TÖBBVÁLTOZÓS ADATELEMZÉS

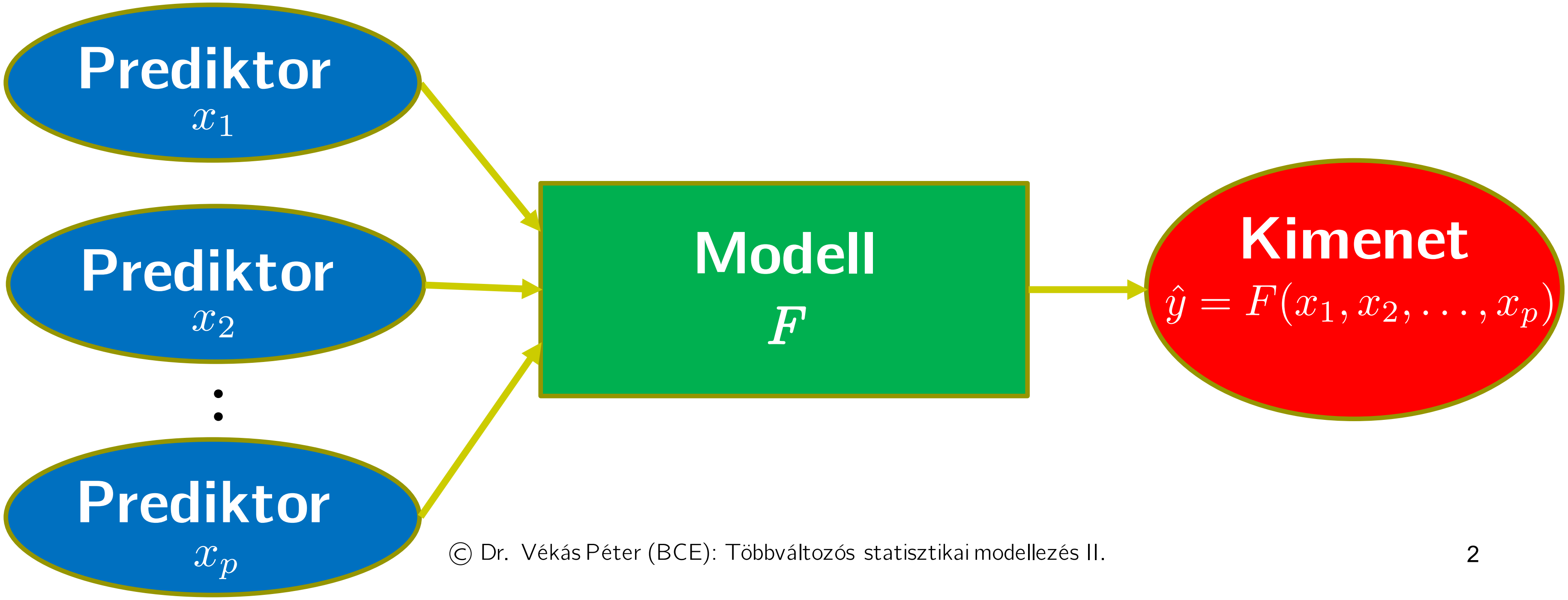
Klaszterelemzés

2020.11.16.

© Dr. Vékás Péter, e-mail: peter.vekas@uni-corvinus.hu
BCE Matematikai és Statisztikai Modellezés Intézet

Gépi tanulási modell mint függvény ("varázsdoboz")

- A gépi tanulási modellek prediktorváltozókból (=inputok, független változókat) hoznak létre egy kimenetet (=output, függő változó).



