2018. szeptember 18.

Bevezetés az adattudományba 1.

Gyakorló feladatsor

1) Döntsd el az alábbi attribútumok típusát kétféle megközelítés szerint! Folytonos, diszkrét, bináris? Kategorikus, ordinális, kvantitatív (intervallum, skála)?
   a) Tengerszint feletti magasság
   b) Férőhelyek száma egy szállodában
   c) Katonai rangok
   d) K épület alulájának középpontjától vett távolság
   e) Könyvek ISBN száma
   f) Szögek fokai (0 és 360 között)
   g) Átlátszóság mértéke: átlátszó, áttetsző, átlátszatlan
   h) Ruhatári szám
   i) Osztályzatok (1-5)
   j) Érmek (bronz, ezüst, arany)
   k) Nem
   l) Kor (években mérve)
   m) pH-érték

2) Bizonyítsd be, hogy
   a) $L_1(x, y) = \sum_{i=1}^{d} |x_i - y_i|$ metrika
   b) $L_2^2(x, y) = \sum_{i=1}^{d} (x_i - y_i)^2$ nem metrika

3) Hogy számolnád ki két embert reprezentáló attribútumvektor távolságát, ha a komponensek a következők:
   - a személy magassága (1,5 m és 1,8 m között)
   - a személy súlya (40 kg és 120 kg között)
   - éves jövedelme (10 ezer $ és 1 millió $ között)

4) Szövegállományok sokszor ún. dokumentum-szó mátrixszal adottak. Például tekintsük az alábbi három dokumentumot, amit 8-dimenziós vektorok reprezentálnak:
   d_1: „ant bee”
   d_2: „dog bee hog ant”
   d_3: „cat gnu dog eel fox”
   Számold ki d_1 és d_2 közötti egyszerű egyezőségi együtthatót (SMC) és a Jaccard-együtthatót is, illetve az ezeknek megfelelő távolságokat! Melyik változat felel meg jobban az elvárásainknak? Miért?

	ant	bee	cat	dog	eel	fox	gnu	hog
d_1	1	1						
d_2	1	1		1				1
d_3			1	1	1	1	1	

(Megj.: Ebben a megközelítésben persze nem tudunk különbséget tenni a „John is quicker than Mary” és a „Mary is quicker than John” dokumentumok között.)

5) Jelölje tf_{ij} a fenti dokumentum-szó mátrix azon elemét, amely az i-edik dokumentumhoz és a j-edik szóhoz tartozik, pl. $tf_{11} = 1$, ahol szó1 = „ant”. Tekintsük az alábbi tf.idf nevű változótranszformációt. Legyen df_j a nemnulla elemek száma a j-edik oszlopban, azaz a j-edik szót tartalmazó dokumentumok száma. Például $df_1 = 2$. Legyen m a dokumentumok száma. Ekkor a transzformáció:

$$tf.idf_{ij} = tf_{ij} \cdot log(m/df_j)$$

Mi ennek a transzformációnak a hatása? Valós dokumentum-szó mátrixokat elképzelve mi lehet a célja?

6) Az alábbiakban azt láthatod, hogy mely felhasználók (A, B, C) mely termékeket (a, b, …, h) vásárolták. Ez alapján határozd meg A és B felhasználó Jaccard és koszinusz hasonlóságát!

	a	b	c	d	e	f	g	h
A	1	1	0	1	1	0	1	1
B	0	1	1	1	1	1	1	0
C	1	0	1	1	0	1	1	1

7) Feltéve, hogy a nyújtás/összenyomás költsége 0, számold ki a következő két idősor DTW távolságát (belső távolságfüggvényként tekintsd a szokásos eltérést)! Határozd meg a „vetemítő utat” is!

$$t_1 = (3, 2, 5, 7, 8, 9)$$
$$t_2 = (2, 3, 2, 3, 6, 8)$$

2018. október 1.

Bevezetés az adattudományba 1.

Gyakorló feladatsor II.

1. Hány logikai függvényt (Boole-függvényt, $f: \{0,1\}^N \rightarrow \{0,1\}$) lehet képezni N bináris attribútumon? Add is meg N=2 esetben a lehetséges függvényeket!

