Adatbányászat oktatási segédlet

A segédletben található esetleges hibákkal kapcsolatos visszajelzéseket szívesen veszem.

1. gyakorlat

1.1. feladat

Bonferroni-elv: Gyanúsnak definiálunk egy vásárlói párost, ha egységnyi idő (pl. 1 év) alatt a páros mindkét tagja pontosan egyező termékeket kíván megvásárolni a piacon.[1] A piacon t termék kapható, melyből egy alkalommal v darabot vásárol egy vevő. Amennyiben m darab vásárló fejenként n darab vásárlását figyeljük meg, várhatóan hány gyanús párost fogunk találni?

2. gyakorlat

2.1. feladat

Adott elemek egy n-elemű U univerzuma, melyből véletlenszerűen kiválasztjuk annak m-elemű S és T részhalmazait.

a) Mekkora lesz $|S \cap T|$ várható értéke?
b) Mekkora lesz $|S \cup T|$ várható értéke?
c) Mekkora lesz a Jaccard hasonlóság várható értéke?

2.2. feladat

Elem	S_1	S_2	S_3	S_4
0	0	1	0	1
1	0	1	0	0
2	1	0	0	1
3	0	0	1	0
4	0	0	1	1
5	1	0	0	0

a) $h_1(x) = 5x + 2 \mod 6$ és $h_2(x) = 2x + 1 \mod 6$ hasítófüggvények használata mellett határozd meg az adatpontok minhash lenyomatait!

[1] A gyanússág már onnantól kezdve értendő, hogy két ember **egy alkalommal** ugyanazt a v terméket vásárolja meg, nem kell, hogy mind az n darab vásárlásuk tételről-tételre megegyezzen.

b) Egyformán hasznosnak tűnik mindkettő hasítófüggvény?
c) Mekkora az S_1 és S_4 pontok tényleges és a minhash lenyomataik alapján becsült Jaccard hasonlósága?

2.3. feladat

Lássuk be, hogy amennyiben két halmaz metszete m méretű, szimmetrikus differenciájuk n nagyságú (továbbá a pluszpontot érő verzióban: komplementereik metszete pedig k nagyságú), úgy a karakterisztikus mátrix sorainak (összes!) lehetséges permutációja mellett kiszámított minhash értékek egyezésének aránya éppen a két halmaz Jaccard hasonlóságát adja!

2.4. feladat

Lássuk be, hogy a szimmetrikus differencia halmazművelet segítségével definiált távolság eleget tesz a távolságmetrikák axiómarendszerének!

2.5. feladat

Amennyiben az adatpontjainkhoz rendelt k-hosszú minhash lenyomatainkat b darab egyenként s soros minhash lenyomatok blokkjaiként képzeljük el, mekkora valószínűséggel fog két egymással $0,8$ Jaccard hasonlóságot mutató pont ugyanabba a kosárba esni az LSH algoritmus végrehajtása során? (Feltéve, hogy két pont azonos kosárba képződésének feltétele az, hogy legyen legalább egy blokk, amelyen a két pont mind az s minhash értéke teljesen megegyezik.)

3. gyakorlat

3.1. feladat

Egy gyár, amely 2 terméket gyárt a $C(x, y) = 6x^2 + 12y^2$ költségfüggvénnyel dolgozik. Hogyan határozza meg a következő időszakra vonatkozó termelését a gyár, ha a két (helyettesítő)termék együttes kereslete előre ismert (180 egység), továbbá a két terméktípus eladási árai megegyeznek?

3.2. feladat

Lagrange szorzók használatával lássuk be, hogy az entrópia függvény maximuma egy bináris jellemző esetében a (0.5, 0.5) pontban van!

3.3. feladat

Lagrange multiplikátorokat használva határozzuk meg annak a k kerületű szimmetrikus trapéznak az oldalhosszait, mely maximális területtel rendelkezik!

3.4. feladat

Számold ki az x=[3, 4, 5, 6], y =[4, 3, 2, 1] pontok koszinusz hasonlóságát és távolságát!
A $v_1 = [1, -1, 1, 1], v_2 = [-1, 1, -1, 1]$ és a $v_3 = [1, 1, -1, -1]$ irányokba történő véletlen projekciók alapján mekkorának becsülnénk a 2 vektor által bezárt szög nagyságát?
Hogy változik a becslésünk, amennyiben az összes lehetséges módon elvégezzük a 2 vektor projekcióját? (érdemes lehet használni a generateAllBinaryOutcomes.m fájlt, ami az összes szóba jöhető módon létrehozza a +/-1 értékekből álló vektorokból képzett mátrixot)

5. gyakorlat

5.1. feladat

Ha d termékünk van, akkor belőlük hány $A \to B$ típusú ($A \cap B = \emptyset, A, B \neq \emptyset$) asszociációs szabály generálható le?

