2. Tulajdonságtér

2.1. A lineáris térről

A *lineáris tér*, vagy *vektortér* halmaz, amelyben bizonyos műveletek értelmezettek, és amelynek elemeire meghatározott tulajdonságok érvényesek [1]. Szám-*n*-esek, vektorok *ilyen* elemek, ezek tehát lineáris tér elemei. Ha a térben van *n* olyan elem, amelyek a lineáris kombináció szabályai szerint más elemekből nem állíthatók elő, akkor ezek az elemek egy *n*-dimenziós tér *bázisainak* tekinthetők, és velük a tér minden további eleme előállítható. Az *n*-dimenziós térben az $\mathbf{e}_1$, $\mathbf{e}_2$,...$\mathbf{e}_n$ egymásra merőleges egységvektorok *n*-méretű derékszögű koordináta rendszert feszítenek föl, amelyben bármely **x** vektor

$$x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + ... + x_n\mathbf{e}_n = \mathbf{x} \qquad (2.1)$$

alakban eőállítva *pontként*, vagy az arra mutató nyílként ábrázolható. A pont koordinátái az egyes bázisvektorok x_1, x_2, ...x_n súlyát jelentik.

Az *n*-dimenziós teret képező elemek egy kisebb dimenziójú részhalmaza a tér *altere*, amelynek természetesen megmaradnak lineáris tér tulajdonságai.

A mátrixok oszlopai vagy sorai vektorok, amelyekkel összefüggésben felvetődik a *lineáris függetlenség* és a mátrix *rangjának* kérdése.

Egy *I* x *J* (*I* > *J*) méretű **D** mátrix oszlopai egymástól lineárisan függetlenek, ha a

$$c_1\mathbf{d}_1 + c_2\mathbf{d}_2 + ... + c_J\mathbf{d}_J = \mathbf{0} \qquad (2.2)$$

egyenlőség *nem* teljesül (kivéve a $c_1=c_2=...=c_J = 0$ triviális esetet), azaz egyik $\mathbf{d}_j$ vektor sem állítható elő a többi vektor lineáris kombinációjaként. Egy mátrix lineárisan független vektorainak számát a mátrix rangjának nevezik. Mivel a lineárisan független vektorok száma nem lehet nagyobb, mint a teret képező vektorok mérete, egy mátrix rangja sem lehet nagyobb, mint kisebbik mérete:

$$\text{rang}(\mathbf{D}) \le \min(I, J) \qquad (2.3)$$

Ha (2.3) reláció egyenlőségként teljesül, akkor a mátrixot *teljesrangúnak* szokás nevezni.

Ha **D** mátrix, mint szokásos, álló téglalap mátrix, azaz *I* › *J*, *és* nem teljesül (2.2) feltétel, akkor rangja *J* és *teljesrangú*.

A *J* darab lineáriasan független vektor *J* dimenziós lineáris teret feszít fel. Ha **D** mátrix csak *R* rangú volna, akkor a sorvektorok által kijelölt pontok a *J*-dimenziós tér *R* méretű alterében foglalnának helyet.

2.2. Tulajdonságok lineáris tere

Az *I* x *J* méretú *adatmátrix* tulajdonságvektorait tekinthetjük olyan lineáris tér elemeinek, amelyeket a *tulajdonságok* egységnyi hosszúságú vektorai feszítenek fel. Ebben a *tulajdonságtérnek* nevezett térben az *objektumok J* koordinátájú vektorai egy-egy pontot, *objektum-pontokat* jelölnek ki.