2. Képesek megtanulni a Boole-függvényeket a döntési fák?
   Hogy lehetne döntési fával reprezentálni az alábbi függvényeket:
   A OR B, A AND B, A XOR B

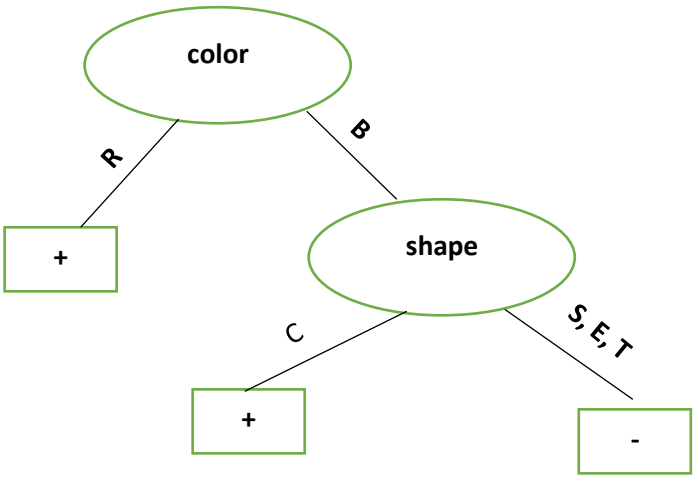
3. A következő táblázatban három attribútumú (A, B, C) adatokat látunk, két lehetséges címkeváltozóval (+, -). Építs kétmélységű döntési fát!
   a) Használd a félreosztályozási hibát mint inhomogenitási mértéket. Mindhárom attribútumra számítsd ki az információs nyereséget (gain)! Melyik attribútum szerint vágjunk először?
   b) Ismételd meg az előző lépést a gyermek csúcsokra is. Mely csúcsokat kell vágnunk, és mi lesz a második vágó attribútum?
   c) Számítsd ki a következőket: Accuracy, Error rate, Precision, Recall, F-mérték!
   d) Hogyan alakulna a döntési fa, ha rögzítenénk, hogy először a C attribútum szerint vágunk?

A	B	C	Number of Instances	
			+	−
T	T	T	5	0
F	T	T	0	20
T	F	T	20	0
F	F	T	0	5
T	T	F	0	0
F	T	F	25	0
T	F	F	0	0
F	F	F	0	25

4. Rajzold fel az alábbi osztályozó ROC görbéjét, és számold ki az AUC értéket? Mit javasolnál az eredményt látva? A táblázatban balról jobbra a pontérték („pozitív osztály valószínűsége”) nő.

	Label	+	+	-	+	-	-	+	+	-	+
TP	6										
FN	0										
TN	0										
FP	4										
TPR	1										
FPR	1										

5. Egy a tanítóadatokon épített lehetséges döntési fa sematikus ábráját látod az alábbiakban.
   a) Add meg a levelek pontértékét (pozitív osztályba tartozás valószínűségét) a tanítóadatok alapján!
   b) Állítsd sorrendbe a tesztadatokat pontértékük alapján!
   c) Ábrázold a tesztadatok ROC görbéjét és számítsd ki az AUC értéket!
   d) Ábrázold a ROC görbét két új tesztadat hozzáadása után! Vigyázz, megegyező pontértékek esetén a ROC görbe diagonálisan változik!



Train	Shape	Color	Size	Class
T1	S	R	L	+
T2	C	R	H	+
T3	C	B	H	+
T4	T	R	L	+
T5	S	B	M	-
T6	E	B	L	-
T7	C	R	M	-