5.2. feladat

Egy tranzakciós adatbázisban a termékek $T = \{t_1, t_2, \ldots t_{10}\}$ halmazának t_i eleme a többi elemtől függetlenül $p(t_i) = i^{-1}$ valószínűséggel fordul elő egy vásárlói kosárban.
a) 1%-os relatív gyakorisági küszöbértéket alkalmazva, mely elemhalmazok lesznek gyakoriak?
b) Melyek lesznek az érdekes szabályok?

5.3. feladat

Tfh. a tranzakciós adatbázisunkban 100 termékünk és ugyanennyi kosarunk van, és i termék akkor található meg b kosárban, ha i osztója b-nek.
a) Mekkora lesz az összes kosárban megtalálható termékek száma?
b) 5-ös támogatottsági küszöbérték mellet mely termékek lesznek gyakoriak?

c) 5-ös támogatottsági küszöbérték mellet mely kételemű termékhalmazok lesznek gyakoriak?
d) 5-ös támogatottsági küszöbérték mellet mely n-elemű termékhalmazok lesznek gyakoriak?
e) Mekkora lesz az $\{5, 7\} \to 2$, illetve a $\{2, 3, 4\} \to 5$ szabályok bizonyossága?
f) Melyik kosárban lesz a legtöbb elem?
g) Hogy írhatók le azok a szabályok, amelyek bizonyossága 1,0?

5.4. feladat

Tfh. a tranzakciós adatbázisunkban 100 termékünk és ugyanennyi kosarunk van, és i termék akkor található meg b kosárban, ha b osztója i-nek. A 13. kosár tehát pl. a {13, 26, 39, 52, 65, 78, 92} termékeket fogja tartalmazni.
a) Mekkora lesz az összes kosárban megtalálható termékek száma?
b) 5-ös támogatottsági küszöbérték mellet mely termékek lesznek gyakoriak?
c) 5-ös támogatottsági küszöbérték mellet mely kételemű termékhalmazok lesznek gyakoriak?
d) 5-ös támogatottsági küszöbérték mellet mely n-elemű termékhalmazok lesznek gyakoriak?
e) Mekkora lesz az $\{5, 7\} \to 2$, a $\{2, 3, 4\} \to 5$, illetve a $\{24, 60\} \to 8$ szabályok bizonyossága?
f) Melyik kosárban lesz a legtöbb elem?
g) Hogy írhatók le azok a szabályok, amelyek bizonyossága 1,0?

6. gyakorlat

6.1. feladat

Mi lesz az $\{1, 2, 3\}$ csúcsok rangja abban a G=(E,V) gráfban, melyben a következő élek találhatók?
$E = \{\{1 \to 1\}, \{1 \to 2\}, \{1 \to 3\}, \{2 \to 1\}, \{2 \to 3\}, \{3 \to 2\}, \{3 \to 3\}\}$
Irredúcibilis, és aperiodikus-e a gráf?

6.2. feladat

Igazoljuk, hogy a sztochasztikus mátrixok legnagyobb sajátértéke 1,0!

6.3. feladat

Vegyük azt a gráfot, amely egy k-klikkből áll, továbbá egy pontból, amelyre minden klikkbéli csúcsból vezet él.
Hogy fog kinézni az átmenetmátrix? Igazoljuk, hogy a sorok (többségének) összege 1 lesz!
Adjuk meg az egyes pontok rangát beta és k függvényében?
Mit tapasztalunk a rangokkal kapcsolatban hosszú távon?

6.4. feladat

Link farmok - Álljon a web n oldalból, amelyből legyen m « n oldal a mi kezünkben. Az m oldal mindegyike kölcsönös lineket alakít ki egy t oldallal, amit a valóságosnál jobb színben szeretnénk feltüntetni. Legyen x azon (m ponton túli) oldalak (pl. blogok, fórumok) rangjának összege, amelyeken el tudunk helyezni t oldalra mutató linket. Mi lesz rang(t)?
a) Mi lesz az m támogató oldal rangja?
b) Hogy áll ebből össze rang(t)?

