Examen 2e session (1h30) - 22 juin 2021

Rappels : **Aucun document n'est autorisé.** Les calculatrices et autres appareils électroniques doivent être éteints et rangés. Le barème (sur 22) n'est donné qu'à titre indicatif.

Exercice 1 *Huber Loss (5 points)*

On hésite souvent en machine learning entre l'optimisation des fonctions de coût L_1 (écarts en valeur absolue) et L_2 (écarts des moindres carrés). Pour mêler le meilleur des deux mondes, les chercheurs utilisent parfois la *Huber loss*, définie comme suit :

$$L_\delta(a) = \begin{cases} \frac{1}{2}a^2 & \text{pour } |a| \leq \delta, \\ \delta(|a| - \frac{1}{2}\delta), & \text{sinon.} \end{cases}$$

Q. 1. Afin de prendre le temps d'étudier la fonction, donner la valeur du coût si $\delta = 0.3, y = 2, \hat{y} = 3$

Q. 2. Montrer que cette fonction coût est continue (indépendamment du choix de δ).

Q. 3. Proposer une implémentation d'algorithme stochastique (type perceptron) reposant sur cette fonction de coût et sur un modèle linéaire $f(\mathbf{x}) = \mathbf{wx}$.
Note : question difficile demandant le calcul du gradient.

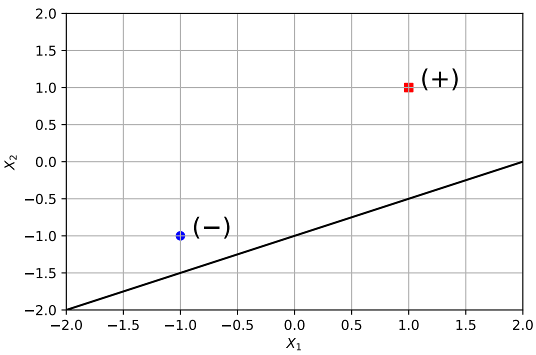
Q. 4. Quel est l'intérêt d'une telle fonction de coût selon vous ?

Exercice 2 *Classification et frontière de décision (5 points)*

Soit un jeu de données supervisé

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,d} \\ x_{2,1} & x_{2,2} & \dots & x_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,d} \end{pmatrix}, Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, y_i \in \{-1, 1\}$$

Les données et la frontière de décision associée à $f(\mathbf{x}) = \mathbf{wx} + b$ sont représentées ci-contre.



Q. 1. Donner les valeurs de d et n correspondant à l'illustration ci-dessus.

Q. 2. Donner les valeurs de $\mathbf{w}$ et b associés au tracé de la frontière [plusieurs valeurs sont possibles, mais l'une des solutions est plus facile à calculer].

On se fixe la fonction de coût de la première session :

$$C_{L_1} = \sum_i |f(\mathbf{x}_i) - y_i|$$

Q. 3. Rappeler le calcul du gradient sur cette fonction puis dérouler l'algorithme avec $\epsilon = 1$. Calculer les deux mises à jour des paramètres du modèle après les passages successifs sur les points $(-1, -1)$ et $(1, 1)$. Vous détaillerez en particulier la mise à jour de b.

Q. 4. Donner le code permettant de créer la base de données de la figure et d'appliquer cet algorithme. [=donner la solution en python aux questions précédentes]

Exercice 3 *Clustering (4 points)*

Soit un ensemble d'exemples $\mathbf{X} \in \mathbb{R}^{n \times d}$

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,d} \\ x_{2,1} & x_{2,2} & \dots & x_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,d} \end{pmatrix}$$

Q. 1. Quel est le but de l'algorithme des k-moyennes ?

Q. 2. Quelles sont les 2 grandes familles d'approches possibles pour déterminer des clusters à partir d'un ensemble de données ?

Q. 3. Soient A et B deux sous-ensembles d'exemples de $\mathbf{X}$ non-vides. Donner les expressions permettant de calculer la distance entre A et B pour les approches a) *complete linkage*, b) *average linkage*, c) *simple linkage* et d) *centroid linkage*. Pour chacune de ces 4 approches, faire un schéma permettant d'illustrer la distance calculée.

Exercice 4 *Apprentissage supervisé (8 points)*

On considère la base d'apprentissage représentée dans la figure donnée en Annexe. Cette base contient 20 exemples, dont la description est le couple représenté par leurs coordonnées (x, y), et la classe est soit *rond* (notée R) soit *carré* (notée C).

Dans ce qui suit : utiliser l'Annexe (à rendre) pour les réponses graphiques demandées.

Q. 1. En utilisant l'algorithme des k plus proches voisins, avec $k = 1$ et la distance euclidienne, représenter graphiquement la frontière de séparation des classes. En cas d'égalité de distances, les points de classe R seront considérés en priorité.

Q. 2. Même question mais cette fois-ci en utilisant l'algorithme des k plus proches voisins avec $k = 3$.

Q. 3. On considère les 2 seuils de coupure suivants :

— sur X : coupure c_1 : $x \leq 5.5$

— sur Y, coupure c_3 : $y \leq 3.5$

En utilisant uniquement ces 2 seuils, construire un arbre de décision binaire en donnant le détail des calculs d'entropie réalisés. Donner une représentation graphique de cet arbre et représenter graphiquement la frontière de séparation entre les classes correspondante.

