Adatbányászati technikák 1. zárthelyire gyakorló feladatok
2018. március 22.

1. a) A $p = (1, 1, 1, 0, 0, 0)$ és $q = (0, 0, 1, 1, 1, 0)$ vektorokra határozza meg az L_1, L_2, L_∞ távolságot, az SMC-t és a Jaccard-hasonlóságot.

   b) A $p = (2, 1, 0, 7)$ és $q = (1, 3, 0, 0)$ vektorokra határozza meg az L_1, L_2, L_∞ távolságot, és a cosine-hasonlóságot.

2. Van két modellem ugyanazon feladatra, egy teszthalmazon tesztelem a jóságukat. Az alábbi táblázat azt mutatja, hogy a 12 tesztsoron a két modell milyen előrejelzést tett, a táblázat első sora pedig a célfüggyvény valódi értékét tartalmazza.

Valóság	-	-	+	-	-	-	+	-	-	-	+	-
M_1	-	-	-	-	-	-	-	-	-	-	-	-
M_2	+	+	+	-	+	-	+	-	-	-	-	-

   a) Írja fel mindkét modellre a confusion mátrixot, majd mindkét modellre számítsa ki az accuracy-t, a precisiont, a recallt és az F-value-t. Melyik modellt válasszuk az egyes mértékek szerint?

   b) Melyik modellt válasszuk, ha az alábbi költségmátrixot használjuk?

	Algo +	Algo -
Valóság +	-1	10
Valóság -	1	0

3. Az alábbi táblázat arról tartalmaz adatokat, hogy adott korú, súlyú, sportolási státuszú emberek érdeklődtek-e egy adott termék iránt. Ezen training halmaz alapján döntési fát akarunk építeni az "érdekli-e" változó előrejelzésére.

	kor	súly	sportol	érdekli
1.	fiatal	alacsony	igen	igen
2.	idős	közepes	nem	nem
3.	középkorú	magas	nem	igen
4.	idős	közepes	igen	nem
5.	fiatal	magas	nem	igen
6.	középkorú	alacsony	nem	nem
7.	idős	alacsony	nem	nem
8.	fiatal	közepes	nem	igen
9.	középkorú	magas	igen	igen
10.	idős	közepes	igen	nem

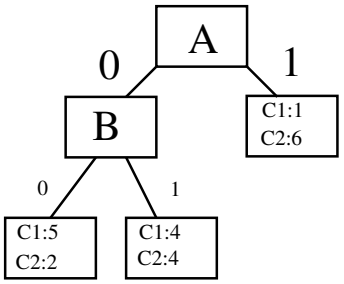
a) Melyik változó szerint vágjunk először, ha Gini alapján mérjük az inhomogenitást? (A kor és a súlynál a hármas multiway split-et, a sportolnál a bináris vágást vegyük figyelembe.)

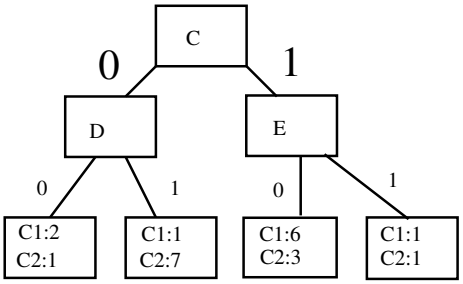
b) Hol legyen a 2. vágás és mi alapján történjen (továbbra is Ginivel dolgozunk)?

c) Mi lesz a végső fa? (Továbbra is Ginivel dolgozunk)

d) Ugyanez a kérdés (első vágás, második vágás, végső fa) classification error-ral.

4. Egy training halmazon az alábbi két döntési fát építettük fel:





A levelekben használ jelölés magyarázata: hány darab C1 és C2 címkéjű sor került oda.

a) Hány sorból áll a training halmaz?

b) Mekkora a két fa training error-ja? (Optimista hibának is hívtuk.) Melyik fa jobb eszerint?

c) Mekkora a két fa pesszimista hibája, ha a levelek $\ell = 1$-es büntetést kapnak? Melyik fa jobb eszerint?

d) Mekkora lehet a levelenkénti ℓ büntetés nagysága, ha az első fa jobb a pesszimista hibát tekintve?

e) Mekkora a két modell validation errorja az alábbi validation set-et használva?

A	B	C	D	E	cél
0	1	1	1	0	C1
0	1	0	0	1	C2
1	1	1	1	0	C1
0	1	0	1	0	C1
1	1	0	1	1	C2
0	0	0	0	0	C1

5. A 3. feladat táblázatát tanító halmazként használva Bayes-osztályozót akarunk készíteni.

(a) A sima Bayes osztályozót használva milyen címkét kap egy fiatal, közepes súlyú, sportoló ember?

(b) Mi a döntés, ha Laplace módszert használunk a Bayes eljáráson belül?

6. Bayes osztályozót szeretnék használni egy olyan helyzetben, ahol az A, B, C, D attribútumok alapján akarom jósolni a célváltozó értékét. Az A attribútum folytonos, a tanítóhalmaz $+$ címkéjű soraiban 10, 60, 70, 80, 130, 150 és 200 értékek vannak az A oszlopban, a $-$ címkéjű sorokban pedig 40, 50 és 60.

Hogyan kell kiszámolni az A=100, B = 1, C = 0, D = 3 sor címkéjét, ha $P(B = 1 \,|\, +) = 0.3$, $P(B = 1 \,|\, -) = 0.2$ $P(C = 0 \,|\, +) = 0.7$,$P(C = 0 \,|\, -) = 0.2$, $P(D = 3 \,|\, +) = 0.5$,$P(D = 3 \,|\, -) = 0.5$, $P(+) = 0.7$, $P(-) = 0.3$?

7. Osztályozza kNN osztályozóval $k = 5$ értéket használva a következő tanítóhalmaz segítségével a 0.9 értékű pontot.
(a) Használjon többségi szavazást!
(b) Használja a távolságokat is figyelembe vevő módszert a címke meghatározására!

$$0.1\,(-),\ 0.2\,(+),\ 0.4\,(+),\ 0.6\,(+),\ 0.8\,(-),\ 1.1\,(-),\ 1.7\,(+),\ 1.9\,(+),$$