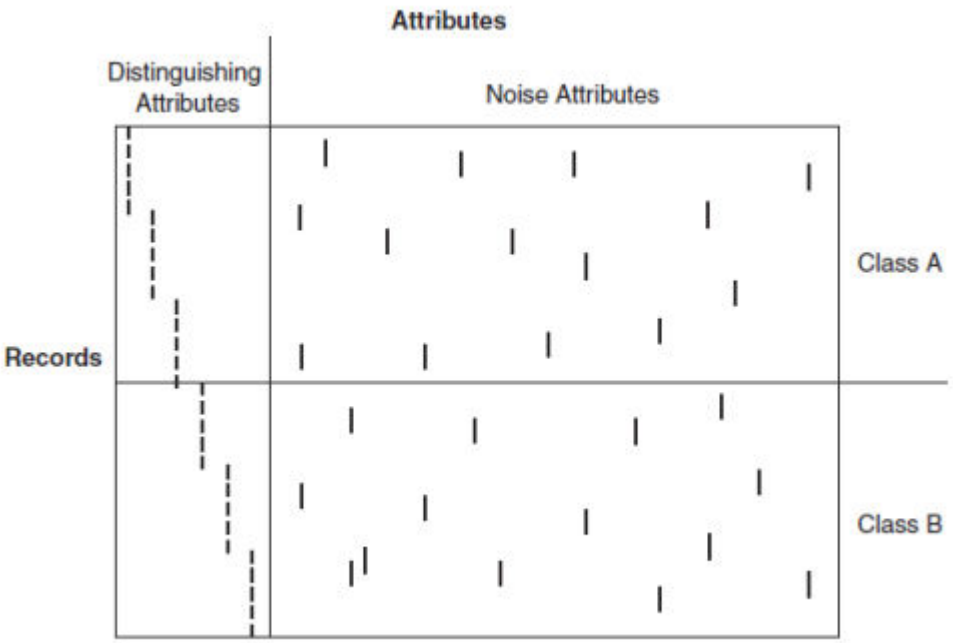
Test	Shape	Color	Size	Class
Test1	C	R	H	+
Test2	C	B	L	-
Test3	E	B	H	-

Test	Shape	Color	Size	Class
Test4	C	R	L	+
Test5	E	R	H	-

6. Osztályozd naiv Bayes-módszerrel az alábbi táblázatban szereplő tanítóadatok alapján a következő rekordot: (Marital Status = Single, Income = 90K). Az éves bevételt diszkretizáld 20K hosszúságú intervallumok szerint!
   a) Használd az eredeti becsléseket!
   b) Használj Laplace-becslést!

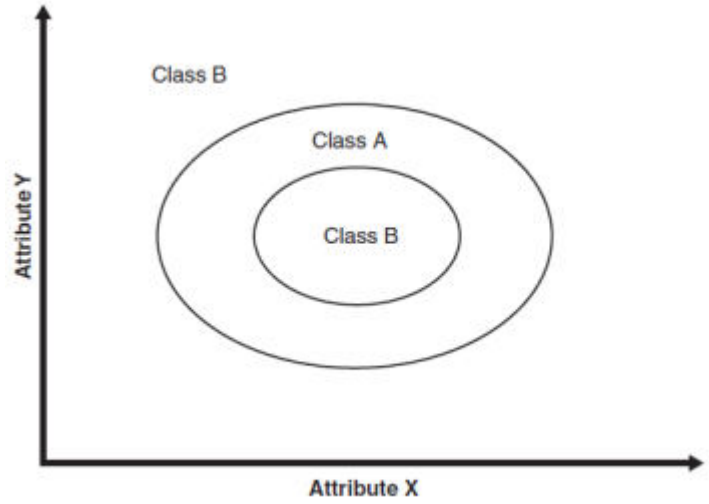
Marital status	Annual Income	Default
Single	125K	No
Married	95K	No
Single	70K	No
Married	120K	No
Divorced	75K	Yes
Married	60K	No
Divorced	220K	No
Single	85K	Yes
Married	75K	No
Single	90K	Yes

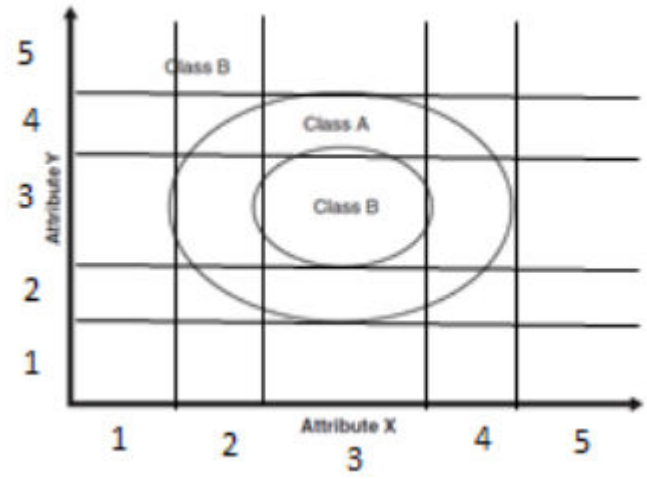
7. Tegyük fel, hogy az alábbi adattábla 1000 A osztályú és 1000 B osztályú rekordot tartalmaz. Van néhány megkülönböztető erővel bíró bináris változónk X_1, X_2, …, ezen felül pedig sok zajos bináris attribútumunk, amik véletlenszerűen vesznek fel 1 értéket.
   a) Vázolj egy döntési fát, ami ilyen jellegű adaton tanul! Mit gondolsz, jól tud teljesíteni egy ilyen adaton a döntési fa?
   b) Mely sorok vannak közel az első sorhoz? Hogy teljesítene ezen az adaton a kNN osztályozó?
   c) Hogy alakulnak a feltételes valószínűségek az egyes attribútumokra? A naiv Bayes osztályozó képes jól osztályozni ilyen jellegű adatot?



