Disciplina: Învățare automată **Timp: 1 oră și 10 minute**	**TEST**	**Probabilități și statistică** **ID3** **Clasificare bayesiană** **AdaBoost** **Clusterizare ierarhică**

1. (0.05p) Verificați dacă următoarea definiție pentru funcția masă de probabilitate pentru variabila aleatoare X [care ia valori în mulțimea {1,2,3}] este validă:
   p(1) = 0.2; p(2) = 0.3; p(3) = 0.4.
2. Fie următorul experiment aleator: aruncarea unei monede. Fie X variabila aleatoare asociată [cu valorile posibile 0 și 1: 0 pentru *tails*, 1 pentru *heads*].
   a. (0.05p) **Presupunând** că rezultatele sunt echiprobabile, asignați probabilități ca X să ia o anumită valoare.
   b. (0.1p) Experimentul a fost repetat de 10 ori și s-au înregistrat următoarele date: 1,1,1,1,1,1,1,0,0,1. **Estimați** în sensul verosimilității maxime (MLE) probabilitățile ca X să ia o anumită valoare.
3. (0.15p) Presupunem ca functia densitate de probabilitate a unei variabile aleatoare continue X este definita astfel: $$p(x) = \begin{cases} \frac{4}{3}(1-x^3) & \textbf{pentru } 0 \le x \le 1 \\ 0 & \textbf{în caz contrar.} \end{cases}$$. Cat este P(X<0.5)?
4. (0.2p) Fie X si Y variabile aleatoare continue avand functia densitate de probabilitate comuna definita astfel: $$p(x,y) = \begin{cases} 1 & \textbf{pentru } 0 < x < 1,\ \mid y \mid < x \\ 0 & \textbf{în caz contrar.} \end{cases}$$.
   a) Cat este p(y|x=0.5)?
   b) Este variabila X independenta de variabila Y?

5. (0.2p) Fie o variabila aleatoare cu distributia de probabilitate uniforma, definita pe intervalul [0,½], si anume f(X)=2 pentru x din intervalul [0,½]. Intrucat suma probabilitatilor care corespund unei distributii aleatoare trebuie sa fie 1, Mike este nedumerit de ce valoarea lui f(x) este mai mare decat suma totala. Explicati acest paradox. Altfel spus, aratati ce anume nu stie Mike.

6. (0.3p) Formula de inlanturie a entropiilor pentru cazul general este:
   $$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y).$$
   a) (0.2p) Demonstrati ca daca X si Y sunt variabile aleatoare independente discrete, atunci H(X,y)=H(X)+H(Y)
   b) (0.1p) Este adevarata si reciproca acestei afirmatii? Adica, atunci cand are loc egalitatea demonstrata la punctul a) rezulta ca variabilele X si Y sunt independente?

7. (0.5p) În cadrul algoritmului **ID3**, fie următorul tabel în care doar atributele C și D sunt **continue**:

A	B	C (continuu)	D (continuu)	Y (output)
0	0	2	1	0
0	1	2	2	1
1	0	1	3	1
1	1	5	4	1
1	0	6	4	0
0	1	5	6	0

a. (0.15p) Câte praguri distincte trebuie să considerăm pentru C atunci când căutăm atributul (optim) care trebuie pus în rădăcină? Dar pentru D? **Justificați**.

b. (0.05p) Care sunt nodurile ce trebuie luate în calcul atunci când se caută nodul rădăcină? **NU mai scrieți partițiile [a+,b-], ci doar numele din cercuri.**