6.5. feladat

Lássuk be, hogy amennyiben a véletlen bolyongást azonos valószínűséggel kezdjük a G=(V,E) gráf bármely pontjából, melyre $V = \{1, 2, 3, 4\}$, továbbá $E = \{\{1 \to 2\}, \{1 \to 3\}, \{1 \to 4\}, \{2 \to 1\}, \{2 \to 4\}, \{3 \to 1\}, \{4 \to 2\}, \{4 \to 3\}\}$, abban az esetben a 2,3, valamint 4. pontra vonatkozó rangok szükségszerűen meg fognak egyezni a hatványiteráció minden egyes iterációjában (az egyes iterációk között persze különböznek értékeik a konvergenciáig)!

7. gyakorlat

7.1. feladat

A $G = (\{1, 2, 3, 4\}, \{(1, 2), (1, 3), (3, 2), (3, 4), (4, 1)\})$ gráfon hajtsuk végre a hubs and authorities algoritmust!

7.2. feladat

Hajtsd végre a hierarchikus klaszterezést a $A = \binom{4}{8,5}, B = \binom{4}{10}, C = \binom{6}{8}, D = \binom{7}{10}$ pontokon úgy, hogy két klaszter összevonása a klaszterek közötti

a) legközelebbi pontpárok
b) klaszterek átlagos távolsága
alapján történjen, az összevont klasztereket pedig centroidjaikkal reprezentáljuk.

7.3. feladat

$x = (0, 0); y = (10, 10), a = (1, 6); b = (3, 7); c = (4, 3); d = (7, 7), e = (8, 2); f = (9, 5)$. x és y már klaszterközepek és kell még 2-t válasszunk (k=4). Melyek legyenek azok?

7.4. feladat

Adott klaszterközepek mellett lehetséges-e, hogy egy x pontot L_i és L_j normák szerint különböző klaszterekbe soroljuk?

8. gyakorlat

8.1. feladat

Egy klaszterünk a $(7, 5), (2, 2), (3, 8)$ pontokból áll. Hogyan bontanánk szét 2 részre, ha az a célunk, hogy a keletkező klasztereken belüli négyzetes hibaösszegeket szeretnénk mindeközben minimalizálni?

8.2. feladat

Az $E = \{(1, 2), (2, 3), (2, 4), (3, 4), (3, 5), (4, 6), (5, 6), (5, 7), (5, 8), (6, 7), (7, 8)\}$ élek által meghatározott gráfon hajtsuk végre a spektrálklaszterezést! Mekkora lesz a spektrálklaszterezés által javasolt particionáláshoz tartozó vágás, illetve normalizált vágás értéke?

8.3. feladat

Hajtsuk végre a Grivan-Newman algortimust az alábbi élek által meghatározott gráfon. $E = \{(A, B), (A, C), (B, C), (B, D), (D, E), (D, F), (D, G), (E, F), (G, F)\}$!

9. gyakorlat

9.1. feladat

Adott 3 bináris véletlen változó (J, B és R), melyek rendszere egy Bayes-hálózatot alkot.
A Bayes-hálózat (feltételes) valószínűségei a következők szerint alakulnak.

$P(B = i|R = i) = P(B = i|R = h) = P(B = i) = 0,8$

$P(R = i|B = i) = P(R = i|B = h) = P(R = i) = 0,2$

b	r	P(J=i|B=b,R=r)
i	i	1,0
i	h	0,9
h	i	0,8
h	h	0,2

Feltétlesen független-e B változó R-től J változó értékének ismeretében?

9.2. feladat

Az Y osztályváltozó együttállása az M magyarázó változóval a következők szerint alakul?

	Y=+	Y=-	
M	40	5	45
$\neg M$	20	35	55
	60	40	100

a) Mi az osztálycímkére vonatkozó entrópia értéke?
b) Mi az osztálycímkére és az M változóra vonatkozó feltételes, valamint együttes entrópia értéke?
c) Mekkora az M és Y változók kölcsönös információtartalmának értéke?
d) Milyen χ^2 mutató tartozik az M változóhoz?

Megoldások

1.1

P(X és Y teljesen azonos vásárlást hajt végre) $= \binom{t}{v}^{-1} \approx \frac{v!}{t^v}$ (, ha $v \ll t$)

Lehetséges váráslás-vásárló párosok száma $= n^2\binom{m}{2} \approx \frac{(nm)^2}{2}$

Ahonnan a gyanús párosok száma $\approx \frac{(nm)^2}{2}\frac{v!}{t^v}$

2.1

a) 1. m.o.: Z véletlen változó $= |S \cap T|$

$$E[Z] = \sum_{i=1}^{m} P(|S \cap T| = i)i = \sum_{i=1}^{m} \frac{\binom{m}{m-i}\binom{n-m}{m-i}}{\binom{n}{m}} i$$

$$E[Z] = \frac{((n-m)!m!)^2}{n!} \sum_{i=1}^{m} \frac{1}{(i-1)!(m-i)!^2(n-2m+i)!}$$