2018. október 1.

8. Tekintsük a következő két attribútummal (X és Y) és két osztályváltozóval (A és B) rendelkező adatot, ahol az A és B osztályú rekordok X-Y térben való elhelyezkedését az alábbi ábra szemlélteti.
   a) Hogy működne a döntési fa egy ilyen adaton? Jelöld be a döntési határokat!
   b) Hogy teljesítene az adaton a kNN osztályozó? Mitől függ?
   c) Hogy teljesítene a naiv Bayes osztályozó? Vázold az egyes feltételes valószínűségeket! Feltehetjük, hogy a két osztályban kb. ugyanannyi elem van, és azon túl az eloszlások egyenletesek. Segítségül megadtunk egy lehetséges diszkretizálást, ezt használd a továbbiakban!



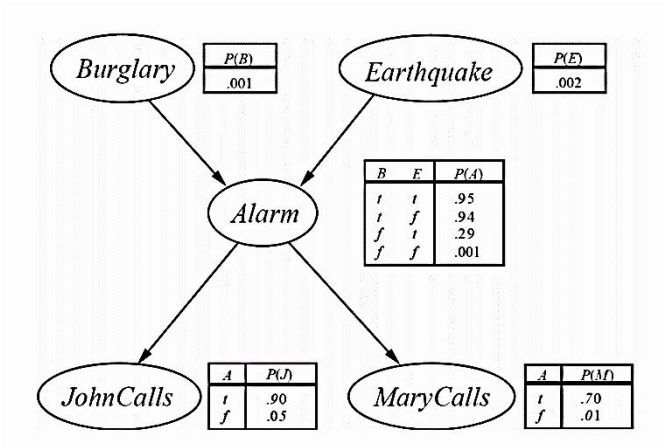


2018. október 16.

Bevezetés az adattudományba 1.

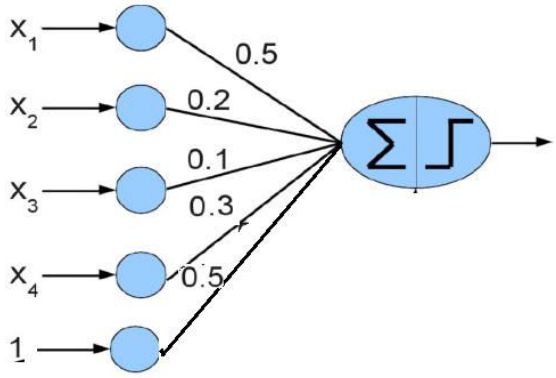
Gyakorló feladatsor III.

1. Adottak a következő bináris kimenetelű valószínűségi változók
   B = betörés a lakásodba
   E = földrengés a lakásod környékén
   A = a riasztó megszólalása
   J = John hívása a riasztó miatt
   M = Mary hívása a riasztó miatt
   Az egyes változók közötti összefüggést és az egyes valószínűségeket a következő Bayes-háló szemlélteti!
   a. Hány paraméterrel írható le így a teljes együttes eloszlás? Hány paraméter kéne, ha nem lenne ismert az egyes változók közötti összefüggőségi hálózat?
   b. A Burglary és az Earthquake függetlenek egymástól? Feltételesen függetlenek-e az Alarm értékét ismerve?
   c. A JohnCalls és a MaryCalls függetlenek egymástól? Feltételesen függetlenek-e az Alarm értékét ismerve?
   d. Vázold, hogy hogyan kéne meghatározni a $P(B = t \mid M = t, J = f)$ feltételes valószínűséget!