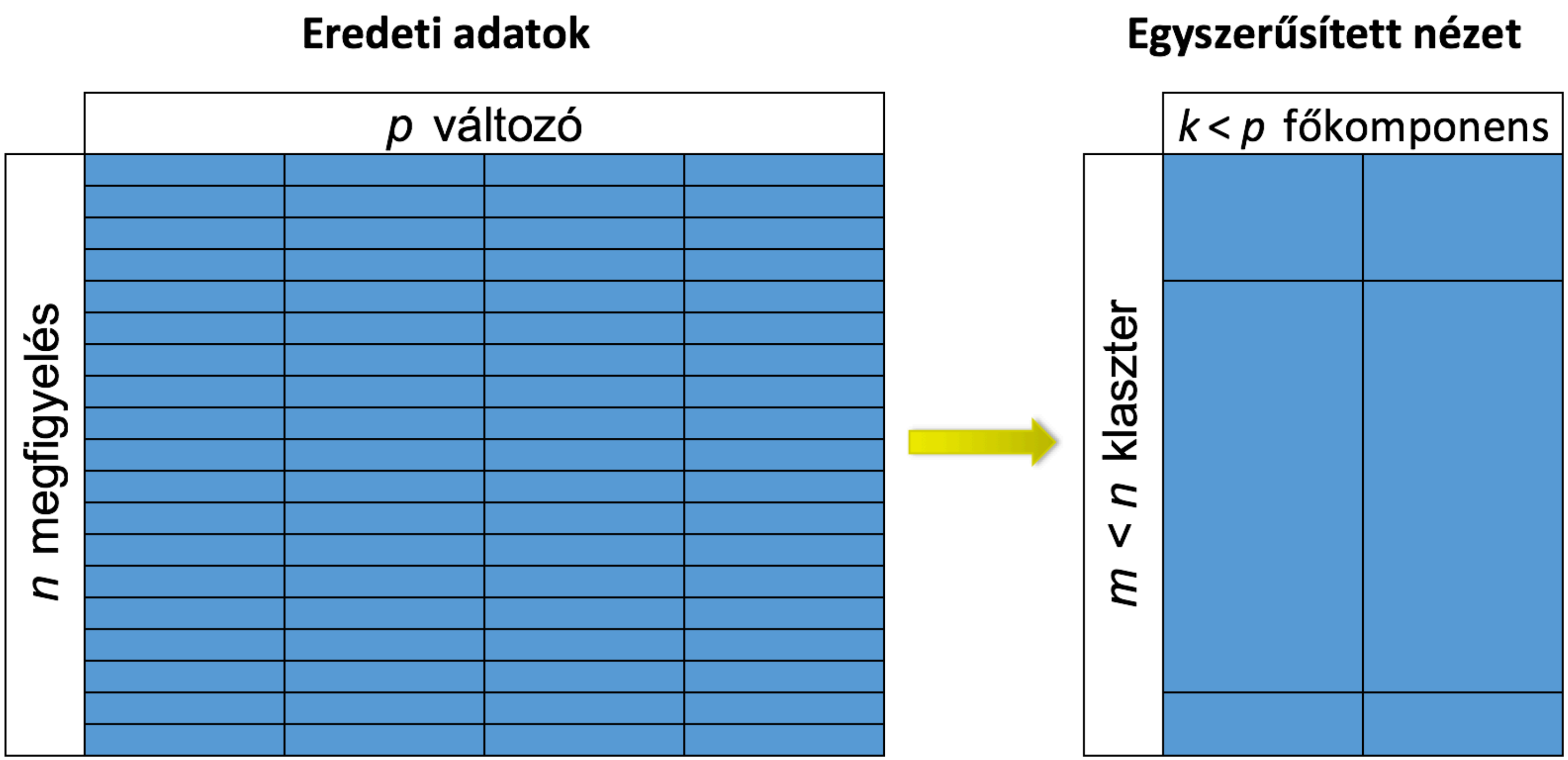
Gépi tanulási problémák típusai

- **Felügyelt tanulás:**
  Van egy adatfájlunk, ahol ismert a “valódi” kimenet. A cél ennek a becslése (predikció).
  - Osztályozás (klasszifikáció): kategorikus kimenet.
  - Regresszió: numerikus kimenet.
- **Nem felügyelt tanulás:**
  A kimenet ismeretlen, “láthatatlan” változó. A cél az adatok struktúrájának megismerése.
  - Klaszterezés: kategorikus kimenet.
  - Dimenziócsökkentés: numerikus kimenet.

