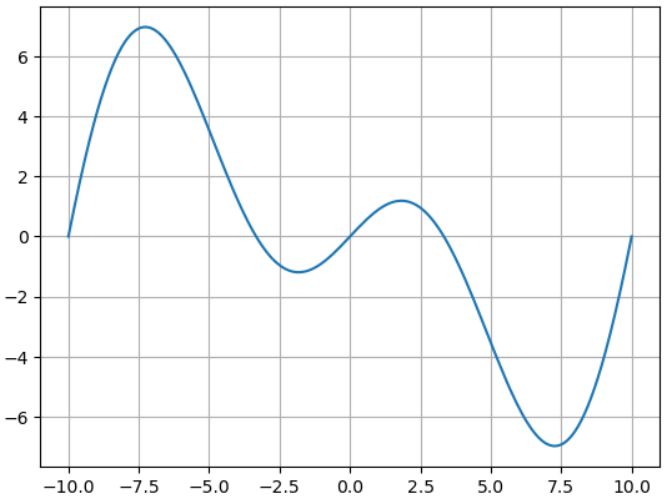
université de BORDEAUX Collège Sciences & Technologies	Année universitaire : **2024/2025** **DS Analyse, Classification, Indexation des Données** Code UE : **4TIN703U** Date : **8 novembre** Heure : **8h00** Durée : **1h30** **Documents non autorisés** **4** pages

Nom :
Prénom :
Groupe :

Exercice 1 : Descente de gradient

1. Quel est l'objectif de l'algorithme de la descente de gradient ? Est-ce que cet objectif est toujours atteint ? Rappeler les différentes étapes de la descente de gradient en donnant l'algorithme en pseudo-code. Détailler l'intérêt du pas d'apprentissage η et expliciter le critère d'arrêt qui dépendra d'une valuer ϵ.
2. On se donne la fonction $f(x) = x \cdot \cos(0.15\pi x)$ dont la représentation graphique pour $-10 \leq x \leq 10$ est la suivante :