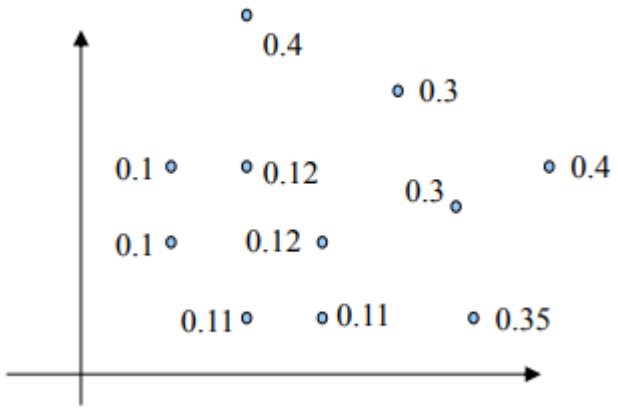


2. Legyen adott három címkézett rekord a síkon. A harmadik koordináta a címke: (0,0,-2); (0,1,1); (1,0,2). Határozzuk meg a következő lineáris regresszióban az együtthatókat, a cél a négyzetes eltérés minimalizálása: $y = w_1 x_1 + w_2 x_2 + w_0$
   a. Határozd meg az optimális együtthatókat analitikusan!
   b. Határozd meg az optimális együtthatókat gradiens módszerrel!
   c. Határozd meg az optimális együtthatókat sztochasztikus gradiens módszerrel! A gradiens módszerek esetén indulj ki a $w_1 = w_2 = w_0 = 1$ kezdeti állapotból, és a tanulóráta legyen 0.25.

3. Adott az alábbi perceptron, amit bináris osztályozási feladat megoldására használunk. A két osztálycímkét +1-gyel és -1-gyel jelöljük, az aktivációs függvény a szignum (előjel) függvény.
   a. Hogyan osztályozza ez a perceptron az alábbi példányt:
   $$(x_1 = 1\,,\ x_2 = -0.8\,,\ x_3 = -0.3,\ x_4 = 1.5)$$
   b. Tételezzük fel, hogy a fenti példány valódi osztálycímkéje -1, és ezt a rekordot a perceptron tanításához használjuk, a tanítási ráta λ=0.1. Hajts végre egy javító lépést! Változnak-e az élek súlyai, ha igen, hogyan?



4. Regressziós problémát szeretnénk megoldani döntési fával. A levelek maximális számát 3-ra állítjuk. Vázold, hogy hogyan működhetne egy ilyen döntési fa az alábbi adatokon, hol húzódnak a döntési határok? Mi alapján dönt a fa? (Nem kell pontos számolásokat végezned, a lényeg csak, hogy derüljön ki, hogy mi alapján dönt a fa.)



5. Tekintsük az alábbi adatokat, ahol az első két koordináta bináris attribútumok, a harmadik koordináta pedig a címke: (1, 1, -); (1, -1, +); (-1, 1, +); (-1, -1, -). Melyik Boole-függvényt ismered föl az adatokban? Lineárisan szeparálható? Ha igen, add meg annak a szeparáló egyenletnek az egyenletét, amelyik esetén a legnagyobb a margó! Ha nem, akkor alkalmazd a $\Phi = (x_1, x_2, x_1 x_2)$. transzformációt, és add meg így a legnagyobb margójú szeparáló sík egyenletét!

2018. október 16.

6. Készíts perceptront a következő logikai függvényhez vagy mutasd meg, hogy ez nem lehetséges! Ez utóbbi esetben készíts egy rejtett rétegű neurális hálózatot!
    a. A AND B AND C
    b. (A XOR B) AND (A OR B)

7. Ábrázold, hogy a $2 + x + y = 0$ egyenes hogy szeli a síkot pozitív és negatív részekre. Az alábbi pontok közül melyek szupport vektorai az egyenesnek: $(-3, 0);\ (0, -3);\ (-1,0);\ (0, -1);\ (0, 0)$?