Megjegyzések:
- számláló első tagja: az m kiválasztott elemből $m-i$ lecserélése szükséges ahhoz, hogy a metszet mérete i legyen
- számláló első tagja: $n-m$ lehetséges helyettesből $m-i$-t használunk föl

2. m.o.:

$$X_i = \begin{cases} 1, \text{ ha i. pont a két halmaz metszetében található} \\ 0, \text{ különben} \end{cases}$$

$$E[Z] = E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} P(X_i = 1) \cdot 1 + P(X_i = 0) \cdot 0 = n\Big(\frac{m}{n}\Big)^2 = \frac{m^2}{n}$$

b) 1. m.o.: Y véletlen változó $= |S \cup T|$

$$E[Y] = \sum_{i=1}^{m} P(|S \cap T| = i)(2m-i) = \sum_{i=1}^{m} \frac{\binom{m}{m-i}\binom{n-m}{m-i}}{\binom{n}{m}} (2m-i)$$

2. m.o.:

$$X_i = \begin{cases} 1, \text{ ha i. pont a két halmaz uniójában található} \\ 0, \text{ különben} \end{cases}$$

$$E[Y] = E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} P(X_i = 1) \cdot 1 + P(X_i = 0) \cdot 0 =$$

$$= n\Big(1 - \Big(1 - \frac{m}{n}\Big)^2\Big) = n\Big(1 - \frac{n^2 - 2nm + m^2}{n^2}\Big) = \frac{m(2n-m)}{n}$$

c) 1. m.o.:

$$E[Jaccard] = \sum_{i=1}^{m} P(|S \cap T| = i)\frac{i}{2m-i} = \sum_{i=1}^{m} \frac{\binom{m}{m-i}\binom{n-m}{m-i}}{\binom{n}{m}}\frac{i}{2m-i}$$

2. m.o.: Jó is lenne, ha az működne, hogy vesszük a Jaccard együttható számlálójának és nevezőjének várható értékét, de ahogy azt a gyakorlaton is vett példa jól mutatta, ez sajnos nincs így.

2.2

a)

Elem	S_1	S_2	S_3	S_4	$h_1(x) = 5x + 2 \mod 6$	$h_2(x) = 2x + 1 \mod 6$
0	0	1	0	1	2	1
1	0	1	0	0	1	3
2	1	0	0	1	0	5
3	0	0	1	0	5	1
4	0	0	1	1	4	3
5	1	0	0	0	3	5

	S_1	S_2	S_3	S_4
h_1	0	1	4	0
h_2	5	1	1	1

b) $h_1(x)$ hasítófüggvény egyenletesebben "szórja szét" az elemek eredeti sorszámait, így az értelmesebbnek mondható.

c) Tényleges és becsült $sim_{Jaccard}(S_1, S_4)$ értékek: $\frac{1}{4}$, illetve $0,5$.

2.3

$$P(\text{egyezik a minhash értékük}) = \frac{\frac{(m+n-1)!}{(m-1)!n!}}{\frac{(m+n)!}{m!n!}} = \frac{(m+n-1)!}{(m-1)!n!}\frac{m!n!}{(m+n)!} = \frac{m}{m+n}$$

2.4

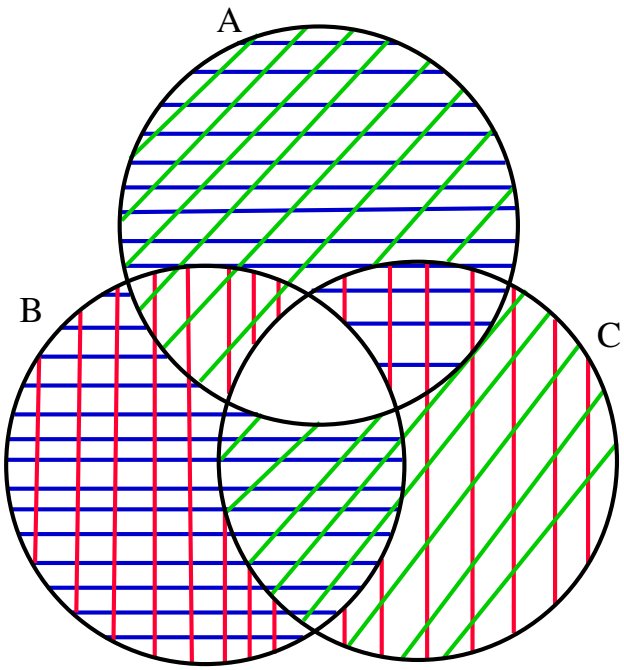
$d(A, B) = |A \triangle B| = |A \setminus B| + |B \setminus A| = |A| + |B| - 2|A \cap B|$

a) nemnegatív, mivel $2|A \cap B| \leq |A| + |B|$, illetve már csak azért is, mert a szimmetrikus differencia műveletének végrehajtása egy halmazt eredményez, amelyekről tudjuk, hogy csak valamilyen nemnegatív egész elemet tartalmazhatnak.