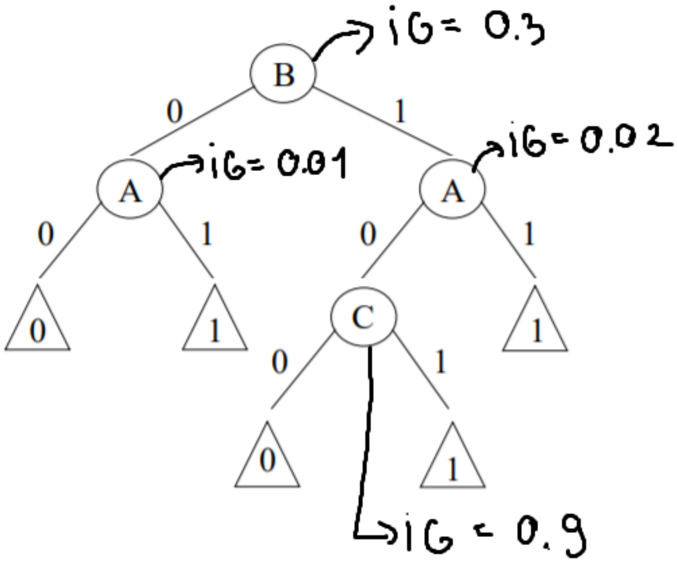
c. Presupunem că valorile câștigurilor de informație dintre atributul de ieșire și fiecare atribut candidat pentru nodul rădăcină sunt: 0.9 pentru A, 0.8 pentru B, 0.7 pentru restul.

   i. (0.05p) Ce atribut/nod va fi ales ca rădăcină? **Justificați**.

   ii. (0.1p) Care sunt nodurile ce trebuie luate în calcul atunci când se caută nodul corespunzător ramurii *Rădăcină* == 0? **NU mai scrieți partițiile [a+,b-], ci doar numele din cercuri.**

d. (0.15p) Fie tabelul în care apar doar coloanele C și Y. Aplicăm ID3 **doar** pe aceste date. Cum vor fi clasificate instanțele (C=0) și (C=7)? **Justificați fără a face arborele!**.

8. (0.5p) Fie următorul arbore produs de ID3 pe un set de date. Pentru fiecare nod este trecut și IG-ul corespunzător calculat în cadrul algoritmului.



a. (0.2p) Pentru un astfel de arbore de decizie, o strategie simplă de trunchiere (engl., pruning) în vederea contracarării fenomenului de "overfitting" constă în a parcurge arborele de sus în jos, începând deci cu nodul-rădăcină şi identificând fiecare nod de test pentru care câştigul de informaţie are o valoare mai mică decât o valoare pozitivă, mică, fixată de la început, ε. Orice astfel de nod de test este imediat înlocuit — împreună cu subarborele corespunzător lui — cu un nod de decizie, conform etichetei majoritare a instanţelor asignate nodului de test **(în acest exercițiu veți lăsa nodul de decizie, adică triunghiul, gol, pentru că nu aveți informații suficiente)**. Această strategie se numeşte "top-down pruning".
Care este arborele de decizie obţinut aplicând această strategie pe arborele de mai sus, dacă se consideră ε = 0.03?
b. (0.2p) O altă posibilitate de a face pruning este să parcurgem arborele de decizie începând cu părinţii nodurilor-frunză şi să eliminăm în mod recursiv acele noduri de test pentru care câştigul de informaţie (sau un alt criteriu ales) este mai mic decât ε. Aceasta este strategia de "bottom-up pruning".
Observaţie: Spre deosebire de strategia top-down, în varianta de pruning de tip bottom-up nu vor fi eliminate noduri (cu IG < ε) pentru care există descendenţi al căror câştig de informaţie este mai mare sau egal cu ε.
Ce arbore se obţine făcând "bottom-up pruning" pe arborele dat mai sus, dacă se consideră ε = 0.03?
c. (0.1p) În acest exercițiu am folosit următoarea metodă pentru a determina dacă un atribut de intrare este independent de atributul de ieșire: IG < ε, unde ε e valoare mică, pozitivă apropiată de zero. Există o altă metodă de a verifica o astfel de independență în contextul trunchierii arborilor de decizie. Menționați numele ei.

9. (0.2p) Se consideră variabilele aleatoare X1, X2, X3 şi X4. Aceste variabile sunt independente condiţional două câte două în raport cu variabila Y , cu excepţia perechii X3, X4. (Aşadar, dacă am aplica algoritmul Bayes Naiv, acesta ar produce

erori de clasificare.) Cum am putea modifica regula de decizie a algoritmului Bayes Naiv pentru a ţine cont de această particularitate a datelor?

