

Gépi Tanulás (vimim136)
Zárthelyi

2013. március 25.

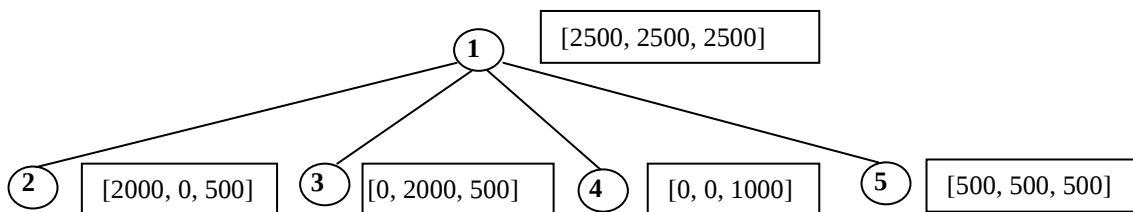
8:30 - 9:50 (80 perc !)

Minden válaszhoz rövid, de áttekinthető indoklást is kérek! Mindkét részből el kell érni a 40-40 %-ot! (Ez az első résznél 6 pont, a másodiknál 4 pont.)

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

Aláírás:.....

1. Az alábbi ábrán egy 7500 minta alapján tanítással előállított döntési fa látható. A csomópontok mellett 3 szám látható egymás mellett bekeretezve, szögletes zárójelben, ez rendre a C_1 , C_2 , illetve C_3 osztályba jutó tanítóminták száma az adott csomópontban.



Mekkora lesz az információnyereség az 1-es csomópontban elvégzett – 4 lehetséges kimenettel rendelkező – teszt elvégzésekor? **(6 pont)**

2. Egy bináris osztályozási feladatra tanított döntési fa építése során a 17-es csomópontba eljutott mintáink: 1213 tartozik az C_1 osztályba és 3785 a C_2 osztályba. A teszt elvégzése után keletkező három csomópontba a következő minták jutottak:

17/A csomópont C_1 : 403	C2: 1205
17/B csomópont C_1 : 445	C2: 1537
17/C csomópont C_1 : X	C2: Y

Úgy döntöttünk, hogy ha a null-hipotézist (a teszt nem releváns) nem tudjuk legalább 95% valószínűséggel elvetni, akkor bezárjuk a csomópontot, azaz a tesztet már nem végezzük el. A khi-négyzet szignifikancia teszt alapján megtartjuk-e a tesztet? **(4 pont)**

3. Milyen lehetőségeket ismer annak a problémának a megoldására, ha a tanulórendszerünk tanító mintáinak egy részénél hiányzik a 20 leíró paraméter (attribútum) közül néhánynak az értéke? (Legfeljebb 2 paraméter érték hiányzik, ennél soha nem több, persze lehet, hogy kevesebb, vagy egyáltalán nem hiányzik egy sem.) Még a hiányos mintáinkat se akarjuk eldobni, mert kevés a hibátlanul regisztrált minta. **(5 pont)**

4. Szakértő együttest tanítunk, az i -dik szakértő kimenetét jelölje y_i , az eredő kimenetet az egyes hálók

válaszainak súlyozott lineáris összegeként kapjuk: $y = \sum_{i=1}^K w_i \cdot y_i$, ahol $\sum_{i=1}^K w_i = 1$. Vezesse le az órán

tanult összefüggést: hogyan írható fel a teljes rendszer átlagos négyzetes hibája az egyes hálók átlagos négyzetes hibája és a végső kimenettől való eltérése alapján! Mit mutat ez az összefüggés? **(5 pont)**

5. Kétsztályos (bináris) osztályozási feladatot oldunk meg. 11 valós, skalár paramétert mértünk minden mintán, és 15.000 mintás tanítóhalmazt állítottunk össze. (Tanítómintáink eloszlása jól közelíti a lehetséges minták eloszlását.) Ezzel egy olyan eszközt tanítottunk, amely lineárisan (hipersíkkal) bontja két részre a teret, így osztályoz. A tanítóhalmazon elért pontosság 96% volt.

Az eszközünket el tudjuk adni 87 forint nyereséggel, de ha kiderül, hogy nem teljesít legalább 95% pontosságot, akkor 10.000,- Ft kártérítést kell fizetnünk. Várható értékben nyereséges lesz-e az üzlet, ha – jobb híján – a pontosságra vonatkozó felső korlátot pontos értéknek fogadjuk el? **(5 pont)**

Jó munkát !

Megadott paraméterek, összefüggések

$$I(P(v_1), \dots, P(v_K)) = - \sum_{i=1}^K P(v_i) \cdot \log_2(P(v_i))$$

$$M(A) = \sum_{i=1}^K \frac{n_{i,1} + \dots + n_{i,N}}{n_1 + \dots + n_N} \cdot I(P(v_{i1}), \dots, P(v_{iK}))$$

$$D = \sum_{i=1}^k \left(\frac{(p_i - \hat{p}_i)^2}{\hat{p}_i} + \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i} \right)$$

$$R(T_n) + \alpha \cdot |T_n| = R(\{T_n\}) + \alpha \cdot |\{T_n\}|$$

$$\frac{1}{\varepsilon} (\ln(|H|) - \ln(\delta)) \leq N$$

$$\mathbf{R}(\mathbf{w}) \leq \mathbf{R}_{\text{emp}}(\mathbf{w}) + \frac{\varepsilon(h)}{2} \sqrt{1 + \frac{4 \cdot \mathbf{R}_{\text{emp}}(\mathbf{w})}{\varepsilon(h)}}$$

$$\varepsilon(h) = 4 \cdot \frac{\ln(2 \frac{L}{h} + 1) - \ln(\frac{\eta}{4})}{L/h}$$

χ^2 táblázat

SzF	P _{0,5}	P _{0,2}	P _{0,1}	P _{0,05}	P _{0,01}	P _{0,005}	P _{0,001}
1	0,46	1,64	2,71	3,84	6,64	7,88	10,83
2	1,39	3,22	4,61	5,99	9,21	10,60	13,82
3	2,37	4,64	6,25	7,82	11,35	12,84	16,27
4	3,36	5,99	7,78	9,49	13,28	14,86	18,47
5	4,35	7,29	9,24	11,07	15,09	16,75	20,52
6	5,35	8,56	10,65	12,59	16,81	18,55	22,46
7	6,35	9,80	12,02	14,07	18,48	20,28	24,32
8	7,34	11,03	13,36	15,51	20,09	21,96	26,12
9	8,34	12,24	14,68	16,92	21,67	23,59	27,88
10	9,34	13,44	15,99	18,31	23,21	25,19	29,59