Gépi tanulási problémák típusai

Kimenet / Modell	Felügyelt $\mathbf{x} \to y$	Nem felügyelt $\mathbf{x} \to ?$
Kategorikus 	Osztályozás	Klaszterezés
Numerikus 	Regresszió	Dimenzió-csökkentés

Az elemző a lényegre kíváncsi



Klaszterelemzés

- Természetes, jól elkülönülő csoportokat (*klasztereket*) keresünk a sorok között.
- Két egyidejű cél:
  - ① egymástól minél markánsabban különböző
  - ② és belül minél homogénebb klaszterek.
- A klaszterek diszjunktak, és lefedik a teljes fájlt.
- A megfigyelések közötti távolságokra épül.
- Adatvezérelt (data-driven) módszer: a csoportok nem önkényesek, hanem hagyjuk, hogy az adatok „beszéljenek”.

Definíció (Google Translate)

cluster

/ˈklʌstə/

noun

1. a group of similar things or people positioned or occurring closely together.
   "**clusters of** creamy-white flowers"

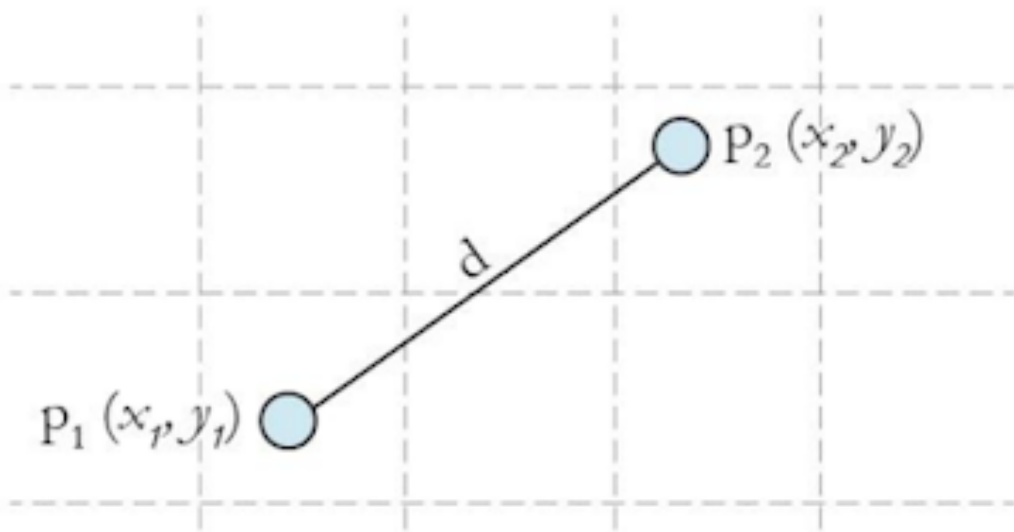
Source: http://www.cfgphoto.com

Source: https://www.artfire.com

Példák alkalmazásokra

- Egy vállalat ügyfélköre milyen jellegzetes ügyféltípusokból áll (piackutatás)?
- Milyen országcsoportokra tagozódik az Európai Unió a gazdasági-szociális mutatók terében?
- Milyen klikkek figyelhetők meg egy döntéshozó testületben vagy egy személyes kapcsolathálóban?
- Regressziós becslés esetén javíthatja a pontosságot, ha klaszterenként külön-külön egyenletek alapján történik a becslés.

Távolság mérése: euklideszi távolság



$$\text{Euclidean distance (d)} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Forrás: http://technokarak.com

Mennyire különbözik két ember?

Name	Age (years)	Income ($)	Height (cm)
Debbie	24	30 000	160
Patrick	40	41 000	175

$$d = \sqrt{(24-40)^2 + (30\,000 - 41\,000)^2 + (160-175)^2} =$$

$$= \sqrt{16^2 + 11\,000^2 + 15^2} \approx \mathbf{11\,000}$$

Mi van, ha a jövedelmet 1000$-ban mérjük?

Name	Age (years)	Income ($1000)	Height (cm)
Debbie	24	30	160
Patrick	40	41	175

$$d = \sqrt{(24-40)^2 + (30-41)^2 + (160-175)^2} =$$
$$= \sqrt{16^2 + 11^2 + 15^2} \approx \mathbf{24.54}$$

És ha a jövedelmet 1000$-ban és a magasságot m-ben mérjük?