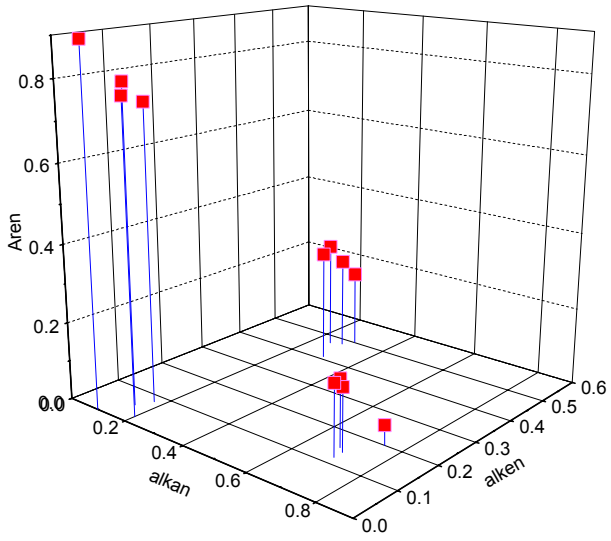
2.1 Példa: Legyen adott három benzin, darabonként négy mintában. Minden mintában megmérik az alkán, alkén és arén móltörtet. Van tehát *I* = 12 objektum és *J* = 3 tulajdonság. Az adatmátrix legyen a következő:

2.1 táblázat. Benzinösszetétel

I / J	1	2	3
1	0.7	0.12	0.18
2	0.68	0.15	0.17
3	0.75	0.2	0.05
4	0.7	0.14	0.16
5	0.3	0.5	0.20
6	0.28	0.48	0.24
7	0.3	0.41	0.29
8	0.25	0.47	0.28
9	0.15	0.05	0.80
10	0.17	0.08	0.75
11	0.1	0	0.90
12	0.2	0.02	0.78

Az adatmátrix objektum vektorainak (pontjainak) képe a háromdimenziós térben a 2.1 ábrán látható.

Az objektumpontok elhelyezkedése a tulajdonságtérben az objektumok kapcsolatát is tükrözi. Ezt az elhelyezkedést természetesen háromnál több dimenzióban nem látni, az azonban mindenképpen kézenfekvő, hogy az egymáshoz közelfekvő pontok rokon objektumokhoz tartoznak. A pontok helyzetére kis (2 - 3) dimenziójú sikra, térbe való vetületeikből lehet következtetni. Ha a pontok csomókba, “fürtökbe”, clusterekbe tömörödnek, az objektumok közös sajátságúak, ha a csomók jól elválnak, idegenek.



2.1 ábra. Objektumok a tulajdonságtérben

Tekintsünk határeseteket. Tegyük fel, hogy az objektumokat olyan, együttesen eloszló, egymással akár korreláló tulajdonságok jellemzik, amelyeknek becsülhető várható értéke, szórása, kovarianciája. Ilyen esetben a tulajdonságoknak az (1.4) összefüggésben már említett J x J méretű szimmetrikus kovariancia mátrixa **C**:

$$\mathbf{C} = \begin{bmatrix} s_{11}^2 & c_{12} & \dots & c_{1J} \\ c_{21} & s_{22}^2 & \dots & c_{2J} \\ \dots & & & \\ c_{J1} & c_{J2} & \dots & s_{JJ}^2 \end{bmatrix} \qquad (2.4)$$

ahol

$$c_{ij} = c_{ji} = \frac{\sum_k \left(d_{ki} - \bar{d}_i\right)\left(d_{kj} - \bar{d}_j\right)}{J-1} \tag{2.5}$$

az *i*-edik és *j*-edik tulajdonság becsült kovarianciája és

$$c_{ii} = s_{ii}^2 = \frac{\sum_k \left(d_{ki} - \bar{d}_i\right)^2}{J-1} \tag{2.6}$$

az *i*-edik tulajdonság becsült varianciája.

Az objektumok ebben az esetben a J-dimenziós térben a

$$\mathbf{x}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{x} = k \tag{2.7}$$

egyenletű hiperellipszoiddal burkolható térrészbe kerülnek. Ennek a (szélsőséges esetben, egyenlő szórásoknál, zérus kovarianciáknál hipergömb) alakú térrésznek vetülete valamely sikra ellipszis (szélsőséges esetben kör).