8. Tekintsünk két neuront, az egyik bemente x_1 és kimenete $y_1 = ax_1 + b$, a másik bemenete x_2 és kimenete $y_2 = c\ x_2 + d$ (mindkettő esetben az aktivációs függvény az identitás). Most kapcsoljuk össze a két neuront úgy, hogy a második neuron bemente y_1 legyen, azaz $x_2 = y_1$.
    a. Rajzold fel ezt a neurális hálót!
    b. Írd fel a végső y_2 kimenetet x_1 függvényeként!
    c. Tekintve egy x bementet egy elvárt y kimenettel, vezesd le, hogy hogy alakul egy rekordot látva a gradiens módszerrel való lépés *a, b, c, d* súlyokat illetően, ha a minimalizálandó célfüggvény a négyzetes hiba!

2018. november 27.

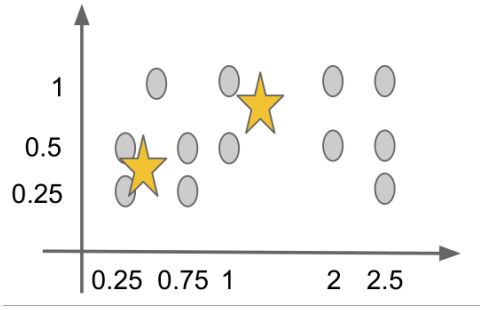
Bevezetés az adattudományba 1.

Gyakorló feladatsor IV.

1. Tekintsük a következő vásárlói kosarakat (tranzakciókat).
   a. Határozd meg a gyakori, maximális és zárt elemhalmazokat, ha a minimális gyakoriság 0,3 (minsupp = 0,3)? Szemléltesd a példán az apriori algoritmus működését!
   b. Határozd meg a következő szabályok megbízhatóságát (confidence) és lift mutatóját!
      i. kenyér -> tej
      ii. {kenyér, kávé} -> tej

TID	Kosár
1	tej, sör, kávé
2	tej, kávé, kenyér
3	tej, kávé, sör
4	kenyér, sör
5	alma, kávé
6	kenyér
7	alma, kenyér
8	alma
9	kávé
10	tej

2. A k-közép klaszterezőt a megjelölt klaszterközéppontokkal inicializáljuk. Hajts végre egy iterációt, és add meg az *i+1*-dik iteráció végén az új klaszter-középpontokat! Mutasd meg a számításokat, amiket végeztél!

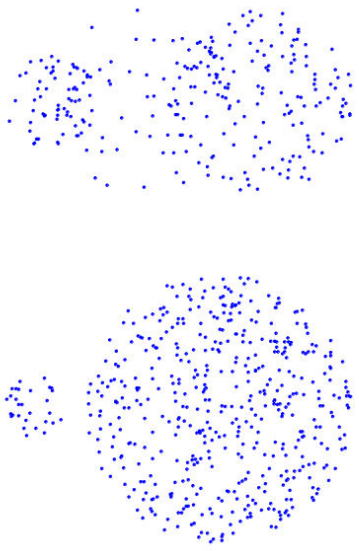


2018. november 27.

3. Adott a következő értékelési mátrix, ahol a sorok a felhasználóknak, a filmek az oszlopoknak felelnek meg. A kitöltetlen értékeléseket szeretnénk megbecsülni rejtett faktormodell segítségével. Két rejtett faktort tételezünk fel, a tanulórátát $\varepsilon = 0.1$-re állítjuk, és eltekintünk a regularizációtól!
   a. Mutass be egy javító lépést az A felhasználót és a TW filmet használva! A felhasználóhoz és a termékekhez tartozó faktormátrixokat csupa 1 elemmel inicializáld!
   b. Mutasd meg, hogy ez a javító lépés megfelel egy gradiens módszer szerinti lépésnek!
   c. Végezd el a fenti javító lépést, de most úgy, hogy használsz regularizációs tagot is!
   d. Bónusz: Melyik filmeket jelölhetik a rövidítések?

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

4. Mutasd meg, hogy hogyan particionálná két klaszterre az alábbi adathalmazokat a *k*-közép klaszterező és a hierarchikus klaszterező MIN (single linkage), illetve MAX (complete linkage) távolsággal számolva!