Name	Age (years)	Income ($1000)	Height (m)
Debbie	24	30	1.60
Patrick	40	41	1.75

$$d = \sqrt{(24-40)^2 + (30-41)^2 + (1.60-1.75)^2} =$$
$$= \sqrt{16^2 + 11^2 + 0.15^2} \approx \mathbf{19.42}$$

Mértékegységek

- Minden mértékegység-átváltáskor változik a távolság. ☹
- Pedig nem kellene függnie a mértékegységtől.
- Valahogyan el kell tüntetni a mértékegységeket!

Sztenderdizálás

- Az x változó átlaga és szórása:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}$$

- Sztenderdizált változó:

$$z_i = \frac{x_i - \bar{x}}{s}$$

- Átlaga 0, szórása 1.
- Jelentése: x_i az átlagnál z_i szórással nagyobb.

Hierarchikus klaszterezés

- **Lépések:**
  1. Kezdetben minden megfigyelés önálló klaszter.
  2. A két *leghasonlóbb* klaszter összevonása.
  3. Az előző lépést ismétlése addig, amíg végül egyetlen klaszter nem marad (*lépések száma?*).
- A kívánt számú klaszter elérésénél megállítható az eljárás.
- **Két klaszter különbözőségének mérése:**
  - Ward-távolság
  - Egyéb módszerek (legközelebbi/legtávolabbi szomszéd, medián/centroid, átlagos lánc stb.)

Ward-távolság

- Az **a** és **b** p-dimenziós térbeli pontok (megfigyelések) közötti d(**a**,**b**) **euklideszi távolság:**

$$d(\mathbf{a},\mathbf{b}) = \sqrt{\sum_{j=1}^{p} (a_j - b_j)^2}.$$

- Az A klaszter $H(A)$ **belső heterogenitása:** a klaszter $\mathbf{a}_k$ ($k = 1,2,\ldots,n_A$) elemei és $\mathbf{c}_A$ koordinátánkénti átlagpontja között mért távolságok négyzetösszege.

$$H(A) = \sum_{k=1}^{n_A} d^2(\mathbf{a}_k, \mathbf{c}_A).$$

- Az A és a B klaszterek $\mathrm{D}(A,B)$ **Ward-távolsága:** mennyivel növekedne a teljes belső heterogenitás a klaszterek egyesítése következtében.

$$D(A,B) = H(A \cup B) - H(A) - H(B).$$

Dendrogram

- A hierarchikus klaszterezés menetének vizuális megjelenítése.
- A vízszintes tengelyen a megfigyelések, a függőleges tengelyen pedig a klaszterek összevonásakor a klaszterek között mért távolságok láthatók.
- Látható rajta, hogy hogyan tömörülnek homogén csoportokba a megfigyelések.
- Támpontot adhat a klaszterek számának megállapításához.

k-középpontú klaszterezés

- Véletlenszerűen kiválaszt *k* darab kezdeti klaszterközéppontot a *p*-dimenziós térben.
- Minden megfigyelést a hozzá legközelebbi középpontú klaszterbe sorol be.
- A klaszterközéppontokat felülírja a hozzájuk legközelebb eső pontok átlagpontjával.
- Bebizonyítható, hogy így csökken a klasztereken belüli eltérés-négyzetösszeg.
- Addig ismétli a frissítést, míg a csökkenés már elhanyagolható (konvergencia).