b) Pozitív definitség: $d(A, B) = 0 \Leftrightarrow A = B$
$\Rightarrow d(A, B) = 0 \Leftrightarrow |A| + |B| - 2|A \cap B| = 0 \Leftrightarrow |A \cup B| + |A \cap B| = 2|A \cap B| \Leftrightarrow$
$|A \cup B| = |A \cap B| \Leftrightarrow A = B$
$\Leftarrow A = B \Leftrightarrow d(A, A) = |A| + |A| - 2|A \cap A| = 0$

c) Szimmetria: melyik halmazművelet alapján is határozzuk meg a távolságot? ☺

d) Háromszög egyenlőtlenség: szemléltető ábra
Nincs vízszintes kék vonal, ami ne lenne áthúzva keresztbe egy másik vonallal.



Q.E.D. ■

2.5

P(x és y ugyanabban a kosárban köt ki$|sim_{Jaccard}(x,y) = 0,8) = 1 - (1 - 0,8^s)^b$

3.1

$L(x, y, \lambda) = 6x^2 + 12y^2 - \lambda(x + y - 180)$

$\nabla L(x, y, \lambda) = \vec{0}$ megoldásában lehet a maximum, ami a $(120, 60, 1440)$ pontban van.

3.2

A bináris véletlen változók entrópiáját ($H(p_1, p_2) = -p_1 \log_2 p_1 - p_2 \log_2 p_2$) maximalizáló argumentumait szeretnénk meghatározni amellett a megkötés mellett, hogy $p_1 + p_2 = 1$ teljesüljön. $L(p_1, p_2, \lambda) = -p_1 \log_2 p_1 - p_2 \log_2 p_2 - \lambda(p_1 + p_2 - 1)$

$$\frac{\partial L}{\partial p_1} = -\log_2 p_1 - \frac{1}{\ln 2} - \lambda = 0 \tag{1}$$

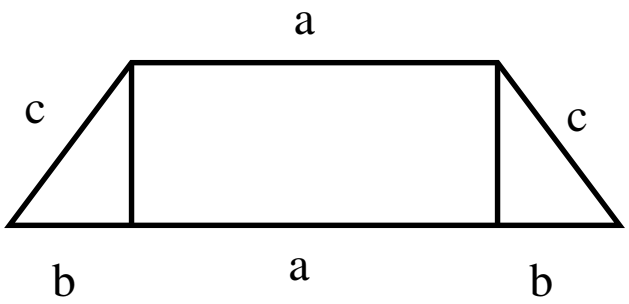
$$\frac{\partial L}{\partial p_2} = -\log_2 p_2 - \frac{1}{\ln 2} - \lambda = 0 \tag{2}$$

$$\frac{\partial L}{\partial \lambda} = -p_1 - p_2 + 1 = 0 \tag{3}$$

$$(1) \wedge (2) \Rightarrow p_1 = p_2 \tag{4}$$

$(3) \wedge (4) \Rightarrow p_1 = p_2 = 0,5$

3.3



$$f(a,b,c) = \frac{a+(a+2b)}{2}\sqrt{c^2-b^2} = (a+b)\sqrt{c^2-b^2}$$
$$\text{f.h. } g(a,b,c) = 2(a+b+c) - k = a+b+c-\tfrac{k}{2}$$

Ebből a Lagrange függvényünk tehát,

$$L(a,b,c,\lambda) = a\sqrt{c^2-b^2} + b\sqrt{c^2-b^2} - \lambda a - \lambda b - \lambda c + \lambda\frac{k}{2} \tag{5}$$

, amely gradiense

$$\nabla L(a,b,c,\lambda) = \begin{bmatrix} \sqrt{c^2-b^2} - \lambda \\ \frac{-ab}{\sqrt{c^2-b^2}} + \sqrt{c^2-b^2} - \frac{b^2}{\sqrt{c^2-b^2}} - \lambda \\ \frac{ac}{\sqrt{c^2-b^2}} + \frac{bc}{\sqrt{c^2-b^2}} - \lambda \\ -a-b-c+\frac{k}{2} \end{bmatrix}.$$