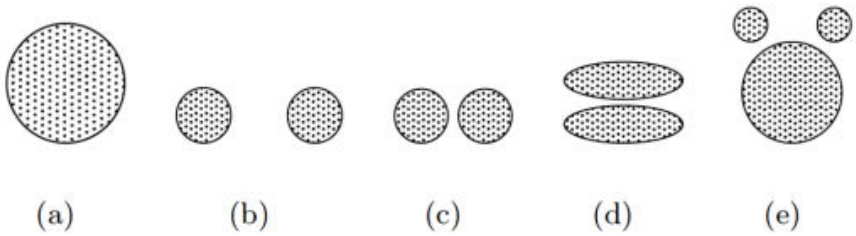


2018. november 27.

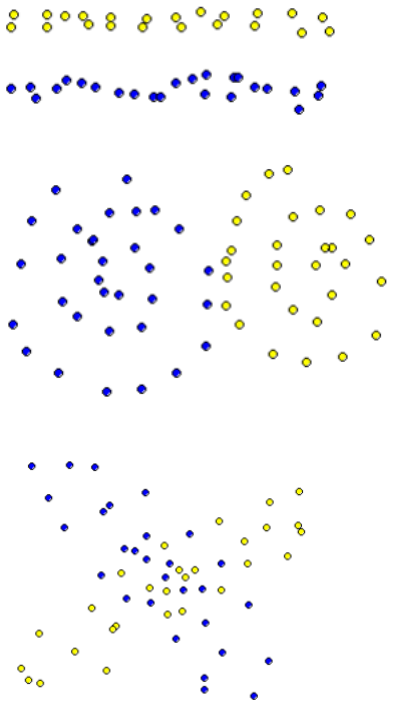
5. Adott az alábbi hasonlósági mátrix. Mutasd meg dendogramok segítségével, hogy hogyan klaszterezi a példányokat egy hierarchikus klaszterező egyszerű láncmódszert (single linkage), illetve teljes láncmódszer (complete linkage) használva!

	1	2	3	4	5
1	1	0,15	0,6	0,15	0,95
2	0,15	1	0,5	0,2	0,2
3	0,6	0,5	1	0,05	0,7
4	0,15	0,2	0,05	1	0,85
5	0,95	0,2	0,7	0,85	1

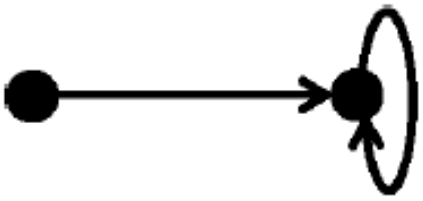
6. A következő kétdimenziós adatpontok esetén vázold, hogy milyen eredményt adna a *k*-közép klaszterező algoritmus euklideszi távolságot használva! Jelöld azt is, hogy körülbelül hol lennének a végső centroidok! Ha úgy gondolod, hogy van olyan eset, ahol több lehetséges megoldás is van, ott gondold át azt is, hogy melyik ad lokális, melyik globális optimumot!
   a. k=2
   b. k=3
   c. k=3
   d. k=2
   e. k=3



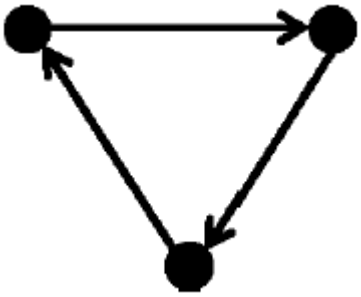
7. Melyik klaszterező algoritmus bontaná az alábbi adatokat a legjobban két klaszterre, ha a szakértő által színekkel jelzett klaszterek megtalálása lenne a cél? Gondold át a következőket: hierarckikus klaszterező (MIN, ill. MAX távolság esetén), *k*-közép, Gauss-keverékmodell.



8. Számold ki az alábbi gráfhoz tartozó PageRank értékeket a stacionárius eloszlás meghatározásával $\alpha = 1$ és $\alpha = 0.8$ esetben!



9. Számold ki a PageRank algoritmus első néhány iterációs lépését $q^0 = (1,0,0)$ állapotvektorból indulva az alábbi gráf esetén. Határozd meg $\alpha = 1$ és $\alpha = 0.7$ esetben is. Mi a fő különbség? Milyen gond lép fel, ha a teleportáció valószínűsége 0?



10. Mi a valószínűsége, hogy pontosan l lépést tesz meg a PageRank bolyongás éleken bolyongva (azaz linkekre való kattintást imitálva) két teleportálás között. Határozd meg a valószínűséget α függvényében!