10. (0.3p) În cadrul unui experiment de ML un student a împărțit setul de date de care dispunea în 3: antrenare, validare și testare. Studentul trebuie să aleagă între 3 modele: ID3, ID3 cu index Gini, Bayes Naiv și calculează următoarele erori la antrenare și validare după ce a antrenat pe setul de antrenare.

Model	Eroare la antrenare	Eroare la validare
ID3	0.05 = 5%	0.06 = 6%
ID3 cu index Gini	0.05 = 5%	0.4 = 40%
Bayes Naiv	0.54 = 54%	0.55 = 55%

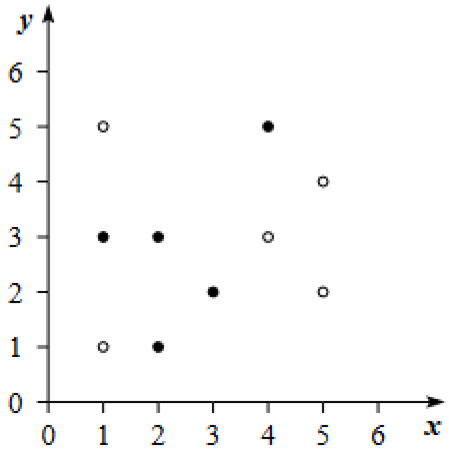
a. (0.1p) Ce model va alege din cele 3? De ce?
b. (0.1p) Ce probleme (underfitting/overfitting) prezintă celelalte două modele?
c. (0.1p) Pentru modelul cu overfitting ce s-ar putea face pentru a scăpa de această problemă?

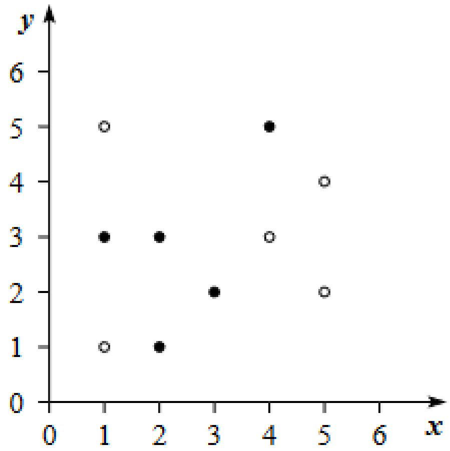
11. (0.25p) Fie următorul set de date, unde A,B,Y sunt discrete:

A	B	Y (output)
0	**0**	**0**
1	**1**	**1**
2	**0**	**1**
1	**1**	**0**

a. (0.1p) Estimați P(A = 0|Y=1) în sensul verosimilității maxime (MLE).
b. (0.15p) Estimați P(A = 0|Y=1) folosind regula *add-one* de netezire a lui Laplace.

12. (0.4p) Pe setul de date de mai jos desenaţi graniţele de decizie şi apoi haşuraţi suprafeţele de decizie produse de:
   a) algoritmul 1-NN (veţi obţine deci diagrama Voronoi);
   b) algoritmul ID3 extins cu capacitatea de a procesa atribute cu valori continue

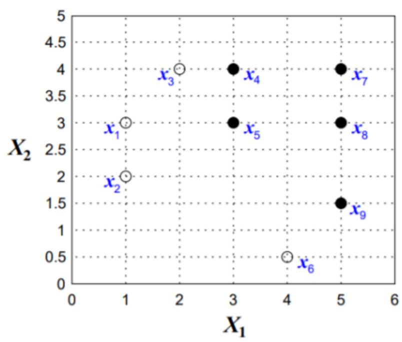




13. (0.4p) Folosind metoda 1-NN cu distanţa euclidiană, învăţăm un clasificator cu două valori pentru atributul de ieşire, Y = 0 şi Y = 1, pornind de la datele de antrenament din tabelul de mai jos (X1 şi X2 sunt atribute de intrare).
    a) (0.1p) Care este eroarea la antrenare (exprimată ca număr de exemple clasificate eronat)?
    b) (0.1p) Care este eroarea la cross-validare folosind metoda “Leave-One-Out”?
    c) (0.2p) Răspundeţi la întrebările de mai sus, considerând acum arbori de decizie cu atribute numerice continue în locul metodei 1-NN.

