Adatbányászat – Klaszterezés
Ellenőrző kérdések

– Mit nevezünk klaszterezésnek?
Az üzleti intelligencia szempontjából a klaszterezés egy eszköz arra, hogy a teljes adathalmazt hasonló elemeket tartalmazó részhalmazokra particionáljuk. Számos üzleti területen alkalmazzák, mint például a marketingkutatás vagy a CRM (customer relationship management). Alkalmazási példák: piacszegmentálás, alkatrészek meghibásodásanalízise, csalásdetektálás, nemfizetés-elemzés.
Alapvetően tehát a klaszterezés az információk feltárására szolgál, segítségével jól megkülönböztethető profilok alakíthatóak ki a klaszterre jellemző viselkedésminták, tulajdonságok alapján. Ezek alapján a klaszter tulajdonságaihoz illesztett cselekvési tervet lehet készíteni pl.: egy adott vásárlói csoport jobb elérése érdekében a termék portfólió átalakítása.

– Mi a k-közép klaszterezés lényege?

1.2. A K-közép klaszterezés

Az egyik legelterjedtebb K-közép algoritmus az objektumokat K klaszterbe próbálja rendezni, hogy minimalizálja az összesített klaszteren belüli varianciát, nevezetesen például a következő négyzetes hibafüggvényt

$$E = \sum_{i=1}^{K} \sum_{j \in S_i} |x_j - \mu_i|^2$$

ahol μ_i jelöli a klaszterek átlagait.

Require: adathalmaz D_n^X, klaszterszám K
Ensure: K klaszterközép
Ini: a μ_k-k véletlen vagy a priori tudáson alapuló megválasztása
repeat
minden egyes minta hozzárendelése a legközelebbi klaszterhez
a klaszterek középenek ismételt becslése
until NincsJavulás(E_{t+1},E_t,t)

Hátrányai a klaszterszám megadásának szükségessége, lokális optimumokba való beragadás, a nagy $2^{\Omega(\sqrt{n})}$ időigény rossz esetben. Robosztusabb verziója például a k-medoid, ami létező mintákat használ a klaszterek közepe helyett

Az algoritmus alapötlete valójában az úgynevezett Expectation-Maximization eljárással hozható kapcsolatba és az adatok tekintetében pedig k normális eloszlás lineáris kombinációjának a feltételezésével. Továbbá végiggondolható, hogy a k-közép, az Expectation-Maximization eljárás alkalmazása ebben az esetben, és az úgynevezett Gibbs mintavételen alapuló Monte Carlo eljárás alkalmazása ebben a feladatban egy egyre egzaktabb és több lehetőséget kínáló hierarchiát alkotnak.

MI almanach-ból:

A *k-közép klaszterezés (k-means clustering)* pontosan *k* darab kategória halmazát állítja elő. A következő elven működik:

1. Vegyünk véletlenszerűen *k* dokumentumot a *k* kategória reprezentálására.
2. Rendeljünk minden dokumentumot a legközelebbi kategóriához.
3. Számoljuk ki minden egyes kategória átlagát, és használjuk a *k* átlagot a *k* kategóriák új értékeinek reprezentálására.
4. Ismételjük a 2-es és 3-as lépéseket, amíg konvergálnak.

(http://project.mit.bme.hu/mi_almanach/books/aima/ch23s02)

– Mit nevezünk hasonlósági mértéknek s(x_k, x_j), és mi a szerepe a klaszterezésnél?

Az objektumok közt tipikusan - de nem szükségszerűen - definiált egy pozitív valósokra leképező különbözőség (dissimilarity) $d(x^{(k)}, x^{(j)})$. Ez a távolság axiómákat tipikusan nem teljesíti, sok esetben ennek „inverze" a hasonlóság érhető el $s(x^{(k)}, x^{(j)})$. Ezeknek és a klaszterbe tartozásnak a „fuzzy" vagy „valószínűségi" értelmezésére nincs szükségünk.

<**a**,**b**> = szum(i=1 - n)[a_i * b_i]
s(**a**, **b**) = <**a**,**b**> / (||**a**|| * ||**b**||)

– Mit jelent az interklaszter távolság (klaszter elválás)?
Szerintem a válasz: a klaszerek közötti távolság.

– Mit jelent az intraklaszter távolság (klaszter koherencia)?
//előző talán
Szerintem a klasztereken belüli távolságok

– Mit nevezünk klaszterek közötti homogenitásnak?
//tipp: klaszterek középpontjai közötti távolságok mennyire térnek el

– Mit nevezünk klaszteren belüli homogenitásnak?
//tipp: klaszter elemei mennyire vannak távol a középponttól
(ez és az előző sincs benne a jegyzetbe, de google is csak a beugró kérdéseket dobta ki)

– Hogyan határozható meg az optimális klaszterszám általánosan? (Lehetséges-e egyáltalán?)
K-means-nél priori információnak számít, hierarchikus klaszterezésnél a fa belső élei alapján becsül az algoritmus.
Segédletben:
A klaszterek számának megfelelő kialakítása: Nagyszámú, kisméretű klaszter valószínűleg nem hoz értelmes eredményt, ugyanúgy, ahogy egyetlen mindent magába foglaló klaszter sem. Meg kell találni az „egyensúlyt" a klaszterek száma és mérete közt. Adott problémára lehet jó megoldást adni, általánosan viszont nem.

Szóval szerintem nem, mert általánosan kérdezi.
("Ha a klaszterek száma nem ismert, egy kínálkozó módszer a hierarchikus klaszterezés használata.")

– Milyen mértékben lapolódhatnak át a klaszterek?
Semennyire, mert különben az ún. modultanuláshoz jutnánk.

– Mit értünk sziluett együttható alatt?

A „sziluett" (Silhouette) együttható egy elterjedt jellemzője a klaszterek koherenciájának (azaz a klaszteren belüli távolságoknak) és klaszterek elválásának (azaz a klaszterek között távolságoknak). A k klaszter i elemére a sziluett érték

$$sc_{ik} = \frac{b(i) - a(i)}{max(a(i), b(i))}$$

ahol $a(i)$ az i átlag különbözősége/távolsága a klaszterének az összes többi eleméhez képest és $b(i)$ pedig a legközelebbi klaszter legközelebbi elemét jelöli i-hez. Ha n_k jelöli a k klaszter méretét és n az objektumok összes számát, akkor a sziluett együttható SC_k a k. klaszterre és az összesített sziluett együttható SC a következő (-1 és 1 közötti értékkel)

– Mit jelent a SOM, és mi a lényege?
Self-Organizing Maps
Ekkor nem csupán a klaszterek számát, hanem azok topológiáját, azaz a klaszthereknek a klaszter szomszédjait is előre megadhatjuk (például egy síkbeli négyzetrács mintája szerint). A tanulás során az objektumok közül véletlenszerűen sorsolunk egyet, majd a legközelebbi klasztert reprezentáló vektort és a térkép topológiája szerinti szomszédos klaszthereknek a reprezentáns vektorait is az objektumhoz közelítjük. A megszokott felhasználása, hogy nagy dimenziós adatokat egy alacsonyabb dimenzióbeli szabályos hálóra képezzük le.