Más határesetben az objektumok a térben vagy annak vetületeiben valamely jellegzetes, pl. vonal-, síkmenti mintázat mentén helyezkednek el. Ilyenkor az objektumok tulajdonságai között függvénykapcsolat sejthető. Előfordulhat az is, hogy az objektumpontok egyenletesen, homogén módon szórják be a teret.

2.3 Objektumok távolsága a tulajdonságtérben

A csoportosítás alapja az objektumok közötti hasonlóság. Az objektumok hasonlósága *alkalmas* tulajdonságvektor definíciók (alkalmas "reprezentáció") esetén egyenértékű az objektumok tulajdonságvektorainak hasonlóságával. A tulajdonságok terében a hasonló tulajdonságvektorú objektumok (pontok) –mint láttuk– egymáshoz közel helyezkednek el, a különbözők messze. Az objektumokat képviselő pontok közötti *távolságnak* tehát a csoportosítás szempontjából döntő jelentősége van.

A távolságként definiált mennyiségeknek –fajtájuktól függően– több-kevesebb feltételnek kell eleget tenniök. A legfontosabbak:

$d_{st} \geq 0$ (2.8)

$d_{ss} = 0$ (2.9)

$d_{st} = d_{ts}$ (2.10)

$d_{st} = 0$ akkor és csak akkor, ha $s = t$ (2.11)

Távolságokat sok szempontnak megfelelően sokféleképpen lehet definiálni. A távolság elsősorban függ attól, hogy milyen természetűek az adatok: nevesítők (nominálisak), binárisak, rendezők (ordinálisak) vagy mérhetők (metrikusak).

2.3.1 Kategorikus tulajdonságok rokonsága

Nominális tulajdonságok távolságáról akkor lehet beszélni, ha azokat bináris tulajdonságúakká alakítjuk. Tegyük fel, hogy az objektumoknak *J* tulajdonsága van. Minden objektumnak megfeleltetünk egy rendezett, J elemű bináris (Boole) vektort, amelyik azon a helyen tartalmaz 1-értéket, amely az objektum adott tulajdonságának a helye, máshol zérus.

Bináris (J darab 1 vagy 0 értéket tartalmazó) vektorok hasonlóságára az elemek egyezésének és eltérésének leszámlálása alapján következtethetünk [2]. Jelölje két vektor esetén

a az 1 - 1 egyezések számát

b az 1 – 0 eltérések számát

c a 0 – 1 eltérések számát

d a 0 – 0 egyezések számát.

A négy számértékből valamely $t(a,b,c,d)$ távolságmértéket számolnak, gyakran célszerűen kerek határok közé normálva.

A 0 – 0 egyezéseket óvatosan kell kezelni. Két objektum között ugyanis nem jelent szükségképpen hasonlóságot az, hogy egyiküknek sincs meg ugyanaz az adott tulajdonsága. A *d* szám tehát csak akkor vehető figyelembe, ha a tulajdonságok távolléte rokonságot igazol.

A legegyszerűbb rokonságmértékek az (1 - 1) egyezések számát figyelik, viszonyitva valamilyen bázishoz:

$$t = \frac{a}{a+b+c} \qquad \text{(Jaccard szám)} \qquad (2.12)$$

$$t = \frac{a}{a+b+c+d} \qquad \text{(Russel-Rao szám)} \qquad (2.13)$$

$$t = \frac{2a}{2a+b+c} \qquad \text{(Sorenson szám)} \qquad (2.14)$$

$$t = \frac{a}{a+2(b+c)} \qquad \text{(Edmonston szám)} \qquad (2.15)$$

$$t = \frac{a}{b+c} \qquad \text{(Kulczinsky szám)} \qquad (2.16)$$

$$t = \frac{1}{2}\left(\frac{a}{a+b}+\frac{a}{a+c}\right) \qquad \text{(Módosított Kulczinsky szám)} \qquad (2.17)$$

Az számok 0 értékűek, ha a vektorelemek között nincs egyezés ($a = 0$), és (2.16) kivételével 1 értékűek, ha a vektorelemek között nincs eltérés ($b = 0$ és $c = 0$)

