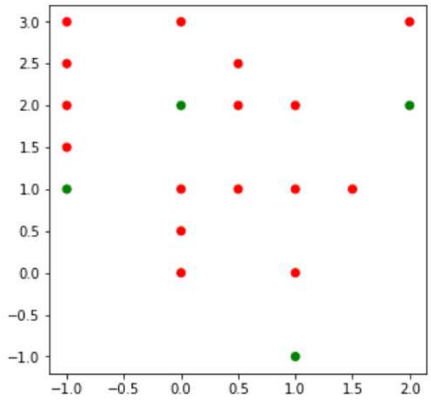
Final exam. Practical exercises

Total: 5p. Minimum: 1.25p

1. (0.5p) The following dataset is retained in the code variables X and Y.

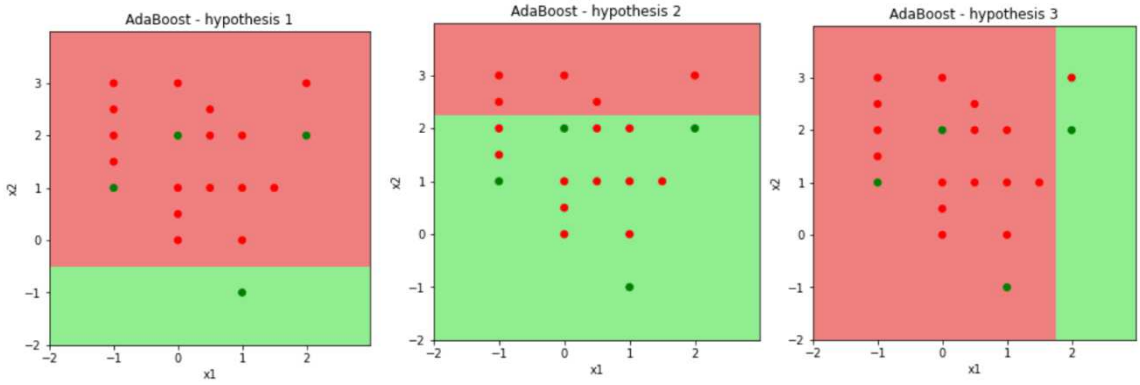


We run the following code

```
# [...] X and Y are initialised as usual with the features
# and the classes, respectively
from sklearn.ensemble import AdaBoostClassifier
ab = AdaBoostClassifier(n_estimators=3,algorithm="SAMME").fit(X, Y)
alphas = ab.estimator_weights_/2
print(alphas)
```

and obtain `[0.86730053 0.36879947 0.44804401]`.

Furthermore, we visualize the decision surfaces of each hypothesis/estimator h_t in the AdaBoost ensemble:



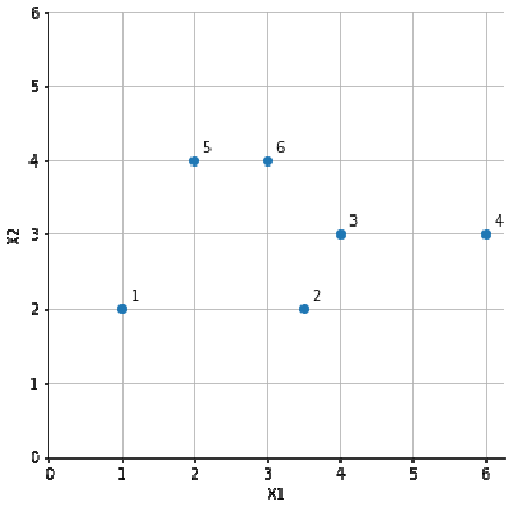
a. What is the output of the following code? The red class is encoded by 1 and the green class by 0. Justify your answer.