FIGURE 1 – $f(x) = x \cdot \cos(0.15\pi x)$

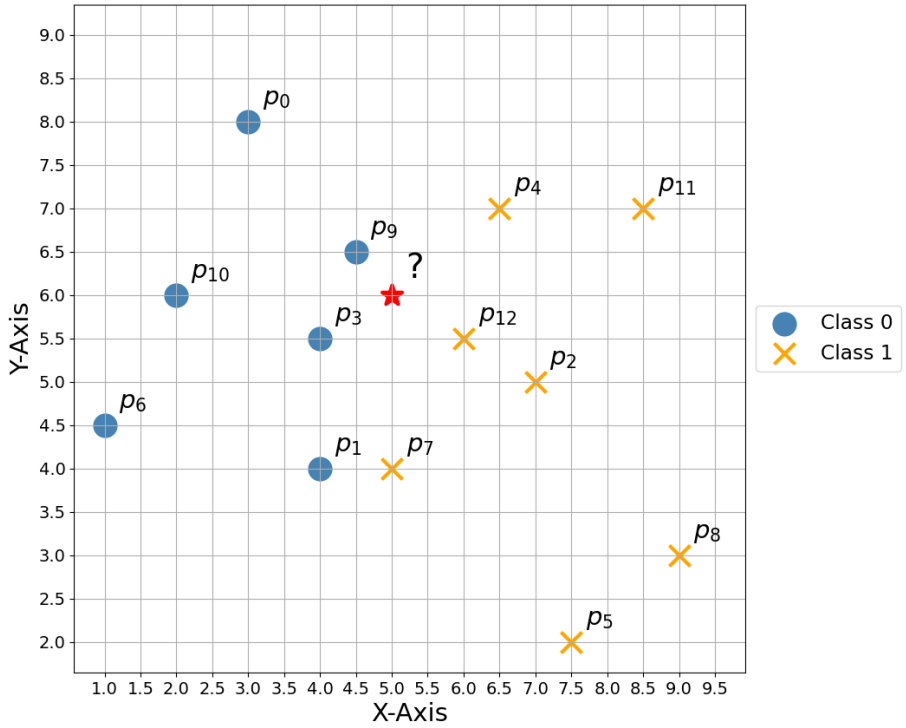


   Soit $f'(x)$ la dérivée de f par rapport à x. Sachant que $\cos(u(x))' = -u'(x) \cdot \sin(u(x))$ et que $(u(x) \cdot v(x))' = u'(x) \cdot v(x) + u(x) \cdot v'(x)$, donner la formule de $f'(x)$.
3. Appliquer l'algorithme de la descente de gradient à la fonction f en prenant comme point de départ $x = 5$ et en utilisant un pas d'apprentissage $\eta = 1$ et $\epsilon = 0.5$ comme critère d'arrêt. Pour simplifier les calculs vous pouvez arrondir tous les nombres à 3 chiffres après la virgule.
4. Que se passe-t-il si on prend comme point de départ $x = -5$ avec les mêmes paramètres η et ϵ qu'à la question précédente ?
5. Plus généralement, que peut-il se passer si le pas η est trop grand ? Trop petit ? Quelle solution peut-on envisager pour éviter ces problèmes ?

Exercice 2 : $k-$NN

On dispose d'un jeu de données composé de 13 observations $p_i, i \in \{0, 12\}$. Pour chacune de ces observations, 2 attributs numériques sont mesurés. Ces observations sont représentées par des points 2D sur la figure suivante :

FIGURE 2 – Représentation du jeu de données sous forme de points 2D



Les classes des points sont données en légende en fonction des symboles. On souhaite à présent déterminer la classe d'appartenance de la nouvelle observation ★, grâce à l'algorithme des k plus proches voisins ($k-$NN). Les attributs des points pourront être lus directement sur la figure, grâce aux coordonnées dans la grille. Dans l'algorithme de $k-$NN, on utilisera la distance de Manhattan, définie entre deux points $u = (u_1, u_2, \cdots, u_n)$ et $v = (v_1, v_2, \cdots, v_n)$ en dimension n comme suit :

$$d(u, v) = \sum_{i=1}^{n} \mid u_i - v_i \mid$$

1. Déterminez la classe de l'observation ★ en appliquant l'algorithme du $k-$NN avec $k = 5$, **et en donnant le détail des calculs**.
2. Que se passe-t-il si $k = 3$? Discutez de l'impact de k sur la prédiction.
3. Expliquez comment on peut choisir la meilleure valeur de k pour un certain jeu de données.

Exercice 3 : Régression linéaire

Une entreprise souhaite analyser les facteurs influençant les ventes de son produit dans différentes régions. On dispose des données suivantes pour chaque région :

— Dépenses publicitaires (en milliers d'euros),
— Nombre de distributeurs (en unités),
— Population de la région (en milliers d'habitants),
— Température moyenne (en Celsius) dans la région,
— Ventes (en milliers d'unités), la variable dépendante.

On ajuste un modèle de régression linéaire multiple pour prédire les ventes en fonction des quatre variables explicatives. On obtient les résultats exposés dans le rapport suivant :

OLS Regression Results

Dep. Variable:	Ventes (en milliers d'unités)	**R-squared:**	0.916
Model:	OLS	**Adj. R-squared:**	0.912
Method:	Least Squares	**F-statistic:**	258.0
Date:	Tue, 05 Nov 2024	**Prob (F-statistic):**	4.22e-50
Time:	09:48:59	**Log-Likelihood:**	-355.47
No. Observations:	100	**AIC:**	720.9
Df Residuals:	95	**BIC:**	734.0
Df Model:	4		
Covariance Type:	nonrobust		