Az (1 – 1) és (0 – 0) egyezéseket is figyeli a

$$t = \frac{a+d}{a+b+c+d} \qquad \text{(Egyszerű egyezési szám)} \qquad (2.18),$$

$$t = \frac{a+d}{a+d+2(c+b)} \qquad \text{(Rogers-Tanimoto szám)} \qquad (2.19),$$

$$t = \frac{2(a+d)}{2(a+b)+c+d} \qquad \text{(Sokal-Sneath szám)} \qquad (2.20).$$

Az számok 0 értékűek, ha a vektorelemek között nincs egyezés (a = 0, és d = 0), és (2.15) kivételével 1 értékűek, ha a vektorelemek között nincs eltérés (b = 0 és c = 0).

A vektorelemek közötti eltéréseket méri a [0,1] tartományban a

$$t = \frac{b+c}{a+b+c+d} \qquad \text{(Tanimoto szám)} \qquad (2.21)$$

illetve annak négyzetgyöke. A [-1,1] tartományban jellemzi az egyezés mértékét a

$$t = \frac{ad-bc}{ad+bc} \qquad \text{(Yule szám)} \qquad (2.22),$$

amely +1, ha nincsenek eltérő elemek (bc = 0 és ad › 0), 0, ha az egyező és eltérő elempárok száma megegyezik (ad = bc), és -1, ha nincsenek egyező elemek (ad = 0 és bc › 0).

Az eltéréseket extenzív egységekben (darabszámban) méri a Hamming szám

$$t = b + c \qquad (2.23)$$

illetve annak négyzetgyöke. Ezeekt a számokat a manhattan távolság és az euklideszi távolság bináris megfelelőinek tekinthetjük.

2.3.2 Metrikus tulajdonságok rokonsága

Mérhető adatok esetén egyik leggyakrabban használt távolságmérték az *euklideszi távolság*:

$$t_{st} = \sqrt{\sum_j \left(x_{sj} - x_{tj}\right)^2} = \sqrt{\left(\mathbf{x}_s - \mathbf{x}_t\right)^T \left(\mathbf{x}_s - \mathbf{x}_t\right)} \qquad (2.24)$$

A távolságot jobban kiemeli az *euklideszi távolság négyzete*, t_{st}^2.

Gyakran használatos a *manhattan távolság,* amely két objektum négyzetrács mentén megtehető útjának hossza:

$$t_{st} = \sum_j \left| x_{sj} - x_{tj} \right| \qquad (2.25)$$

A távolságfogalmak definiálása során gyakran feltételezik, hogy az objektumok olyan sokaságokhoz tartoznak, amelyeknél megadható a tulajdonságok szórása és kovarianciájuk, igy a távolság definiciójában célszerűen felléphet az **S** szórás- ill. **C** kovariancia mátrix. A távolságot, más szóval, hsznos lehet az objektumok kovariancia ellipszoidjai méreteit felhasználva definiálni.

Használatos emiatt az euklideszi távolság olyan változata, a *Pearson távolság*: amely az egyes objektumok tulajdonságainak különbségeit az adott tulajdonságok szórásához viszonyítja, azokkal “normálja”:

$$t_{st} = \sqrt{\left(\mathbf{x}_s - \mathbf{x}_t\right)^T \mathbf{S}^{-1} \left(\mathbf{x}_s - \mathbf{x}_t\right)} = \sqrt{\sum_j \frac{\left(x_{sj} - x_{tj}\right)^2}{s_j^2}} \qquad (2.26)$$

Itt $\mathbf{S}^{-1}$ a variancia mátrix inverze, a szórásnégyzetek reciprokait tartalmazó J x J méretű diagonális mátrix. A Pearson távolság dimenziómentes. Az egyes objektumok távolsága nagyítható, ha a *Pearson távolság* d_{st}^2 *négyzetét* használják.