k-középpontú klaszterezés online demo

- http://shabal.in/visuals/kmeans/6.html

k-középpontú klaszterezés képletekkel

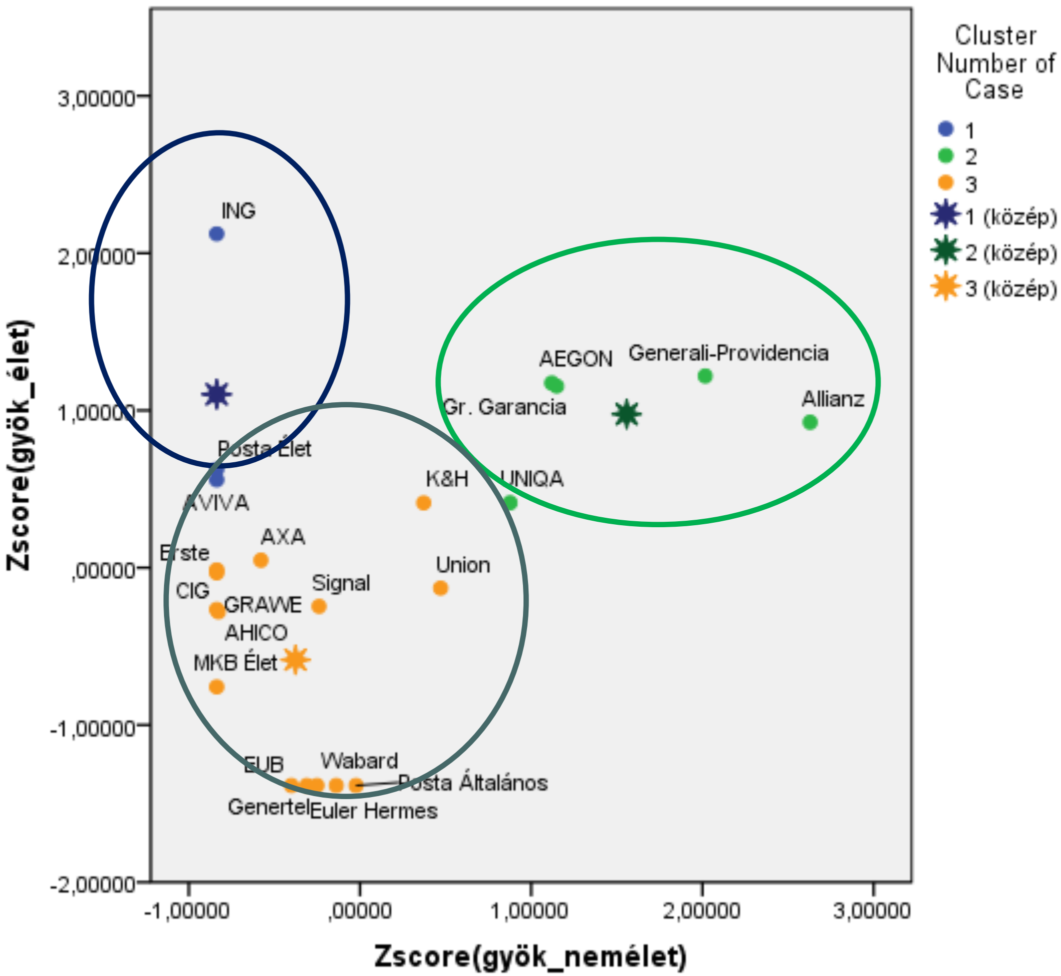
- Adottak: a megfigyelések $\mathbf{x}_i$ $(i=1,2,\ldots,n)$ sorvektorai és a k klaszterszám.
- Meghatározandók: a klaszterközepek $\mathbf{c}_j$ $(j=1,2,\ldots,k)$ vektorai.
- i-edik megfigyelés és j-edik klaszterközép euklideszi távolsága:

$$d(\mathbf{x}_i,\mathbf{c}_j)=\sqrt{\sum_{\ell=1}^{p}\left(x_{i\ell}-c_{j\ell}\right)^2}$$

- Célfüggvény:

$$E(\mathbf{c}_1,\mathbf{c}_2,\ldots,\mathbf{c}_k)=\sum_{i=1}^{n}\min_{j=1,2,\ldots,k} d^2(\mathbf{x}_i,\mathbf{c}_j)\rightarrow \min.$$

Példa: A magyar biztosítók három klasztere



A véletlen szerepének kiküszöbölése

- A kezdeti klaszterközéppontok véletlenek, így a végeredmény is az.
- Stabil végeredményt kapunk, ha sokszor újrafuttatjuk a módszert (*nstart* hiper-paraméter), és a legjobb (legkisebb SSE-t adó) végeredményt választjuk ki.
- A klaszterek sorrendje tetszőleges!
- A megengedett lépések számát (*iter.max* hiperparaméter) is érdemes növelni, hogy biztosan eljussunk az optimumba.