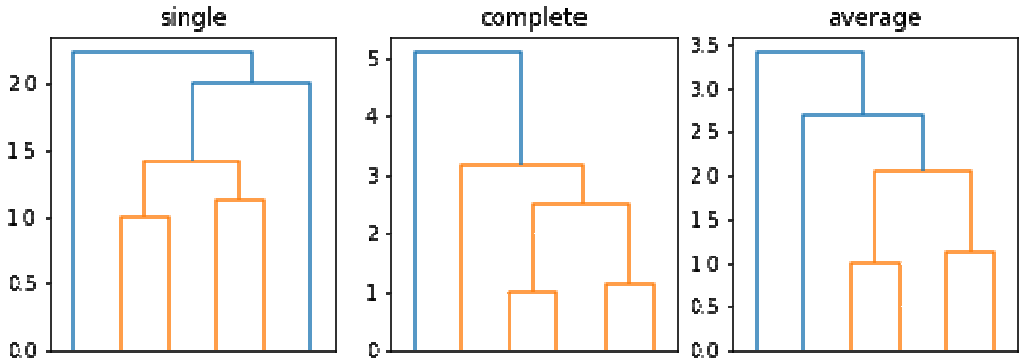
```
ab.predict([[2.0,2.0]])
```

b. Draw the decision boundary of the AdaBoost model/ensemble. You do not need to draw the points. Justify your answer.

2. (0.9p) For the dataset below, the dendrogram for agglomerative clustering has been generated for each type of linkage, but the labels of the dendrogram are missing.

What are the labels for each?





3. (1.2p) What does the following code print? Justify your answer!

```
import pandas as pd
import numpy as np
from sklearn.cluster import KMeans

d = pd.DataFrame({
    'X': [0.0, 4.0, 6.0, 8.0, 10.0],
})

init_centroids = np.array([[-2.0], [20.0]])
model = KMeans(init=init_centroids, n_clusters=2, max_iter=100)
clusters = model.fit_predict(d)
print(clusters)
```

4. (1p) You are using the standard KMeans algorithm to cluster a group of students based on the scores they obtained over four different tests. During the execution of the algorithm, you have the following clusters and centroids. What is the value of the inertia (or J criterion) at this point?

Cluster 1

T1	T2	T3	T4
6	9	10	7
5	9	9	7
7	9	8	8

Cluster 2

T1	T2	T3	T4
8	6	9	9
7	7	9	7
8	7	9	8

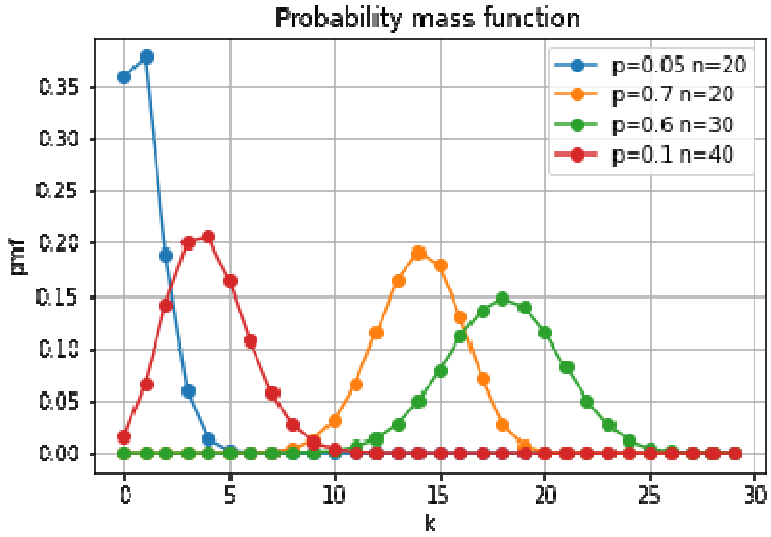
Centroid 1

T1	T2	T3	T4
6	9	9	7

Centroid 2

T1	T2	T3	T4
8	7	9	8

5. (0.5p) A website has roughly the same number of n visitors every day and the same conversion rate p. You learn that this is a binomial distribution of parameters p and n, where a success means that a visitor has made a purchase, thus becoming a customer. If the website is seeing 10 customers in one day and 15 in the next, which of the following four sets of parameters is most likely to apply to this website? *(Please include calculations and reasoning!)*



6. (0.9p) Apply one iteration, i.e. the E step and the M step, of the EM algorithm for the mixture of two Bernoulli distributions:

$$\pi \cdot \underbrace{\text{Bernoulli}(\theta_1)}_{\text{Bernoulli}_1} + (1-\pi) \cdot \underbrace{\text{Bernoulli}(\theta_2)}_{\text{Bernoulli}_2}.$$

Pieces of information:

- We are at iteration t.
- The dataset with which we work is D = {0,0,1}.
- The parameters determined at the previous iteration (t-1) are those in the the following mixture:

$$0.6 \cdot \text{Bernoulli}(0.3) + 0.4 \cdot \text{Bernoulli}(0.8)$$

- At the E step you will have to complete the remaining boxes in the following table.