A tulajdonságok közötti korrelációt is figyelembeveszi a *Mahalanobis távolság,* amely tehát a pontfelhőt burkoló hiperellipszoid valamennyi méretadataival dolgozik:

$$t_{st} = \sqrt{\left(\mathbf{x}_s - \mathbf{x}_t\right)^T \mathbf{C}^{-1} \left(\mathbf{x}_s - \mathbf{x}_t\right)} \qquad (2.27)$$

Itt **C** a tulajdonságok már említett *J* x *J* méretű kovariancia mátrixa.

2.3.4 A távolságmátrix

Az objektumok közötti távolságok tárolására vezették be a *távolságmátrixot*, amely *I* objektumot tartalmazó rendszer esetén egy *I x I* méretű, adott esetben nagy mátrix. A mátrix *s*-edik sorának és *t*-edik oszlopának kereszteződésében az *s* és *t* objektumok t_{st} távolságát tartalmazza. A távolságok matematikai kritériumaiból következik, de könnyen be is látható, hogy a távolságmátrix szimmetrikus, és annak átlós elemei zérusok:

$$\mathbf{T} = \begin{bmatrix} 0 & t_{12} \ldots t_{1I} \\ t_{12} & 0 \ldots t_{2I} \\ \ldots & \\ t_{1I} & t_{2I} \ldots 0 \end{bmatrix} \qquad (2.28)$$

Emiatt a távolságmátrixnak csak I (I – 1)/ 2 elemét kell tárolni:

2.4 Objektumok csoportjainak távolsága a tulajdonságtérben

Objektumok, mint ismeretes, *csoportokba*, *clusterekbe* tömörülhetnek. Ezeknek a csoportoknak felismerése, megkülönböztetése, jellemzése, egyes objektumok besorolása csoportokba, ez a csoportosítás és osztályozás feladata. Felvetődik ennek során az a kérdés, hogyan lehet *csoportok* egymástól való távolságát számítani. Melyek a csoportok azon *pontjai*, amelyek közé az ismert (euklideszi, manhattan stb) távolságokat be kell fektetni.

Több lehetőség közül lehet egy adott feladathoz a leginkább illőt kiválasztani. A csoportegyesítés alapjaként leginkább használatos távolságokat a 2.2 táblázat tartalmazza:

2.2 táblázat. Csoporttávolságok

szám	megnevezés	english term	geometriai tartalom
1	egyszerű lánc	(simple linkage)	a csoportok *legközelebbi* elemeinek távolsága
2	teljes lánc	(complete linkage)	a csoportok *legtávolabbi* elemeinek távolsága
3	átlag távolság	(average linkage)	az egyesítendő csoportok elemei közötti *távolságok átlaga*
4	súlypont	(centroid linkage)	a csoportok *súlyponjainak* távolsága
5	McQuitty távolság	(McQuitty linkage)	az egyesített csoport távolsága: a két egyesített csoport távolságának *átlaga*
6	medián távolság	(median linkage)	a csoport *mediánok* távolsága
7	Ward távolság	(Ward linkage)	azon számított távolság, amely biztosítja, hogy az egyesitett csoport csoportonbelüli *eltérésnégyzete minimális* legyen.

A táblázat *súlyponton* (*centroid, barycenter)* a csoporthoz tartozó objektumoknak a tulajdonságtér origójától mért távolságai (az objektum sorvektorok) átlagát képviselő pontot (*J* elemű vektort) értjük:

$$\mathbf{c}_g = \frac{1}{n_g}\sum_{i=1}^{n_g}\mathbf{x}_i^{(g)} \qquad (2.29)$$

ahol n_g : a *g*-edik csoport objektumainak száma.

Csoport *medián* annak az objektumnak helye a térben, amely objektumnál a csoporthoz tartozó objektumok felének hosszabbak, felének rövidebbek a tulajdonságtér origójától mért távolságai. (Páros számú objektumnál a két középső objektumhossz átlaga). A távolság mindkét esetben valamelyik választott távolságváltozat.

