Retours sur l'apprentissage supervisé. Clustering (1).

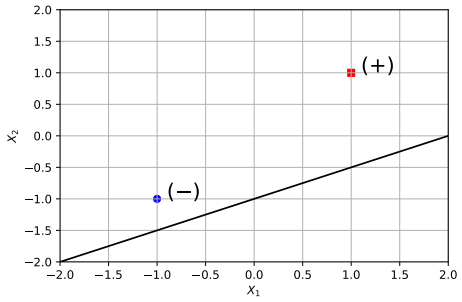
Remarque : les exercices de cette feuille sont à faire par écrit, **sans ordinateur** et sans calculatrice (pour se mettre dans les mêmes conditions qu'à l'examen écrit).

Exercice 1 *Apprentissage supervisé, frontière*
On considère une base d'apprentissage $\mathbf{X}$ contenant n exemples décrits par d variables numériques.

Question 1. Soit $\mathbf{w} = (w_1, \ldots, w_d)$ un vecteur de poids à valeurs dans $\mathbb{R}^d$ utilisé pour prendre une décision linéaire. Donner l'expression permettant de calculer le produit scalaire de chaque exemple de $\mathbf{X}$ avec $\mathbf{w}$.

Question 2. On se place maintenant en langage Python, on note `X` le `numpy.array` qui contient $\mathbf{X}$ et `w` le `numpy.array` qui contient $\mathbf{w}$. Donner les instructions python pour calculer le produit scalaire de tout vecteur de $\mathbf{X}$ par $\mathbf{w}$, sans utiliser de boucle.

Question 3. Soit un jeu de données supervisé $(\mathbf{X}, Y)$. Les données et la frontière de décision associée à $f(\mathbf{x}) = \mathbf{w}\mathbf{x} + b$ sont représentées ci-contre. Donner les valeurs de d et n correspondantes, puis 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).



Exercice 2 *Clustering*

Question 1. Montrer que la distance de Manhattan est bien une mesure de distance.

Question 2. Dans le cours, des approches ont été données pour calculer la distance entre 2 clusters :

- l'approche du chaînage minimum est appelée "simple linkage"
- l'approche du chaînage maximum est appelée "complete linkage"
- l'approche du chaînage moyen est appelée "average linkage"
- l'approche par centre de gravité est appelée "centroid linkage"

On considère une mesure de distance d entre 2 exemples et deux groupes d'exemples $A = \{a_1, a_2, ..., a_{|A|}\}$ et $B = \{b_1, b_2, ..., b_{|B|}\}$, avec pour tout $i = 1, ..., |A|$, $a_i \in \mathbb{R}^p$ et pour tout $j = 1, ..., |B|$, $b_j \in \mathbb{R}^p$. Donner l'expression de la distance D entre A et B pour chacune des 4 approches données.

Question 3. En utilisant la distance euclidienne et l'approche par centre de gravité, appliquer à la main l'algorithme de clustering hiérarchique, méthode par agglomération, sur les données fournies sur le transparent 21 du cours en utilisant la distance euclidienne et l'approche "centroid linkage". Ces données correspondent à 9 points dans l'espace $X_1 \times X_2$. Le point 1 est le point de coordonnées $(-0.5, -1.0)$, etc. Construire le dendrogramme correspondant.

Question 4. On considère la base d'apprentissage de $[0, 10] \times [0, 10]$ contenant les 7 exemples suivants : $\mathbf{X} = \{(1, 2), (1, 4), (3, 4), (3, 5), (6, 2), (6, 5), (8, 3)\}$ (remarque : on considère que cette base est déjà normalisée). En détaillant les étapes et en expliquant les calculs réalisés et les regroupements effectués, appliquer sur $\mathbf{X}$ l'algorithme de classification hiérarchique, version ascendante, en utilisant la distance euclidienne et chacune des 4 approches possibles. Pour chaque approche, donner le dendrogramme obtenu.