$\gamma_{ij} = \mathbb{E}_{Z\|X,\pi^{(t-1)},\theta_1^{(t-1)},\theta_2^{(t-1)}}[Z_{ij}] = \begin{cases} \mathbb{E}[z_i\|x_i]^{(t)} & \text{if } j=1 \\ 1-\mathbb{E}[z_i\|x_i]^{(t)} & \text{if } j=2 \end{cases}$	Bernoulli$_1$ (j=1)	Bernoulli$_2$ (j=2)
$x_1 = 0$	0.84	
$x_2 = 0$		
$x_3 = 1$		

- At the M step you will have to **apply** (not to derive) the formulas already known by you. **Do not compute the final result; just put the numbers in the formulas.**

Teorie: 6.35p. Minim: 1.25p

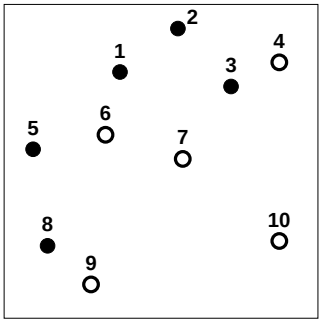
Învăţare automată

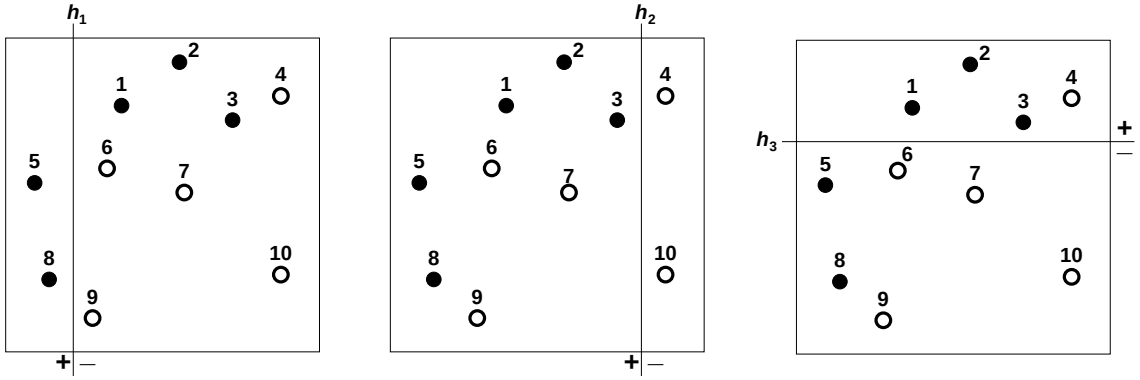
— Licenţă, anul III, 2021-2022, examenul parţial II (seria E) —

Nume student: **Grupa:**

1. [3.5p] **(Algoritmul AdaBoost: aplicare pe un set de date din $\mathbb{R}^2$)**

Se consideră că aplicăm algoritmul AdaBoost pe dataset-ul din figura alăturată. (Pentru uşurinţa exprimării la calcule, am notat pe figură indicii instanţelor de antrenament, în imediata apropiere a acestora.)
Folosim convenţia noastră obişnuită de notare: simbolul $\bullet$ desemnează instanţe pozitive, iar simbolul $\circ$ instanţe negative.
La primele trei iteraţii ale algoritmului au fost selectaţi compaşii de decizie h_1, h_2 şi h_3 (în această ordine), aşa cum se indică în figurile de mai jos.





Obiectivul **acestei probleme este să determinăm dacă la sfârşitul celor trei iteraţii algoritmul AdaBoost reuşeşte să clasifice perfect toate instanţele de antrenament.**

a. Calculaţi

– distribuţiile probabiliste corespunzătoare celor 3 iteraţii, adică $D_1(x_i)$ $D_2(x_i)$ şi $D_3(x_i)$, pentru $i = 1, \ldots, 10$;
– pentru fiecare dintre cele 3 iteraţii ($t = 1, 2, 3$): eroarea ponderată la antrenare (ε_t) produsă de compasul de decizie h_t, precum şi ponderea (α_t) asociată ipotezei / compasului de decizie h_t.

