جامعة العربي بن مهيدي - أم البواقي
كلية العلوم الدقيقة، علوم الطبيعة والحياة
قسم الرياضيات والإعلام الآلي
ماستر الذكاء الاصطناعي وعلوم البيانات - سنة أولى
إمتحان في مادة **التعلم الآلي**

26

اللقب / الإسم : .....
الفوج : .....
الإمضاء : .....

1. Le responsable d'une quincaillerie fait apprendre les différents types de vis à son stagiaire. Il déverse différents sachets sur le comptoir et lui demande de les trier dans des boites neuves. Quelle méthode applique-t-il ? (1 point)

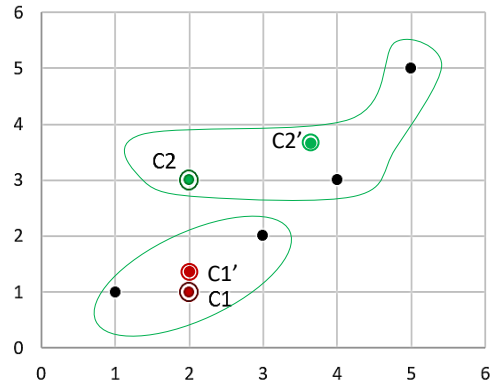
   Il utilise un apprentissage non supervisé, regroupement ou clustering

2. Le responsable ci-dessus fait la même chose avec son fils de 3 ans pour l'amuser sauf qu'il lui donne des boites sur lesquelles il y a des images de vis. Même question. (1 point)

   Il utilise un apprentissage supervisé, classification

3. Donner 3 exemples d'apprentissage par régression dont au moins un qui ne soit pas linéaire. (3 points)
   - Régression linéaire : *Prix d'une maison = a . Surface + b . Age + c*
   - Régression polynomiale (non linéaire) : *Prix d'une maison = a . Surface . Nb_pièces + b . Age² + c . localité + d*
   - Régression non polynomiale (non linéaire) : u = a . sin (b . t + c) qui pourrait modéliser la tension d'un courant alternatif où *a* est la tension max et *b* = 2π.fréquence du signal et *c* la phase du signal en radians
   - Régression logistique (non linéaire) : classer un élève de CEM en fumeur ou non en fonction de *x1* l'existence de fumeur chez lui, *x2* l'existence de fumeur parmi ses copains, *x3* l'existence de vendeurs de cigarettes à l'entrée de l'école $P(élève\ fumeur) = \frac{1}{1-e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3)}}$

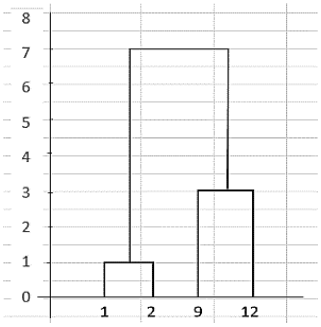
4. Diviser le nuage ci-dessous en 2 clusters avec la distance euclidienne en une itération (ne pas détailler le calcul) (4 points) :



Points	Distance avec C1 (2,1)	Distance avec C2 (2,3)	Cluster d'appartenance
(1,1)	1	2,24	Cluster de C1
C1 (2,1)	0	2	Cluster de C1
C2 (2,3)	2	0	Cluster de C2
(3,2)	1,41	1,41	Cluster de C1 oux C2. C1 considéré
(4,3)	2,83	2	Cluster de C2
(5,5)	5	3,61	Cluster de C2

C1' = ((1+2+3)/3, (1+1+2)/3) = (2, 1.33)
C2' = ((2+4+5)/3, (3+3+5)/3) = (3.67, 3.67)

5. Montrer que le dendrogramme suivant est correct (distance de Manhattan et saut minimum). Détailler les calculs pour les 2 premières cellules de la première matrice (3 points).



$\sum_{i=1}^{n} |x_i - y_i|$ |1-2|=1 |1-9|=8

Matrices des distances :

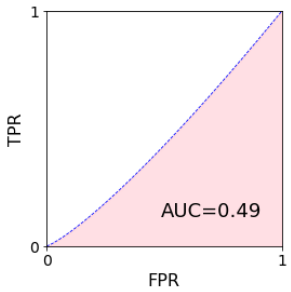
	1	2	9	12
1	0	①	8	11
2		0	7	10
9			0	3
12				0

	1,2	9	12
1,2	0	7	10
9		0	③
12			0

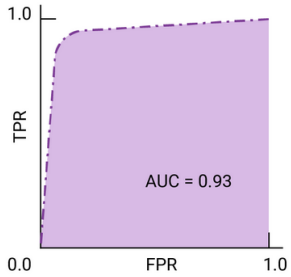
	1,2	9,12
1,2	0	⑦
9,12		0

Dendrogramme correct : regroupement de 1 et 2 (distants de 1 unité) et de 9 et 12 (distants de 3) et des 2 clusters (distants de 7)