(annak az objektumnak *J* elemű tulajdonság vektora), amelynél (páratlan elemszámnál) az objektumokvektorok fele kisebb, fele nagyobb, páros elemszámnál a két középső objektum átlagvektora.

A csoportok közötti távolság számításához vezessük be a 2.3 táblázatban megadott jelöléseket.

2.3 táblázat Jelölések csoporttávolság számításához

t	az egyik egyesítendõ csoport jele,
s	a másik egyesítendõ csoport jele,
k	az egyesített csoport jele,
n_s és n_t	az egyik illetve másik csoport objektumainek száma,
t_{si}	az *s* csoport távolsága valamely *i* objektumtól (a távolság mátrix *s,i* eleme)
t_{ti}	a *t* csoport távolsága valamely *i* objektumtól (a távolság mátrix *t,i* eleme)
t_{st}	az *s* és *t* csoportok távolsága egymástól, (a távolság mátrix *s,t* eleme)
t_{ki}	a *k* csoport távolsága valamely *i* objektumtól (a távolság mátrix *k,i* eleme)

Ezekből a mennyiségekből t_{ki}, a csoportok távolsága (az újabb távolságmátrix eleme) az egyes választásoknál a következő képletekkel számítható:

egyszerű lánc, single (minimum) linkage

$$t_{ki} = 0.5\, t_{si} + 0.5\, t_i - 0.5\, |t_{si} - t_{ti}| \qquad (2.29)$$

teljes lánc, complete (maximum) linkage

$$t_{ki} = 0.5\, t_{si} + 0.5\, t_{ti} + 0.5\, |t_{si} - t_{ti}| \qquad (2.30)$$

csoportátlag, group average linkage

$$t_{ki} = \frac{n_s}{n_s + n_t} t_{si} + \frac{n_t}{n_s + n_t} t_{ti} \qquad (2.31)$$

egyszerű átlag, McQuitty linkage

$$t_{ki} = 0.5\, t_{si} + 0.5\, t_{ti} \qquad (2.32)$$

súlyponti, centroid linkage

$$t_{ki} = \frac{n_s}{n_s + n_t} t_{si} + \frac{n_t}{n_s + n_t} t_{ti} - \frac{n_s n_t}{(n_s + n_t)^2} t_{st} \qquad (2.33)$$

medián, median (weighted centroid) linkage

$$t_{ki} = 0.5\, t_{si} + 0.5\, t_{ti} - 0.25\, t_{st} \qquad (2.34)$$

Ward féle, Ward linkage

$$t_{ki} = -\frac{n_s + n_i}{n_s + n_t + n_i} t_{si} + \frac{n_t + n_i}{n_s + n_t + n_i} t_{ti} - \frac{n_i}{n_s + n_t + n_i} t_{st} \qquad (2.35)$$

A csoportok közötti távolságoknak az objektumok csoportosításánál (cluster analysis) és osztályozásánál (classification) lesz jelentősége. Előrebocsátva, a módszerekről nagyjából most az mondható, hogy az 1) távolság használata lazán összetartozó elemű nagy csoportokhoz vezet, a 2) módszer jól elváló, kicsiny, kerek csoportokat ad, a 3) mód hajlik arra, hogy kilógó értékeket elkülönült csoportokba osszon, a 4) távolság nem kedvez jelentéktelen csoportoknak, a 6) egyesítési módszer megőrzi a kis csoportokat, ugyanúgy, mint az 5) McQuitty eljárás, amely egyenlő súlyt ad minden csoport-nak. A 7) Ward módszer nyilvánvalóan törekszik összetartozó elemek tömör csoportosításá-ra. A súlyponti és Ward egyesítésnél a négyzetes távolságmértékek használata ajánlott.

Irodalom

[1] Rózsa Pál: Lineáris algebra és alkalmazásai. III. kiadás Tankönyvkiadó, Budapest, 1991.

[2] I.E. Frank, R. Todeschini: The data analysis handbook. Elsevier, Amsterdam etc. 1994.