Disciplina: Învățare automată **Timp: 1 oră și 30 minute** **Punctaj minim necesar:** • **Practică: 0,5p** • **Teorie: 1p** **Punctaj maxim: 6p**	**TEST**	**Probabilități și statistică** **ID3 (1)**

- **Citiți cu atenție enunțurile!**
- **De obicei, NU vi se cere să duceți calculele până la capăt!**
- **Când vi se cere să indicați o formulă o puteți indica în 2 moduri**
  - **fie prin nume, (recomand astfel pentru că veți câștiga timp)**
    - **De exemplu, pentru calculul E[X] puteți scrie deasupra egalului „definiție" și după egal doar valoarea lui P(X=1) dacă X ia valori în {0,1}.**
  - **fie prin scrierea formulei într-un caz general (adică fără numere).**

I. Practică

1. (0.2p) Look at the following Google Colab cell:

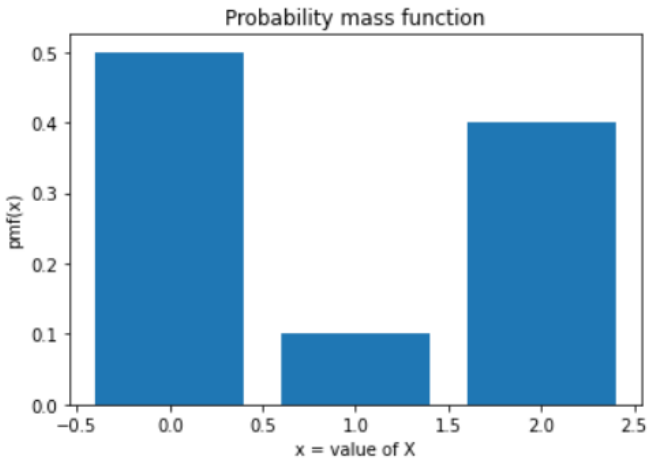
```
[17] from random import sample, seed
     seed(0)
     print(sample({1,2,3}, k=1)[0])
     print(sample({1,2,3}, k=1)[0])
     print(sample({1,2,3}, k=1)[0])
     print(sample({1,2,3}, k=1)[0])
```

```
2
2
1
2
```

What does the following code print?

```
[8]  from random import sample, seed
     seed(0)
     print(sample({1,2,3}, k=1)[0])
```

2. (0.2p) The probability mass function of a discrete random variable is represented below.



If the random variable is programatically stored in a variable X, what does the following code print?

```
print(X.prob(2).numpy())
```

3. (0.6p = 3 * 0.2p) What does the following code print? Justify your answers, by computing by hand the required values.

```
import math
import tensorflow_probability as tfp
tfd = tfp.distributions
X = tfd.Bernoulli(probs=0.5)
print(X.mean().numpy())
print(X.variance().numpy())
print(X.entropy().numpy()/math.log(2), " with 2 as the logar
ithm base.")
```

II. Teorie

1. (0.3p) Fie evenimentele aleatoare A, B, C cu $P(C) \neq 0, P(A, C) \neq 0$. Demonstrați că:

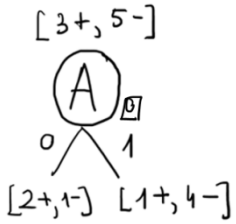
$$P(A, B|C) = P(A|C)P(B|A, C)$$

2. (0.3p) Fie X o variabilă aleatoare. Folosind definiția variației și liniaritatea mediei (E), demonstrați că:

$$\mathrm{Var}[X + 10] = \mathrm{Var}[X]$$

3. (0.3p) Doi jucători de baschet aruncă în paralel la coș (fiecare la coșul său). Primul jucător nimerește coșul cu o probabilitate de 0.1, iar cel de-al doilea cu o probabilitate de 0.7. Câte coșuri dau împreună, în medie, din 20 aruncări (de fiecare)? *[Nu vreau doar numărul final, ci și raționamentul.]*
4. (0.1p) Fie următorul experiment aleator: aruncarea a două monede. Stabiliți care sunt spațiul de eșantionare și spațiul de evenimente pentru acest experiment. *[Nu este necesar să scrieți toate elementele din spațiul de evenimente, ci doar primele și apoi scrieți „...".]*
5. Scrieți sub formă de mulțimi următoarele, unde A și B sunt evenimente:
   a. (0.1p) A și B
   b. (0.1p) fie A, fie B, dar nu ambele
6. (0.1p) Doi țintași trag la țintă. Probabilitatea ca țintașul X să nimerească ținta este de 0.1. Probabilitatea ca țintașul Y să nimerească ținta este de 0.1. Probabilitatea ca ambii țintași să nimerească ținta este de 0.01. Evenimentele „țintașul X nimerește ținta" și „țintașul Y nimerește ținta" sunt independente? De ce?
7. Fie următorul experiment aleator: aruncarea unui zar. Fie X variabila aleatoare asociată [1-6 – valoarea de pe zar].
   a. (0.1p) **Presupunând** că rezultatele sunt echiprobabile, asignați probabilități ca X să ia o anumită valoare [1,2,3,4,5,6].
   b. (0.1p) Experimentul a fost repetat de 10 ori și s-au înregistrat următoarele date: 1,2,3,4,5,6,1,1,1,1. **Estimați** în sensul verosimilității maxime (MLE) probabilitățile ca X să ia o anumită valoare [1,2,3,4,5,6].