Veţi completa tabelele următoare şi veţi indica succint(!) modul în care aţi procedat pentru a ajunge la rezultatele respective.

i	1	2	3	4	5	6	7	8	9	10
$D_1(x_i)$										
$D_2(x_i)$	1/6			1/14						
$D_3(x_i)$	7/66			1/22		1/6				

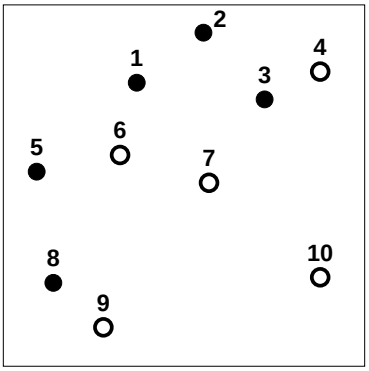
t	1	2	3
ε_t			
α_t			

***Atenţie*! Pentru a vă uşura munca, am completat noi câteva dintre elementele primului tabel. Vă puteţi baza pe valorile indicate de noi, ca să vă simplificaţi raţionamentele / calculele!**

b. Folosind ipoteza combinată obţinută de algoritmul AdaBoost la finalul celei de-a treia iteraţii, stabiliţi

i. **eroarea la antrenare produsă (pentru calcularea ei, puteţi folosi tabelul de mai jos),**

ii. **zonele de decizie corespunzătoare acestui clasificator (veţi justifica modul în care aţi procedat!). Veţi indica aceste zone de decizie, precum şi graniţele de decizie, pe desenul de alăturat.**



t	α_t	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
1	$\alpha_1 = \ldots$										
2	$\alpha_2 = \ldots$										
3	$\alpha_3 = \ldots$										
	$H_3(x_i)$										

***Atenţie*! Pentru a vă uşura munca, vă furnizăm noi următoarele valori numerice:** $\ln\sqrt{7/3} \approx 0.423$, $\ln\sqrt{11/3} \approx 0.649$, $\ln\sqrt{15/7} \approx 0.923$.

2. [2.85p] **(O problemă à la C. Do şi S. Batzoglou: rezolvarea unei mixturi de doi vectori de distribuţii Bernoulli i.i.d. identificarea parametrilor şi a variabilelor neobservabile; scrierea funcţiei de log-verosimilitate a datelor complete)**

Să zicem că avem două monede, A şi B. Sarcina ta este aceea de a afla (a „estima") θ_A şi θ_B, probabilităţile de apariţie a feţei *stemă* (engl., head) pentru fiecare dintre cele două monede. Însă eu sunt cam răutacios şi nu-ţi dau voie să arunci tu monedele. În schimb, decid să procedez astfel: voi arunca eu însumi monedele şi după aceea îţi voi comunica ţie rezultatele aruncărilor. În mod concret, îţi voi spune ceva de forma următoare: am ales una dintre cele două monede (nu-ţi spun care anume),[1] am aruncat-o de 10 ori şi am obţinut în total de 7 ori stema şi de 3 ori banul. Apoi am ales din nou una dintre cele două monede (poate aceeaşi cu cea dinainte, poate nu), am aruncat-o de 10 ori şi am obţinut în total de 5 ori stema şi de 5 ori banul. În total, îţi comunic de N ori câte o astfel de informaţie. (Aşadar, la final vei dispune de rezultatele a $10\,N$ aruncări ale monedelor.)

Facem *presupunerea* că la fiecare dintre aceste N serii cele 10 aruncări sunt independente unele de altele şi că am reţinut ordinea / succesiunea rezultatelor obţinute în urma acestor aruncări şi ţi-o comunic.

a. Formulează aceasta ca pe o problemă de tip EM (Expectation-Maximization).
i. **Care sunt datele observabile?**
ii. **Care sunt variabilele neobservabile / ascunse / latente (engl., hidden variables)?**
iii. **Care sunt parametrii modelului (adică parametrii distribuţiilor probabiliste folosite)?**

b. Calculează [adică, stabileşte pas cu pas care este] expresia funcţiei de *log-verosimilitate a datelor complete* la iteraţia t, pentru această problemă.

(Atenţie! Nu ţi se cere să rezolvi efectiv problema EM.)

[1] Vom considera că moneda A este aleasă întotdeauna cu probabilitatea π, iar moneda B cu probabilitatea $1 - \pi$.