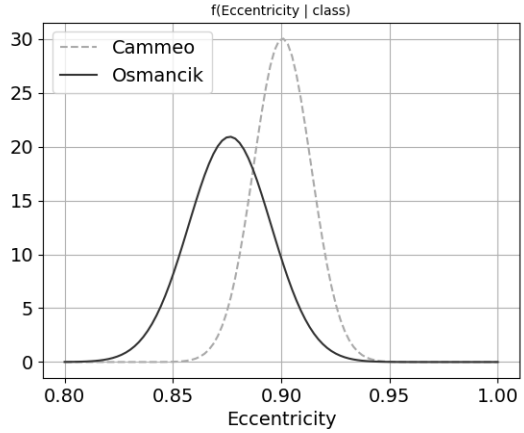
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.8220	5.970	0.305	0.761	-10.029	13.673
Dépenses publicitaires (en milliers d'euros)	0.3951	0.066	5.974	0.000	0.264	0.526
Nombre de distributeurs (en unités)	1.7579	0.313	5.621	0.000	1.137	2.379
Population de la région (en milliers d'habitants)	0.3051	0.010	29.670	0.000	0.285	0.326
Température moyenne (en Celsius)	0.3533	0.169	2.093	0.059	0.018	0.688

1. Intérprétez la première partie du rapport.
2. Quel est le modèle de régression linéaire multiple obtenu ? Donner l'équation du modèle.
3. Cette équation est-elle pertinente ? Justifiez votre réponse.

Exercice 4 : Classifieur de Bayes

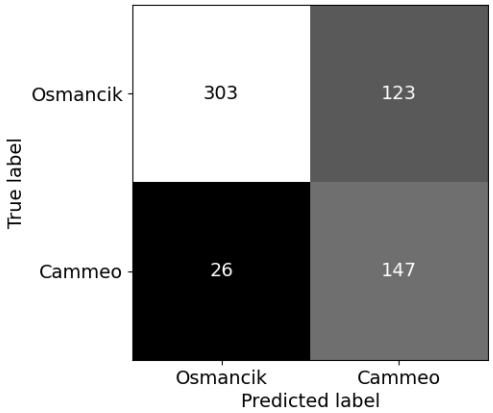
Dans cet exercice, on considère un dataset [1] dont les données sont des caractéristiques géométriques de grains de riz. Une de ces caractéristiques, appelée `eccentricity`, est une valeur numérique entre 0 et 1. Plus le grain de riz est rond, plus la valeur est proche de 0. Plus le grain de riz est allongé, plus la valeur est proche de 1. Le problème est ici de distinguer deux variétés de riz, `Cammeo` et `Osmancik`, à partir de la caractéristique `eccentricity`.

1. On veut tester le classifieur ML (maximum de vraisemblance) sur ce dataset en utilisant 80% des exemples pour l'entrainement et le reste pour les tests. On estime donc sur l'ensemble d'entrainement les paramètres de la loi normale suivie par `eccentricity` pour chacune des deux classes (`Cammeo` / `Osmancik`). On obtient la fonction de vraisemblance ci-dessous :



   Laquelle des deux variétés de riz a les grains les plus allongés ? Quelle sera la classe prédite pour un grain de riz d'`eccentricity` 0.85 ?

2. En appliquant ce classifieur à l'ensemble de test, on obtient les résultats exprimés par la matrice de confusion ci-dessous :



   Expliquez de façon détaillée par quels calculs ces résultats ont été obtenus.

   Calculez la valeur de l'accuracy à partir de ces résultats.

   Calculez les valeurs de recall et de précision en considérant d'abord que la classe `Cammeo` est la classe positive puis en considérant que la classe `Osmancik` est la classe positive.

3. Quel est le nombre d'exemples de chaque classe dans le dataset ? On suppose que la proportion des deux classes dans le dataset complet est similaire à la proportion des deux classes dans l'ensemble de test.

4. On classifie maintenant les grains de riz avec un classifieur MAP (maximum a posteriori). Les probabilités des deux classes sont estimées à partir du nombre d'exemples de chaque classe dans le dataset. Comment évolue le seuil de décision quand on passe du classifieur ML au classifieur MAP ? De même, comment évoluent les valeurs de recall et de précision (classe positive `Osmancik`) quand on passe du classifieur ML au classifieur MAP ? Expliquez.

1. https ://www.muratkoklu.com/datasets/