Adatbányászati technikák 1. zárthelyi
2016. március 31.

Válaszait mindig indokolja meg, a tanult algoritmusok alkalmazásánál jelezze a főbb lépéseket!
Összesen 35 pont szerezhető, a zh 14 ponttól sikeres. Jó munkát!

1. (3 pont) Mi a különbség a kvalitatív és kvantitatív attribútum között? Mutasson példát kvalitatív attribútumra! R-ben milyen típust használunk kvalitatív attribútum megadására?

2. (1 pont) Mi a 20%-os percentilis értékének definíciója?

3. Egy adatbázisban hallgatók adatait tároljuk, az egyik oszlop a Neptun-kód. Egy másik oszlop, mint célváltozó előrejelzésére döntési fát akarunk építeni.
   (a) (2 pont) Az összes lehetséges vágás közül miért választja a döntési fa építő algoritmus az első lépésben a Neptun-kód szerinti multiway vágást?
   (b) (1 pont) Miért baj, hogy ezt választja?
   (c) (1 pont) Hogyan lehet elkerülni ezt a választást?

4. (2 pont) Mi az overfitting jelenség és hogyan próbáljuk ezt döntési fák építése során elkerülni? (Max. 2-3 mondatban válaszoljon.)

5. Egy osztályozó modell jóságát szeretnénk értékelni:
   (a) (3 pont) Hogyan kell kiszámolni a precision-t és a recall értékét? (A definícióban használt fogalmak jelentését is adja meg.) Mi ezen mérőszámok legnagyobb lehetséges értéke? Miért?
   (b) (1 pont) Adjon példát olyan valós életből származó osztályozási feladatra, amikor a hibás pozitív címkézést akarjuk leginkább elkerülni.
   (c) (2 pont) Az alábbi cost (költség-) mátrixban s és t pozitív egész számok. Melyiket kell nagyobbnak választanunk, ha a tévesen pozitívnak címkézett esetek nagyobb gondot jelentenek, mint a tévesen negatívnak címkézettek? Válaszát indokolja is meg.

	Algo: +	Algo: -
Reality: +	0	t
Reality: -	s	0

Folytatás a túloldalon!

6. (a) (1 pont) Adja meg a cosine hasonlóság definícióját n hosszú vektorokra!
   (b) (1 pont) Mutasson egy valós életből vett példát a cosine hasonlóság használatára!
   (c) (1 pont) Számolja ki a cosine hasonlóság értékét a $(3, 4, 0, 2, 0)$ és $(2, 3, 4, 0, 0)$ vektorokra!

7. (a) (2 pont) Adja meg az alábbi fogalmak definícióját: L_∞ és L_1 távolság
   (b) (1 pont) Adjon ellenpéldát vagy bizonyítsa be, hogy tetszőleges, azonos hosszú x és y vektorokra igaz az, hogy $L_\infty(x, y) \leq L_1(x, y)$.

8. (1 pont) Miért/mikor hasznos a Mahalanobis-távolság használata?

9. (6 pont) Az alábbi táblázat arról tartalmaz adatokat, hogy vásároltak-e egy bizonyos terméket az adott paraméterekkel rendelkező ügyfelek. Ezen halmaz alapján szeretnénk egy döntési fát építeni a **vásárol** változó előrejelzésére.
   Döntse el, hogy mely attribútum szerint fog először vágni az órán tanult algoritmus, ha csak multiway split-eket veszünk figyelembe és a classification-errort használjuk.

	kor	jövedelem	saját autó	nem	**vásárol**
1	fiatal	magas	nincs	férfi	nem
2	fiatal	magas	nincs	nő	nem
3	középkorú	magas	nincs	férfi	igen
4	idős	alacsony	van	férfi	nem
5	fiatal	közepes	nincs	férfi	nem
6	fiatal	alacsony	van	férfi	nem
7	idős	magas	van	férfi	igen
8	fiatal	közepes	van	nő	nem
9	középkorú	közepes	nincs	férfi	nem
10	idős	közepes	nincs	férfi	nem
11	középkorú	alacsony	nincs	nő	igen
12	idős	magas	nincs	nő	igen

10. (6 pont) A fenti táblázat alapján naív Bayes osztályozó segítségével határozza meg, hog várhatóan vásárol-e egy olyan ügyfél, aki idős, alacsony jövedelmű, nincs saját autója nő.