

**Adatbányászati technikák 2. zárthelyi**  
**2016. május 10.**

Válaszait mindig indokolja meg, a tanult algoritmusok alkalmazásánál jelezze a főbb lépéseket! Összesen 35 pont szereshető, a zh 14 ponttól sikeres. 10 mmnká!

1. (3 pont) Írja le röviden, hogy hogyan kell a síkon pontokat osztályozni a K<sup>N</sup>N algoritmusmal, L<sub>2</sub> távolságot, k = 11 értéket használva úgy, hogy a címke meghatározáskor nem a többségi szavazást használjuk, hanem figyelembe vesszük a távolságokat is.
2. (a) (1 pont) Mit jelent az, hogy egy 3-dimenzióban adott bináris osztályozási feladat esetén a két osztály lineárisan szeparálható?  
(b) (1 pont) Mi a lineárisan szeparálhatóság kapcsolata a perceptronnal?
3. (3 pont) Egy olyan neurális hálózatot készítettünk, mellyel az  $x_1, x_2, \dots, x_n$  inputértékek alapján akarjuk az egyetlen (bináris) célváltozó értéket előrejelezni. A hálózatnak (az input és output szinten kívül) két belső szintje van, mindegyiken  $m$  darab perceptronnal. Hány különböző  $\Theta$  értéket kell definiálnunk a hálózat működtetéséhez? (A választ  $n$  és  $m$  függvényében adja meg és indokolja is meg.)
4. (a) (1 pont) Hogyan kombinálhatunk több osztályozót egy bináris osztályozási feladat esetén?  
(b) (2 pont) Hogyan készíthetünk ugyanazon training set segítségével több osztályozót ugyanarra a feladatra? Válasszon röviden (max. 1-2 mondat) két különböző megoldást.
5. (5 pont) Az 1, 4, 5, 10, 11, 14, 21, 23, 45 számokat akarjuk K-means algoritmusmal három klaszterbe sorolni. Írja le, hogy hogyan hű le az algoritmus, ha az 1, 4, 5 kezdő-centroidokat használjuk!
6. (a) (2 pont) Adja meg a klaszterezések jóségának mérésére szolgáló SSE memység definícióját és magyarázza el a képletben használt jelöléseket.  
(b) (1 pont) Hogyan használhatjuk az SSE-t két klaszterezés összehasonlítására?
7. (5 pont) Az alábbi hasonlóság-mátrix alapján rajolja fel dendrogramon, hogy a MIN (single link) hierarchikus klaszterező algoritmus hogyan klaszterez. A dendrogram elkészítésének és a hasonlóság-mátrix változásának minden fázisát mutassa be.

	$p_1$	$p_2$	$p_3$	$p_4$	$p_5$	$p_6$
$p_1$	1	0.7	0.7	0.8	0.1	0.95
$p_2$	0.7	1	0.8	0.7	0.3	0.8
$p_3$	0.7	0.8	1	0.2	0.9	0.92
$p_4$	0.8	0.7	0.2	1	0.91	0.6
$p_5$	0.1	0.3	0.9	0.91	1	0.5
$p_6$	0.95	0.8	0.92	0.6	0.5	1

8. (a) (1 pont) Adja meg, hogy hogyan kell kiszámolni az  $X \rightarrow Y$  asszociációs  $\text{conf}(X \rightarrow Y)$ -nal jelölt megbízhatóságot! A használt jelöléseket is magyarázza el.  
(b) (3 pont) Bizonyítsa be, hogy ha  $\text{conf}(Z \setminus X \rightarrow X) < \text{min.conf}$ , akkor  $\text{conf}(Z \setminus X' \rightarrow X') < \text{min.conf}$  is fennáll, amennyiben  $X \subseteq X'$ .
9. (5 pont) Tekintsük az alábbi 10 tranzakciót az 123456 elemhalmazonon: 1346, 135, 146, 1234, 236, 245, 246, 346, 456, 46. Az Apriori algoritmusmal határozza meg a gyakori elemhalmazokat,  $\text{min-sup} = 0.2$  kiszűréssel! Mutassa be, hogy hogyan alakulnak a  $C_k$  és  $F_k$  halmazok az algoritmus futása során.  
(c) (2 pont) Hogyan használjuk ezt az állítást a magas megbízhatóságú szabályok előállításán?