A klaszterezés jósága

- Az R^2 mutató 0 és 1 között jellemzi a klaszterezés jóságát:

$$R^2 = 1 - \frac{SSE}{SST}$$

- Itt SSE a klaszterközepektől, SST pedig az adatok főátlagától számított hibanégyzet-összegek összege minden változóra.

Hány klaszter van?

- Ha a k klaszterszám eléri a megfigyelések n számát, akkor $SSE = 0$ és $R^2 = 1$ adódik.
- Az R^2 mutató alapján nem választhatjuk ki az ideális k klaszterszámot, mert azt maximalizálva annyi klasztert kapunk, ahány megfigyelésünk van.
- Ennek semmi értelme nem lenne!

A klaszterek optimális száma

- Helyette sok más módszert kidolgoztak.
- Itt a gap-statisztikát (Tibshirani, Walther és Hastie, 2001) tanuljuk:

$$GAP = \ln SSE_0 - \ln SSE$$

- Itt SSE_0 egy olyan szimulált adatfájlon végzett középpontú klaszterezés SSE mutatója, ahol egyenletes eloszlás szerint véletlenszerűen szóródnak a pontok (nincsenek klaszterek).
- Az ideális k esetén maximális a gap-statisztika.

k-prototípus klaszterezés

- A k-középpontú klaszterezésben csak numerikus változók használhatók ☹.
- A gyakorlatban a kategorikus változók is fontosak!

k-prototípus klaszterezés

- A *k*-középpontú klaszterezésben csak numerikus változók használhatók ☹.
- A gyakorlatban a kategorikus változók is fontosak!
- A *k*-prototípus klaszterezés a *k*-középpontú klaszterezés kiterjesztése: csak numerikus változók esetén azonos eredményt adnak.
- A távolságfogalmat ki kell terjeszteni kategorikus változókra is!

Mérési szintek és távolságfogalmak

- p numerikus változó terében (euklideszi):

$$d(\mathbf{a}, \mathbf{b}) = \sqrt{\sum_{i=1}^{p} (a_i - b_i)^2}$$

- q kategorikus változó terében (matching távolság, a különböző tulajdonságok száma):

$$d(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^{q} \mathbb{1}_{\{a_i = b_i\}}$$

Vegyes távolságfogalom

- p numerikus és q kategorikus változó terében (vegyes távolság, az euklideszi és matching távolságok súlyozott összege, $\lambda > 0$):

$$d(\mathbf{a}, \mathbf{b}) = \sqrt{\sum_{i=1}^{p}(a_i - b_i)^2 + \lambda \sum_{i=p+1}^{p+q} \mathbb{1}_{\{a_i = b_i\}}}$$

Vegyes távolságfogalom

- p numerikus és q kategorikus változó terében (vegyes távolság, az euklideszi és matching távolságok súlyozott összege, $\lambda > 0$):

$$d(\mathbf{a}, \mathbf{b}) = \sqrt{\sum_{i=1}^{p} (a_i - b_i)^2 + \lambda \sum_{i=p+1}^{p+q} \mathbb{1}_{\{a_i = b_i\}}}$$

- λ becslése (~sztenderdizálás):
  - Kategorikus változó varianciája: $1 - \sum_{i=1}^{K} p_i^2$
  - A becsült érték a numerikus és kategorikus változók átlagos varianciáinak hányadosa.

k-prototípus klaszterezés

- Klaszter prototípusa (fiktív megfigyelés):
  - Numerikus változók esetén átlagos érték.
  - Kategorikus változók esetén módusz.
- *k*-középpontú klaszterezés: a megfigyelések saját (legközelebbi) középpontjaiktól mért euklideszi távolságainak négyzetösszegét minimalizáltuk.
- Az elv itt ugyanaz, csak középpont helyett prototípust, és euklideszi helyett vegyes távolságot számolunk.

Köszönöm a figyelmet!