X_1	X_2	Y
0	0	1
1	0	1
2	0	1
2.5	2	1
3	0	1
1	2	0
1.5	0	0
2	2	0
3	2	0
4	2	0

14. (0.4p) Aplicați 2 iterații de AdaBoost (deci, până la calculul distribuției D3 inclusiv) pe următorul set de date. Care va fi eroarea la antrenare produsă de AdaBoost după



aceste 2 iterații?

15. (0.5p) Adevarat sau Fals? Justificati riguros.

    a) εt, eroarea ponderată produsă la antrenare de către ipoteza ht / clasifica-torul „slab“ A (măsurată relativ la ponderile de la începutul iteraţiei t) tinde să crească în raport cu t.

b) În decursul iteraţiilor executate de algoritmul AdaBoost, erorile ponderate εt (produse la antrenare, pe rând, de către ipotezele „slabe“ ht, în ocurenţă, compaşii de decizie) pe de o parte, şi erorile produse la antrenare de către clasificatorii combinaţi Ht pe de altă parte, variază aproximativ la fel.

c) Ponderile / „voturile“ αt asignate de către algoritmul AdaBoost clasificatorilor „slabi“ ht asamblaţi sunt întotdeauna nenegative.

d) robabilităţile / ponderile Dt(i) alocate de către algoritmul AdaBoost exemplelor de antrenament care au fost clasificate eronat [de către ipoteza

e) Întotdeauna după ce algoritmul AdaBoost execută suficient de multe iteraţii, eroarea la antrenare produsă de ipoteza combinată Ht descreşte la ovaloare care este oricât [dorim să fie] de apropiată de zero, indiferent de tipulde clasificatori „slabi“ folositi.

16. (1,2p) Tabelul de mai jos reprezintă matricea de distanţe pentru [o mulţime formată din] şase obiecte.

	A	B	C	D	E	F
A	0					
B	0.12	0				
C	0.51	0.25	0			
D	0.84	0.16	0.14	0		
E	0.28	0.77	0.70	0.45	0	
F	0.34	0.61	0.93	0.20	0.67	0

a) Aplicaţi algoritmul de clusterizare ierarhică aglomerativă pe aceste date, folosind mai întâi similaritate single-linkage şi apoi similaritate complete-linkage.La fiecare pas al algoritmului, rescrieţi în mod corespunzător matricea de distanţe. (De la o iteraţie la alta, se micşorează cu 1 numărul liniilor precum şi al coloanelor folosite.) La final, desenaţi dendrogramele rezultate [ndicaţie: Înălţimea corespunzătoare fiecărui cluster non-singleton (adică, a fiecărui nod intern) din dendrogramă va fi considerată ca fiind egală cu distanţa(i.e., conform măsurii de similaritate) dintre cele două sub-clustere constitutive.]

b) Dacă aţi lucrat corect, atunci cele două dendrograme obţinute la punctul a nu coincid [nici măcar] ca structură. Modificaţi două valori din matricea de distanţe dată mai sus, în aşa fel încât de data aceasta cele două dendrograme care obţin să fie identice ca structură

c) Procedaţi similar pentru average-linkage. La actualizarea matricei de distanţe (sau, de „proximitate“) veţi ţine cont de formula:

$$\Delta(X \cup Y, Z) \stackrel{def.}{=} \frac{1}{(|X| + |Y|)\,|Z|} \sum_{x \in X \cup Y} \sum_{z \in Z} d(x, z)$$

$$\stackrel{calcul}{=} \frac{1}{|X| + |Y|}(|X|\,\Delta(X, Z) + |Y|\,\Delta(Y, Z)).$$

unde X,Y şi Z sunt clustere disjuncte două câte două, iar notaţia |X| desemnează cardinalul lui X (adică, numărul de elemente din X).

d) Demonstraţi formula enunţată la punctul c.