Mivel $\nabla L(a,b,c,\lambda) = \vec{0}$ megoldását keressük, $\frac{\partial L}{\partial a} = 0$-ból adódik $\sqrt{c^2-b^2} = \lambda$. Ezt fölhasználva $\frac{\partial L}{\partial b} = 0$-ban azt kapjuk, hogy $\frac{-b(a+b)}{\sqrt{c^2-b^2}} = 0$, ami a és b nemnegativitásából következően csak akkor teljesülhet, ha $b = 0$. Az eddigieket fölhasználva $\frac{\partial L}{\partial c} = 0$ miatt $a = c$-nek teljesülnie kell. Az eddigiek fényében a korlátozó feltétel akkor kerül kielégítésre, ha $a = c = k/4$, valamint $b = 0$. A k kerületű szabályos trapézok közül tehát annak a területe maximális, amely valójában nem más, mint egy $k/4$ oldalhosszúságú négyzet.

3.4

$sim_{cos}(x,y) \approx 0,7875 \Rightarrow d_{cos}(x,y) \approx 0,664$ ($\arccos -t$ használva $\Rightarrow \approx 38,05°$

A v_1, v_2, v_3 vektorokra történő projekció alapján a becsült koszinusz hasonlóság mértéke: $1/3 \Rightarrow 120°$-ra becsüljük a két vektor által bezárt szöget
$\forall v \in \{-1,1\}^4$-ra történő projekció alapján a becsült koszinusz hasonlóság mértéke: $12/16 = 0,75 \Rightarrow 45°$-ra becsüljük a két vektor által bezárt szöget

5.1

$$|R| = \sum_{k=1}^{d}\binom{d}{k}\sum_{i=1}^{d-k}\binom{d-k}{i} = \sum_{k=1}^{d}\binom{d}{k}(2^{d-k}-1) = \sum_{k=1}^{d}\binom{d}{k}2^{d-k} - (2^d-1)$$

, ami pedig $(1+x)^d = \sum_{j=1}^{d} \binom{d}{j} x^{d-j} + 2^d$ miatt $|R| = 3^d - 2^d - 2^d + 1 = 3^d - 2^{d+1} + 1$

5.2

a)

$$\left\{ X \in \mathcal{P}(T) : \prod_{t_j \in X} p(t_j) \geq 0,01 \right\}$$

halmazok lesznek gyakoriak, ahol

$$\mathcal{P}(T)$$

a termékek T halmazának hatványhalmazát jelöli
b) p(a,b) = p(a)*p(b) miatt egyik asszociációs szabály sem tud érdekes lenni

5.3

A megoldásokban $\mathcal{P}(T)$ a termékek T halmazának hatványhalmazát, míg $LKT(\cdot)$ a legkisebb közös többszöröst meghatározó műveletet jelöli.
a) 482.
b) Az 1–20. termékek.
c) $\{(i,j) : LKT(i,j) \leq 20\}$
d) $\{X \in \mathcal{P}(T) : LKT(X) \leq 20\}$
e) $c(\{5,7\} \to 2) = \frac{1}{2}$ és $c(\{2,3,4\} \to 5) = \frac{1}{8}$
f) A 60., 72., 84., 90. és 96. kosarakban lesz a legtöbb (egyenként 12) elem.
g) $c(A \to B) = 1,0 \Leftrightarrow s(A,B) = s(A) \Leftrightarrow LKT(A,B) = LKT(A)$

5.4

a) 482.
b) Azok a termékek, amelyek sorszámának legalább 5 osztója van.
c) Azok a termékpárok, amelyek sorszámainak legalább 5 közös osztójuk van.
d) Azok a termék-n-esek, amelyek sorszámainak legalább 5 közös osztójuk van.
e) $c(\{5,7\} \to 2) = 1,0$, $c(\{2,3,4\} \to 5) = 1,0$ és $c(\{24,60\} \to 8) = \frac{3}{6}$
f) Az első kosárban lesz a legtöbb (100) elem.
g) $c(A \to B) = 1,0 \Leftrightarrow \{A \text{ osztói}\} \subseteq \{B \text{ osztói}\}$

6.1

$$\begin{pmatrix} 0,23077 \\ 0,30769 \\ 0,46154 \end{pmatrix}$$

6.2

Belátható, hogy a sztochasztikus mátrixok rendelkeznek 1-es sajátértékkel (mivel sorösszegeik azonosan 1-nek adódnak). Tegyük fel, hogy egy A sztochasztikus mátrixnak létezik $\lambda > 1$ sajátértéke, vagyis $Ax = \lambda x$, valamely $\lambda > 1$-re, ami

azt (is) eredményezi, hogy x vektor legnagyobb értékű x_i eleme az A-val történő szorzás eredményeképp megnő. Ismerve azonban, hogy λx értékei az x-ből képzett konvex kombinációk eredményeiként állnak elő, mindez nem lehetséges, így ellentmondásra jutunk, vagyis sztochasztikus mátrixok nem rendelkeznek 1-et meghaladó sajátértékkel.

