Învăţare automată

— Licenţă, anul III, 2018-2019, re-examinare, parţial I —

Nume student: **Grupa:**

1. **(Câştigul de informaţie — determinarea celui mai „bun“ atribut; arbori de decizie consistenţi cu un set de date)**

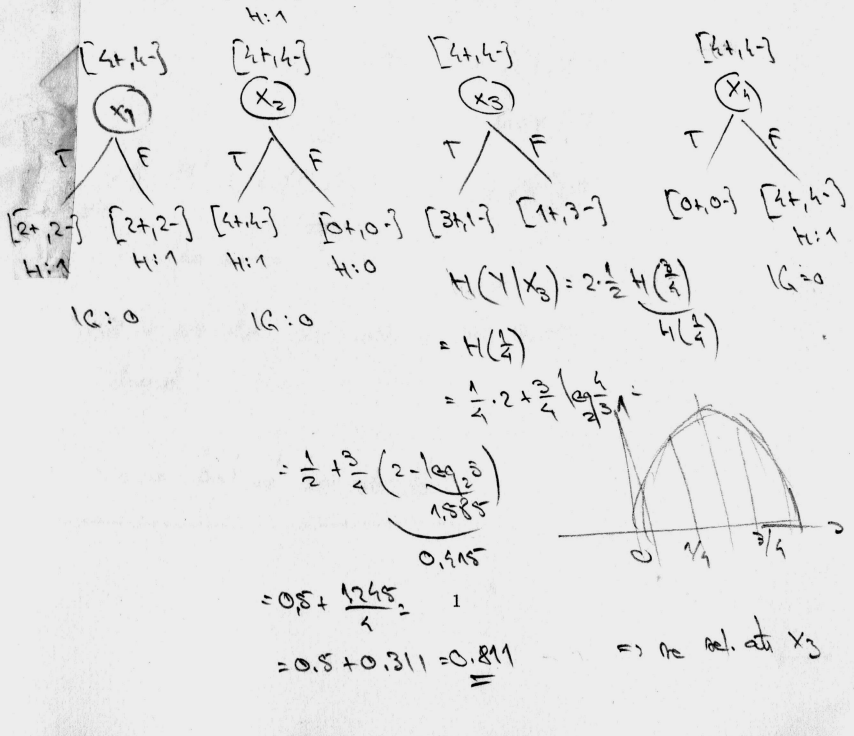
Eticheta instanţei	X_1	X_2	X_3	X_4
1	T	T	T	F
1	T	T	T	F
1	F	T	T	F
1	F	T	F	F
0	T	T	F	F
0	T	T	F	F
0	F	T	F	F
0	F	T	T	F

Fie atributele binare X_1, X_2, X_3, X_4 (ale căror valori pot fi doar T sau F), precum şi două tipuri de etichete, 0 şi 1. Veţi considera cele 8 instanţe din tabelul alăturat.

a. Vrem să învăţăm un arbore de decizie din acest set de exemple. În vederea selectării celui mai bun candidat pentru nodul rădăcină, calculaţi *câştigul de informaţie* pentru fiecare atribut X_i, cu $i = 1, \ldots, 4$. În prealabil, veţi desena compasul de decizie corespunzător fiecărui atribut. (La calcul câştigului de informaţie puteţi folosi aproximaţia următoare: $\log_2 3 = 1.585$.) Ce atribut veţi selecta?

***Observaţie*: Veţi scrie *definiţia* câştigului de informaţie. În cazul în care pentru calculul câştigurilor de informaţie veţi folosi o altă formulă decât definiţia, vă cerem să demonstraţi formula respectivă.**

b. Există oare un arbore de decizie care poate clasifica în mod perfect instanţele date? În cazul afirmativ, desenaţi acel arbore de decizie. În caz contrar, daţi o explicaţie simplă.

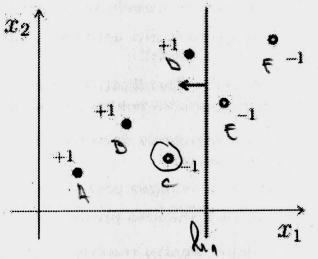


2. **(AdaBoost: întrebări în legătură cu aplicarea algoritmului pe un set de date din $\mathbb{R}^2$)**

Folosind algoritmul AdaBoost, vrem să obţinem un ansamblu de compaşi de decizie (engl., decision stumps) h_t, de forma

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right).$$

În figura alăturată sunt desenate câteva puncte (instanţe) etichetate în planul bidimensional, precum şi primul compas de decizie care a fost ales de către algoritmul AdaBoost. Un compas de decizie oarecare produce valori binare ± 1, ţinând cont doar de un anumit *prag* (engl., the split point). Săgeata mică din figură, care este perpendiculară pe dreapta care reprezintă compasul de decizie indică *zona de decizie* pentru care compasul de decizie va produce valoarea $+1$.



a. Încercuiţi toate acele instanţe din figură pentru care ponderea / probabilitatea [atribuită de către AdaBoost] va creşte ca urmare a incorporării [în ipoteza combinată H a] primului compas de decizie. Justificaţi răspunsul în mod riguros.

b. Desenaţi pe aceeaşi figură un compas de decizie care va putea fi selectat la următoarea iteraţie a algoritmului AdaBoost. Veţi trasa atât dreapta care reprezintă compasul de decizie cât şi [o săgeată care să indice] zona sa de decizie pozitivă. Justificaţi în mod riguros.

c. Va fi oare coeficientul / votul α_2, care este asociat celui de-al doilea compas de decizie mai mare decât α_1, coeficientul [din ansamblul H] pentru primul compas de decizie? Cu alte cuvinte, vom avea oare $\alpha_2 > \alpha_1$? Justificaţi în mod riguros.