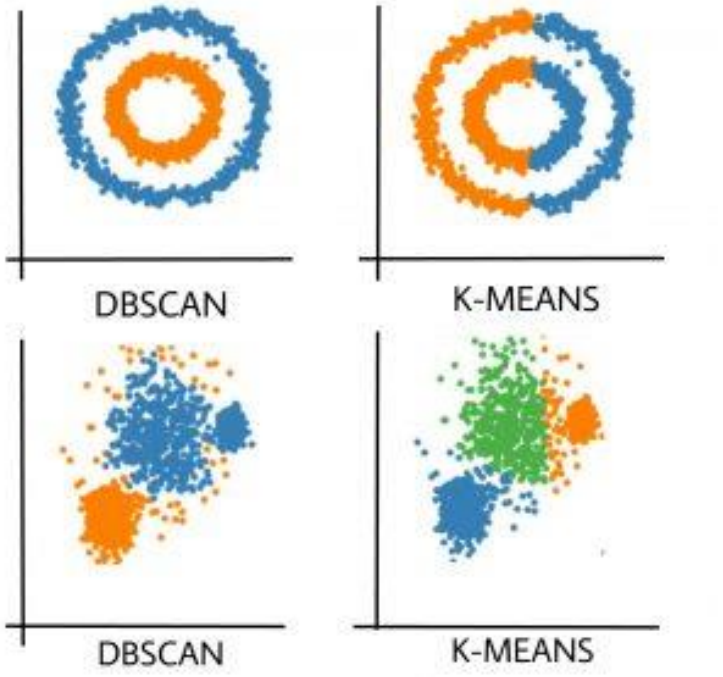
Előnyök

1. Érzéketlen a zajra
2. Kiugró adatokat automatikusan kezeli (figyelmen kívül hagyja)
3. Eltérő méretű és formájú klasztereket jól kezel
4. Nem kell előre meghatároznunk a klaszterek számát

Hátrányok

1. Nem minden adatpont kerül besorolásra
2. Hiperparaméterekre érzékeny
3. Hasonló sűrűségű klaszterekre működik jól
4. Valós adatokon ritkán alkalmazható

DBSCAN vs. K-MEANS



HA KÉRDÉSED VAN, BÁTRAN KERESS MINKET!

HUSZTI DOROTTYA
DATA SCIENTIST
huszti.dorottya@dmlab.hu