6. Interpréter les courbes suivantes (2 points) :



- ROC : la courbe est presque égale à la diagonale y=x donc le modèle est presque équivalent au hasard
- AUC : l'aire sous la courbe est légèrement inférieure à 0.5, valeur d'un tirage aléatoire



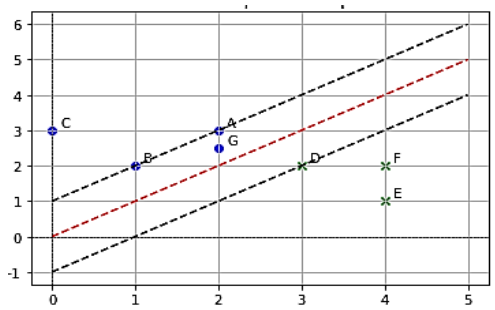
- ROC : la courbe est proche du point (0,1) indiquant que le modèle est très performant
- AUC : en cohérence avec ROC, l'aire sous la courbe est proche de 1 indiquant également que le modèle est très performant

7. Que représente *y* dans l'équation suivante, donner ses différentes valeurs : y (w.x + b) >= 1 ? (2 points)

y représente les 2 classes du modèle SVM

y peut prendre la valeur +1 (classe supérieure) ou -1 (classe inférieure)

8. Quel problème peut-on voire dans cette figure ? (3 points)



Le point G se situe entre les 2 hyperplans marginaux, zone interdite dans la version stricte du modèle SVM

Proposer une démarche pour y remédier

On peut utiliser SVM à marge souple comprenant un coefficient de relâchement : $y_i(w \cdot x_i + b) \geq 1 - \xi_i$

On peut alors régler la quantité d'erreurs tolérées grâce à l'hyperparamètre C qui intervient dans l'expression optimisé par l'algorithme : $\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i$

9. Compléter les tableaux suivants (2 points):

Classe réelle	Classe prédite	TP/FP/TN/FN
C1	C1	TP
C1	C1	TP
C2	C1	FP
C2	C1	FP
C2	C2	TN
C2	C2	TN
C1	C2	FN
C1	C2	FN

Matrice de confusion :

	C1 prédite	C2 prédite
C1 réelle	2	2
C2 réelle	2	2

10. Compléter (ID3) (5 points)

Temps	Temp.	Humidité	Vent	Pêcher
ensoleillé	chaude	haute	non	**non**
ensoleillé	chaude	haute	oui	**non**
nuageux	chaude	haute	non	**oui**

- Entropie avant division (H(S)) : $H(S) = -\sum_{i=1}^{n} p_i \log_2(p_i)$
  H(S) = - p1 * $\log_2$ (p1) - p2 * $\log_2$ (p2)
  H(S) = - 9/14 * $\log_2$ (9/14) - 5/14 * $\log_2$ (5/14) = **0.94**
- Entropie de chaque sous-groupe de chaque valeur de l'attribut Temps (H(Sv)) :

Temps	Pêcher	Pêcher	$\|S_v\|$	$H(S_v)$

pluvieux	douce	haute	non	**oui**
pluvieux	fraîche	normale	non	**oui**
pluvieux	fraîche	normale	oui	**non**
nuageux	fraîche	normale	oui	**oui**
ensoleillé	douce	haute	non	**non**
ensoleillé	fraîche	normale	non	**oui**
pluvieux	douce	normale	non	**oui**
ensoleillé	douce	normale	oui	**oui**
nuageux	douce	haute	oui	**oui**
nuageux	chaude	normale	non	**oui**
pluvieux	douce	haute	oui	**non**

	= oui	= non		
ensoleillé	2	3	5	- 2/5 * $\log_2$ (2/5) - 3/5 * $\log_2$ (3/5) = **0.971**
nuageux	4	0	4	- 4/4 * $\log_2$ (4/4) - 0/4 * $\log_2$ (0/4) « = » **0**
pluvieux	3	2	5	- 3/5 * $\log_2$ (3/5) - 2/5 * $\log_2$ (2/5) = **0.971**

- Entropie après division (Hp(S)) : $H_p(S) = \sum_{v \in A} \frac{|S_v|}{|S|} H(S_v)$

  Hp(S) = 5/14 * 0.971 + 4/14 * 0 + 5/14 * 0.971 = **0.694**

- Gain d'information pour l'attribut Temps : $\mathrm{IG}(\mathrm{S}, \mathrm{A}) = \mathrm{H}(\mathrm{S}) - \mathrm{H}_\mathrm{p}(\mathrm{S})$

  IG(S, Temps) = 0.940 – 0.694 = **0.246**