6.3

Az átmenetmátrix (teleportálás lehetősége nélkül):

$$T = \begin{bmatrix} 0 & \frac{1}{k} & \frac{1}{k} & \dots & \frac{1}{k} & \frac{1}{k} \\ \frac{1}{k} & 0 & \frac{1}{k} & \dots & \frac{1}{k} & \frac{1}{k} \\ \vdots & \frac{1}{k} & \frac{1}{k} & \ddots & \frac{1}{k} & \frac{1}{k} \\ \frac{1}{k} & \frac{1}{k} & \frac{1}{k} & \frac{1}{k} & 0 & \frac{1}{k} \\ 0 & 0 & 0 & \dots & 0 & 0 \end{bmatrix}$$

A teleportálással kiegészített átmenetmátrix:

$$T_\beta = \begin{bmatrix} \frac{1-\beta}{k+1} & \frac{\beta}{k}+\frac{1-\beta}{k+1} & \frac{\beta}{k}+\frac{1-\beta}{k+1} & \dots & \frac{\beta}{k}+\frac{1-\beta}{k+1} & \frac{\beta}{k}+\frac{1-\beta}{k+1} \\ \frac{\beta}{k}+\frac{1-\beta}{k+1} & \frac{1-\beta}{k+1} & \frac{\beta}{k}+\frac{1-\beta}{k+1} & \dots & \frac{\beta}{k}+\frac{1-\beta}{k+1} & \frac{\beta}{k}+\frac{1-\beta}{k+1} \\ \vdots & \frac{\beta}{k}+\frac{1-\beta}{k+1} & \frac{\beta}{k}+\frac{1-\beta}{k+1} & \ddots & \frac{\beta}{k}+\frac{1-\beta}{k+1} & \frac{\beta}{k}+\frac{1-\beta}{k+1} \\ \frac{\beta}{k}+\frac{1-\beta}{k+1} & \frac{\beta}{k}+\frac{1-\beta}{k+1} & \frac{\beta}{k}+\frac{1-\beta}{k+1} & \frac{\beta}{k}+\frac{1-\beta}{k+1} & \frac{1-\beta}{k+1} & \frac{\beta}{k}+\frac{1-\beta}{k+1} \\ \frac{1-\beta}{k+1} & \frac{1-\beta}{k+1} & \frac{1-\beta}{k+1} & \dots & \frac{1-\beta}{k+1} & \frac{1-\beta}{k+1} \end{bmatrix}$$

6.4

a) $\frac{\beta y}{m} + \frac{1-\beta}{n}$

b) $x + \beta m(\frac{\beta y}{m} + \frac{1-\beta}{n}) \Leftrightarrow y = \frac{x}{1-\beta^2} + \frac{\beta m}{(1+\beta)n}$

6.5

$r = R(2) = R(3)$ az átmenetmátrix felépítéséből közvetlen következő módon, valamint $R(4) = \frac{R(2)}{2} + \frac{R(1)}{3} = \frac{r}{2} + \frac{\frac{3r}{2}}{3} = r$

7.1

Aszinkron frissítéssel, és a sajátértékszámításra visszavezethető közvetlen módszerrel a csúcsok autoritás és központiság értékei rendre a következők szerint alakulnak (megint csak 1-re normálást követően):

$$a = \begin{pmatrix} 0,00 \\ 0,50 \\ 0,25 \\ 0,25 \end{pmatrix}$$

$$h = \begin{pmatrix} 0,5 \\ 0,0 \\ 0,5 \\ 0,0 \end{pmatrix}$$

Fontos: itt az átmenetmátrix helyett a szomszédsági mátrix képzi a számítások alapját.

7.2

a) A kialakuló klaszterhierarchia a következő lesz: $(((A, B), C), D)$

b) A kialakuló klaszterhierarchia a következő lesz: $((A, B), (C, D))$

7.3

L_2 norma használata mellett a soron következő klaszterközéppontok a $(8, 2)$, majd a $(3, 7)$ lesznek, ugyanis ezeknek a pontoknak lesz a maximális a legközelebbi klaszterközéppontoktól vett távolságaik (rendre $\sqrt{68}$, valamint $\sqrt{50}$).