8. Fie 2 seturi a câte 10 și, respectiv, 5 de exerciții/subiecte. Studentul X știe să rezolve 9 exerciții din primul set și 4 din al doilea set. El alege un set de exerciții și apoi, din setul respectiv, alege un subiect.
   a. (0.2p) Care este probabilitatea ca el să știe subiectul ales?
      1. Indicați formula folosită.
      2. Înlocuiți numeric în formulă, fără a ajunge la rezultatul final.
   b. (0.2p) Dacă studentul știe subiectul ales, care este probabilitatea ca subiectul să provină din primul set?
      1. Indicați formula folosită.
      2. Înlocuiți numeric în formulă, fără a ajunge la rezultatul final.

9. (0.5p) Fie următorul compas de decizie: 

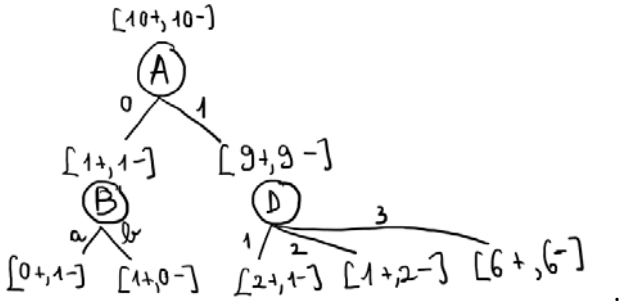
. Calculați entropia condițională medie asociată acestuia ($H_{0|A}$).
   a. Înlocuiți numeric în formula folosită până ajungeți la logaritm. (NU duceți calculele până la capăt!)

10. Fie două variabile aleatoare X și Y. În tabelul de mai jos este dată distribuția probabilistă comună a acestor două variabile.

	X = 1	X = 2
Y = 0	0.4	0.2
Y = 1	0.1	0.1
Y = 2	0.1	0.1

   a. (0.2p) Calculați Cov[X,Y].
      1. Indicați formula folosită.
      2. Înlocuiți numeric în formulă. Duceți calculele până la capăt.
   b. (0.2p) Sunt variabilele X și Y independente?
      1. Scrieți definiția.
      2. Verificați definiția.
   c. (0.3p) Calculați H[Y|X]. *[NU trebuie să duceți calculele până la capăt. Doar înlocuiți numeric în formulă.]*

11. Fie următorul arbore:



a. (0.1p) Etichetați corespunzător nodurile frunză.
b. (0.1p) Cum va fi clasificată instanța (A=0,B=a,C=1,D=2)?
c. (0.1p) Cum va fi clasificată instanța (A=1,B=a,C=2,D=1)?

12. Fie trei variabile aleatoare X, Y și Z. În tabelul de mai jos este dată distribuția probabilistă comună a acestor trei variabile.

	Z = 0		Z = 1	
	X = 0	X = 1	X = 0	X = 1
Y = 0	0.10	0.20	0.10	0.20
Y = 1	0.10	0.10	0.10	0.10

a. (0.1p) Indicați P(X=0,Y=0,Z=0).
b. (0.1p) Calculați P(Z=0).
    1. Indicați formula folosită.
    2. Înlocuiți numeric în formulă. Duceți calculele până la capăt.
c. (0.1p) Calculați P(X=0,Y=0|Z=0).
    1. Indicați formula folosită.
    2. Înlocuiți numeric în formulă. Duceți calculele până la capăt.
d. (0.1p) Sunt variabilele X și Y independente condițional față de Z?
    1. Scrieți definiția.
    2. Verificați definiția.

13. În cadrul algoritmului ID3, fie următorul tabel:

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

a. (0.1p) Setul de date este consistent?
b. (0.2p) Desenați compașii de decizie ce trebuie luați în calcul atunci când se caută nodul rădăcină.
c. (0.2p) Presupunem că valorile entropiilor condiționale medii dintre atributul de ieșire și fiecare atribut candidat pentru nodul rădăcină sunt: 0.8 pentru A, 0.7 pentru B. Desenați arborele care rezultă în urma aplicării algoritmului ID3, considerând adevărate informațiile din fraza anterioară. (nu trebuie să faceți calcule).

14. Fie următoarele două coloane dintr-un set de date (X și Y sunt discrete):

...	X	...	Y	...
...	0	...	0	...
...	1	...	1	...
...	0	...	2	...
...	1	...	0	...
...	0	...	1	...
...	1	...	2	...

    a. (0.1p) Considerând atributul Y atribut de ieșire, desenați compasul de decizie asociat lui X.
    b. (0.1p) Considerând atributul X atribut de ieșire, desenați compasul de decizie asociat lui Y.

15. Marcați cu adevărat sau fals următoarele afirmații/proprietăți (nu trebuie justificări):
    a. (0.05p) $P(A \cup B \cup C) > P(A) + P(B) + P(C)$
    b. (0.05p) O variabilă aleatoare distribuită Bernoulli are 3 valori posibile.
    c. (0.05p) Entropia lui X care ia valori în {1,2,3,4,5,6,7,8} nu poate depăși valoarea 2.
    d. (0.05p) Câștigul de informație poate fi 0.
    e. (0.05p) Dacă X, Y sunt variabile aleatoare independente, atunci IG(X;Y) = 0.
    f. (0.05p) Cov(X,Y) = 0 => X, Y – variabile aleatoare independente.
16. În legătură cu algoritmul ID3 fără extensia legată de lucrul cu atribute de intrare continue, marcați cu adevărat sau fals următoarele afirmații (nu trebuie justificări):
    a. (0.05p) Algoritmul ID3 nu garantează obținerea arborelui de decizie optimal (ca număr de niveluri sau noduri).
    b. (0.05p) Dacă nu mai sunt atribute candidat de testat într-un nod (pentru că toate atributele de intrare apar deja pe drumul de la acel nod la rădăcină), algoritmul ID3 nu mai continuă pe acel drum.
    c. (0.05p) Un atribut de intrare nu poate apărea în arborele produs de algoritmul ID3 de mai multe ori.
    d. (0.05p) ID3 este un algoritm de programare dinamică.