Q. 4. On considère que la classe R correspond à la valeur $+1$ et la classe C correspond à la valeur -1. Sans dérouler l'algorithme, mais en justifiant votre réponse, tracer la frontière de décision qui pourrait être obtenue par l'application de l'algorithme du perceptron avec un nombre d'itérations significatif. Donner l'expression permettant de classer tout point (x_1, x_2) à l'aide de la frontière proposée.

Q. 5. Pour chacun des 4 modèles appris dans les questions précédentes (1-ppv, 3-ppv, arbre de décision, et perceptron), donner le taux d'erreur obtenu sur l'ensemble des données d'apprentissage. Quel est le modèle qui est susceptible de faire du sur-apprentissage ?

Annexes

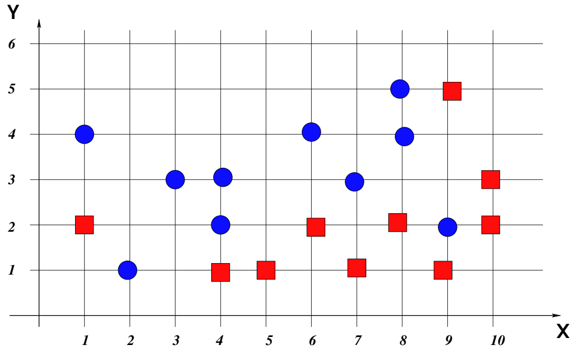
n / d		dénominateur (d)											
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
numérateur (n)	**1**	1	0,50	0,33	0,25	0,20	0,17	0,14	0,13	0,11	0,10	0,09	0,08
	2		1	0,67	0,50	0,40	0,33	0,29	0,25	0,22	0,20	0,18	0,17
	3			1	0,75	0,60	0,50	0,43	0,38	0,33	0,30	0,27	0,25
	4				1	0,80	0,67	0,57	0,50	0,44	0,40	0,36	0,33
	5					1	0,83	0,71	0,63	0,56	0,50	0,45	0,42
	6						1	0,86	0,75	0,67	0,60	0,55	0,50
	7							1	0,88	0,78	0,70	0,64	0,58
	8								1	0,89	0,80	0,73	0,67
	9									1	0,90	0,82	0,75
	10										1	0,91	0,83
	11											1	0,92
	12												1

-(n/d) * log(n/d)		dénominateur (d)											
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
numérateur (n)	**1**	0	0,50	0,53	0,50	0,46	0,43	0,40	0,38	0,35	0,33	0,31	0,30
	2		0	0,39	0,50	0,53	0,53	0,52	0,50	0,48	0,46	0,45	0,43
	3			0	0,31	0,44	0,50	0,52	0,53	0,53	0,52	0,51	0,50
	4				0	0,26	0,39	0,46	0,50	0,52	0,53	0,53	0,53
	5					0	0,22	0,35	0,42	0,47	0,50	0,52	0,53
	6						0	0,19	0,31	0,39	0,44	0,48	0,50
	7							0	0,50	0,28	0,36	0,41	0,45
	8								0	0,15	0,26	0,33	0,39
	9									0	0,14	0,24	0,31
	10										0	0,13	0,22
	11											0	0,12
	12												0

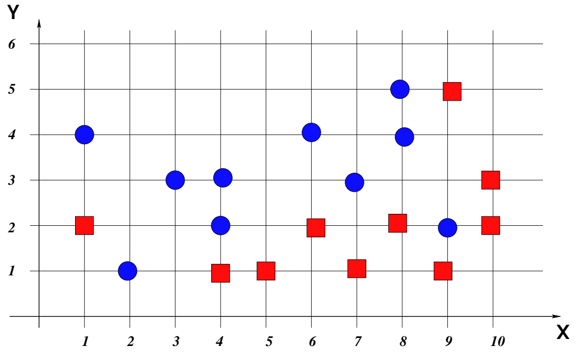
NUMERO D'ANONYMAT :

À rendre avec votre copie

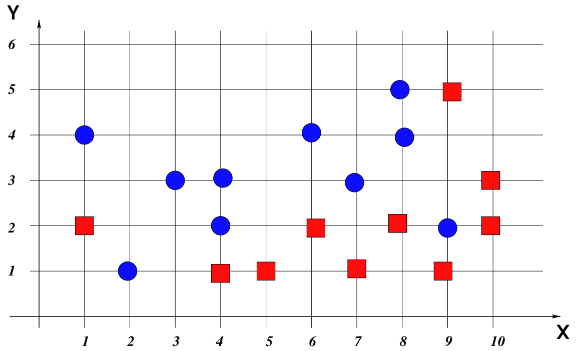
Exercice 4 - Question 1 : (1-ppv)



Exercice 4 - Question 2 : (3-ppv)



Exercice 4 - Question 3 : (arbre)



Exercice 4 - Question 4 : (perceptron)