Gépi tanulás (vimim136) **Vizsgabeugró kérdéssor** **2014. június 5.**

8:30 - 9:30 (60 perc !)

Minden kérdés két pontot ér (kapható 0, 1 vagy 2 pont), kivétel a 9. feladat, amely 4 pontot ér. Minden válaszhoz rövid, de áttekinthető indoklást is kérek- kivéve az 1. (feleletválasztós) feladatot! El kell érni az 50%-ot! (min. 10 pontot)

Név (nyomtatott betűvel):...Neptun-kód:.........................

Aláírás:..

1. **a.** Amennyiben az adatainkat nem tudjuk homogén (vagy közel homogén) klaszterekre bontani, akkor a tanult adat komplexitás mérték nem használható. **a. Igaz Hamis**

b. A döntési fánk egy adott csomópontjánál két teszt közül választhatunk, mindkettő 0,16 bit információ-nyereséget biztosít. Az első tesztnek négyféle lehetséges kimenetele van, és a csomópontba érkező három-ezres mintahalmazunkat ennek megfelelően négyfele osztja, mindegyik gyermek csomópontba 750-750 minta jut. A második teszt két kimenetellel rendelkezik, és 1500-1500 minta jut a kialakuló gyermek-csomópontokba. Ezek alapján a második tesztet célszerű választanunk. **b. Igaz Hamis**

c. Egy problémában két valós attribútum jellemzi az eseteket: x_1 és x_2; tudjuk, hogy $X_{1\mathrm{Lim}1} \leq x_1 \leq X_{1\mathrm{Lim}2}$ és . $X_{2\mathrm{Lim}1} \leq x_2 \leq X_{2\mathrm{Lim}2}$. A feladat 4 kisebb mintatérben megoldható feladatra való mechanikus dekomponálása a két attribútumtartomány felező – $(X_{1\mathrm{Lim}1}+X_{1\mathrm{Lim}2})/2$ és $(X_{2\mathrm{Lim}1}+X_{2\mathrm{Lim}2})/2$ – mentén biztosan nem könnyíti meg a megoldást. **c. Igaz Hamis**

d. Egy döntési fa csomópontjában elvégzett attribútum vizsgálat relevanciáját nézzük meg $\chi 2$ teszttel. A döntési fában elvégzett attribútum vizsgálat akkor releváns, ha a $\chi 2$ tesztben kiszámított valószínűségi változó távol van nullától (nagyobb egy adott limitnél). **d. Igaz Hamis**

(≤ 2 *jó válasz* **0 *pont***, $<$ *jó válasz* ***(jó válaszok száma*-2) *pont***, 4 *jó válasz* 2 ***pont***)

2. Mutassa meg ellenpélda segítségével, hogy a kétbemenetű (x_1 és x_2), eltolással NEM rendelkező, lineáris osztályozási határt megvalósító eszközünknek a 2-dimenziós síkon kevesebb mint 3 a VC dimenziója! Az eszköz kimenete: $\mathrm{h(x)=signum(A \cdot x_1 + B \cdot x_2)}$, ahol A és B az eszköz paraméterei. (**2 pont**)

3. Egy probléma megerősítéses tanulását végezzük, az állapotokat egy tízjegyű (decimális) számmal tudjuk jellemezni. Ha az állapotteret egy 20 bináris paraméterrel megadható rögzített struktúrájú függvénnyel tudjuk approximálni, akkor mekkora tömörítést érünk el? (**2 pont**)

4. 11-osztályos osztályozási feladatot oldunk meg, kétosztályos feladatok tanításával, és az eredmények integrálásával. Mind a 11 osztályból 2000-2000 ismert besorolású mintánk van. Az egyes részosztályozók tanításánál 80%-ot használunk tanításra, a maradék 10%-át tesztelésre és 10%-át validálásra. A C_1 és C_5 osztályok egymástól való megkülönböztetésére tanított részosztályozónál hány mintát fogunk tanításra használni? Milyen választ fog adni ez a részosztályozó a tanítás után, ha egy C_3-beli elem attribútumait kapcsoljuk a bemenetére? (**2 pont**)

5. Három eloszlásból származó kevert minta alapján becsüljük az egyes eloszlások várható értékét. Azonos, $\sigma=2$ szórású Gauss eloszlásokról van szó, és az összes eloszlásra ugyanakkora annak *a priori* valószínűsége, hogy a minta ebből az eloszlásból származik. Az EM algoritmusnál a mellékelt lapon közölt összefüggésekkel végeztük az „expectation”, illetve a „maximization” lépést.