7.4

Igen.

8.1

A gyakorlattól eltérően itt most L_2 normával és távolságnégyzetekkel számolva:
- ha a $(7, 5)$ pont lesz egy önálló klaszter $\Rightarrow SSE = 18,5$
- ha a $(2, 2)$ pont lesz egy önálló klaszter $\Rightarrow SSE = 12,5$
- ha a $(3, 8)$ pont lesz egy önálló klaszter $\Rightarrow SSE = 17$

$\Rightarrow$ A legbölcsebb döntésnek az tűnik, ha a $(2, 2)$ pontot egy külön klaszterbe soroljuk, a további két pontot pedig meghagyjuk egy klaszterbe esőnek.
(Az eredeti SSE értéke $9 + 13 + 10 = 32$ volt a $\mu = \begin{pmatrix} 4 \\ 5 \end{pmatrix}$ centroidhoz viszonyítva.)

8.2

A spektrálklaszterezés a gráf Laplace-mátrixának Fiedler-vekotra alapján particionálja a pontokat, a vektor komponenseinek előjelei alapján.

$$\begin{pmatrix} 0,70 \\ 0,34 \\ 0,07 \\ 0,08 \\ -0,25 \\ -0,21 \\ -0,34 \\ -0,40 \end{pmatrix}$$

$Cut(\{1, 2, 3, 4\}, \{5, 6, 7, 8\}) = 2$

$NormCut(\{1,2,3,4\},\{5,6,7,8\}) = \frac{2}{6} + \frac{2}{7} = 0,619$

8.3

A szélességi bejárások során minden élt kétszer látogatunk meg, így az eredeti köztességértékek kétszeresét kapjuk meg, amennyiben minden csúcsból végrehajtunk egy szélességi bejárást.
$2 * betweenness(A,B) = 5+1+0+1+1+1+1 = 10$
$2 * betweenness(A,C) = 1+0+1+0+0+0+0 = 2$
$2 * betweenness(B,C) = 0+1+5+1+1+1+1 = 10$
$2 * betweenness(B,D) = 4+4+4+3+3+3+3 = 24$
$2 * betweenness(D,E) = 1+1+1+1+0+4,5+0,5 = 9$
$2 * betweenness(D,F) = 1+1+1+1+4+0+0 = 8$
$2 * betweenness(D,G) = 1+1+1+1+0+4,5+0,5 = 9$
$2 * betweenness(E,F) = 0+0+0+0+1+1,5+0,5 = 3$
$2 * betweenness(F,G) = 0+0+0+0+1+1,5+0,5 = 3$

9.1

$B \perp\!\!\!\perp R \,|\, J \Leftrightarrow P(B,R|J) = P(B|J)P(R|J) \Leftrightarrow P(B|R,J) = P(B|J)$
$P(B,R|J) = \frac{0,16}{0,8} = 0,2 \neq 0,2208 = 0,92 * 0,24 = P(B|J)P(R|J)$, azaz B és R változók nem feltételesen függetlenek egymástól J változó értékének ismeretében.

Néhány további marginális valószínűség a hálózatból:
$P(J) = 0,16+0,576+0,032+0,032 = 0,8$
$P(B|R,J) = \frac{0,16}{0,16+0,032} = 0,8333$
$P(B|R,\neg J) = \frac{0,16}{0,16+0,032} = 0,9474$
$P(B|J) = 0,92$
$P(B,\neg R|J) = 0,72$
$P(R|J) = \frac{0,192}{0,8} = 0,24$

9.2

$H(Y) \approx 0,9709$
$H(M|Y) \approx 0,7684$, $H(M,Y) \approx 1,7394$
$MI(M;Y) \approx 0,2243$

Jellemzők elvárt együttállása a megfigyelttel azonos marginálisok, de egyébként egymástól való független viselkedése esetén

	Y=+	Y=-	
M	27	18	45
$\neg M$	33	22	55
	60	40	100

$\chi^2 = \frac{8450}{297} \approx 28,45$

Köszönetnyilvánítás

Az oktatási segédlet a TÁMOP-4.2.4.A/2-11/1-2012-0001 azonosító számú Nemzeti Kiválóság Program – Hazai hallgatói, illetve kutatói személyi támogatást biztosító rendszer kidolgozása és működtetése konvergencia program című kiemelt projekt által nyújtott személyi támogatással valósult meg. A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg.