2. (Calcularea câştigului de informaţie pe "decision stumps")

Studentul Timmy doreşte să ştie cum [ar trebui să procedeze cel mai bine ca] să promoveze examenul de învăţare automată. Pentru aceasta, a cules informaţii de la studenţii care au urmat acest curs în anii precedenţi şi apoi a decis să-şi construiască un *model* bazat pe arbori de decizie. A a colectat în total nouă *instanţe*/exemple, descrise cu ajutorul a două *trăsături* (văzute mai departe ca două variabile aleatoare, S şi A: „este bine să stai şi să înveţi până noaptea târziu înainte de examen" (S) şi „este bine să mergi la toate cursurile şi seminariile" (A). Timmy dispune acum de următoarele „statistici" (care sunt de fapt *partiţionări* ale datelor sale):

$$Set(all) = [5+, 4-]$$
$$Set(S+) = [3+, 2-], Set(S-) = [2+, 2-]$$
$$Set(A+) = [5+, 1-], Set(A-) = [0+, 3-]$$

Presupunând că se foloseşte drept criteriu de selecţie a celei mai bune trăsături câştigul maxim de informaţie, ce trăsătură va alege Timmy? Care este valoarea câştigului de informaţie?

Puteţi folosi la calcule următoarele aproximaţii:

N	1	2	3	4	5	6	7	8	9
$\log_2 N$	0	1	1.58	2	2.32	2.58	2.81	3	3.17

Calculăm mai întâi entropia variabilei All.

Def. $H(X) = -\sum_x p(x) \cdot \log p(x)$.

$H(All) = H(5,4) = -\frac{5}{9} \cdot \log \frac{5}{9} - \frac{4}{9} \cdot \log \frac{4}{9} =$

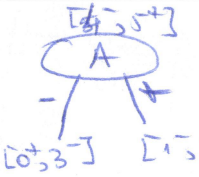
$= \frac{9\log 9 - 4\log 4 - 5\log 5}{9} = \frac{28.52932 - 8 - 11.6096}{9} =$

$\simeq 0.99107$.

$\Rightarrow H_{0|S} = \frac{4}{9} \cdot \underbrace{H(2,2)}_{1} + \frac{5}{9} \cdot H(2,3) =$

$= \frac{4}{9} + \frac{5\log 5 - 2\log 2 - 3\log 3}{9} = \frac{11.60964 - 2 - 4.75488}{9} = \frac{4}{9} =$

$= 0.98386 \Rightarrow IG(0,S) = H(All) - H_{0|S} \Rightarrow \boxed{IG(0,S) = 0.00720}$.



$\Rightarrow H_{0|A} = \frac{3}{9} \cdot \underbrace{H(0,3)}_{0} + \frac{6}{9} \cdot H(1,5) =$

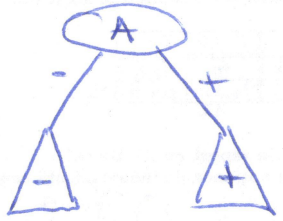
$= \frac{6}{9}\left(\frac{1}{6} \cdot \log 6 + \frac{5}{6} \cdot \log \frac{6}{5}\right) = \frac{6\log 6 - 5\log 5}{9} =$

$= \frac{15.509775 - 11.60964}{9} = 0.4333 \Rightarrow$

$\Rightarrow IG(0,A) = H(all) - H_{0|A} \Rightarrow \underline{IG(0,A) = 0.55772}$.

Cum $IG(0,A) > IG(0,S) \Rightarrow$

$\Rightarrow$ Arborele va arăta astfel:



Concluzia: pt. a promova la examen, este bine să mergi la toate cursurile şi seminariile.

Foarte bine!