A t-dik „maximization”lépés után a következők voltak a várható értékekre vonatkozó EM iteráció eredményei:

$\mu_1(t)=0{,}5$ $\mu_2(t)=4{,}7$ $\mu_3(t)=2{,}9$

Az eddigi tanulás eredményeként mit mondhatunk, milyen valószínűséggel származik az x=1,8 minta az első forrásból? (**2 pont**)

6. MOE struktúrát használunk bináris osztályozási feladatra. A lineáris kapuzó hálózat paramétereit tanítjuk, a két lineáris szakértő paraméterei rögzítettek. Elképzelhető-e, hogy miután ezer iterációt végeztünk (a teljes tanító halmazunkat ezerszer felhasználtuk), az egyik szakértő gyakorlatilag nem tesz hozzá a probléma megoldásához (mondjuk, kevesebbet befolyásol az eredményen, mint 0,00001%)? **(2 pont)**

7. Milyen a kapcsolat a tanított eszköz komplexitása és a túltanulási veszély közt? **(2 pont)**

8. Bináris osztályozási feladatra MINCUT eljárást használunk az $L=\{(x^{(k)},y^{(k)})\}$ 100 darab ismert besorolású minta, ahol $y^{(k)}\in\{0,1\}$, és az $U=\{x^{(j)}\}$ 1 darab ismeretlen besorolású minta alapján. Az ismert besorolású minták fele-fele arányban oszlanak meg a két osztály között. A következő veszteségfüggvényt minimalizáljuk:

$$Loss = \infty \cdot \sum_{x^{(k)}\in L} \left(f(x^k) - y^{(k)}\right)^2 + \frac{1}{10}\sum_{x^{(m)},x^{(n)}\in L\cup U} w_{mn}\cdot\left(f(x^{(m)}) - f(x^{(n)})\right)^2 \quad ; w_{mn} \geq 0\ ,$$

ahol $f(x^{(m)})$ az eljárás által adott besorolás az adott mintapontra

Az $x^{(1)}\epsilon U$ pontra az összes $w_{1,n}$ súly kisebb vagy egyenlő 0,1-nél, kivéve a $w_{1,27}$ súlyt, és tudjuk, hogy $(x^{(27)},y^{(27)}=0)\epsilon L$. Legalább mekkora kell legyen $w_{1,27}$, hogy az ismeretlen pontot biztosan a 0 címkével jelzett osztályba soroljuk, akármekkorák is a többi $w_{1,n}$ értékek (persze 0 és 0,1 között)? **(2 pont)**

9. Együttes tanulással próbálunk megoldani egy négydimenziós *regressziós* feladatot. Két szakértőt tanítunk, mindkettő kNN (k-nearest neighbour, k-legközelebbi szomszéd) regresszor, k=3. Az egyik az x_1-x_2, a másik az x_3-x_4 nézeten dolgozik. Kiinduláskor a következő pontokat ismerjük, d az ismert válasz. Az ismeretlen besorolású pontokat itt d = X jelzi.

k	$\mathbf{p}_k=[\ x_{k1}\ x_{k2}\ x_{k3}\ x_{k4}\ ]$	x_{k1}	x_{k2}	x_{k3}	x_{k4}	d_k
1	$\mathbf{p}_1$	4	1	1	2	-0,3
2	$\mathbf{p}_2$	2	1	1	3	+2,1
3	$\mathbf{p}_3$	3	1	4	3	X
4	$\mathbf{p}_4$	1	2	4	1	+3,2
5	$\mathbf{p}_5$	2	2	3	1	X
6	$\mathbf{p}_6$	3	2	5	3	+2
7	$\mathbf{p}_7$	1	3	2	1	0,25
8	$\mathbf{p}_8$	2	3	2	2	X
9	$\mathbf{p}_9$	3	3	2	4	0,1
10	$\mathbf{p}_{10}$	4	3	2	3	X
11	$\mathbf{p}_{11}$	5	3	3	2	-1,4
12	$\mathbf{p}_{12}$	2	4	3	3	-1,2

a. Rajzolja fel a két szakértő által „látott" nézeteket! **(2 pont)**

b. A két szakértő milyen értéket fog adni a $\mathbf{p}_{10}$ mintának az első tanítási lépés *előtt* (a kiinduló helyzetben)? **(2 pont)**

Jó munkát !