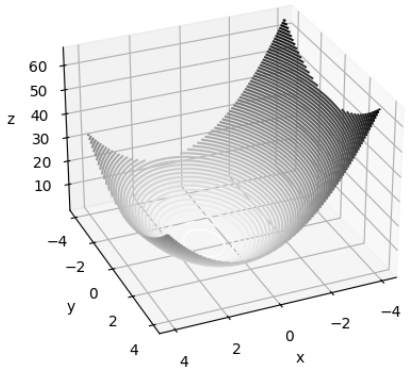
université de BORDEAUX Collège Sciences & Technologies	Année universitaire : **2023/2024** **DS Analyse, Classification, Indexation des Données** Code UE : **4TIN703U** Date : **27 octobre** Heure : **8h00** Durée : **1h30** **Documents non autorisés** **4** pages

Nom :
Prénom :
Groupe :

Exercice 1 : Descente de gradient

Soit le polynôme $p(x, y) = 2x^2 + y^2 - 4x + 2$, dont la représentation graphique est la suivante :

FIGURE 1 – $p(x, y) = 2x^2 + y^2 - 4x + 2$

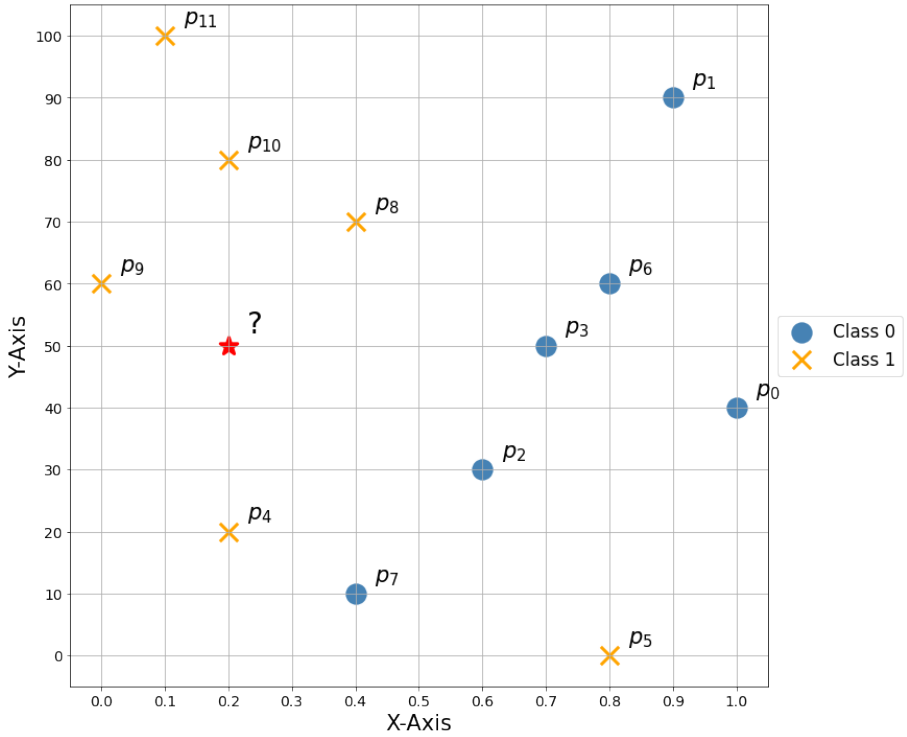


1. Qu'est-ce que le gradient de p ? Donner sa formule et sa valeur au point (1,2). On notera $\boldsymbol{\nabla} p(x, y)$ sa valeur au point (x, y)
2. Quel est l'objectif de l'algorithme de la descente de gradient ? Donner le pseudo-code de l'algorithme en utilisant comme critère d'arrêt la "norme euclidienne du gradient". On notera ϵ la valeur de la norme euclidienne du gradient à partir de laquelle on considère que l'algorithme a convergé, et η le pas d'apprentissage.
3. Appliquer l'algorithme de la descente de gradient à la fonction p en partant du point $(1.1, 0.1)$ et en utilisant un pas d'apprentissage $\eta = 0.1$ et $\epsilon = 0.2$ 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 part du point $(1, 2)$ avec un pas $\eta = 1$?
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 12 observations $p_i, i \in \{0, 11\}$. Pour chacune de ces observations, 2 attributs numériques sont mesurés. Ces observations sont représentées par des points 2D, sur la figure .

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 $v = (v_1, v_2, \cdots, v_n)$ comme suit :

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

1. Déterminez la classe de ★ en appliquant l'algorithme du $k-$NN avec $k = 5$, **et en donnant le détail des calculs**.
2. La classe de la nouvelle observation vous semble-t-elle raisonnable visuellement ? **Justifiez votre réponse**.
3. Donnez **au moins** un avantage et un inconvénient de l'algorithme du $k - NN$.

Exercice 3 : Régression linéaire [1]

Une société automobile Geely Auto aspire à pénétrer le marché américain en y installant son unité de fabrication et en produisant des voitures localement pour faire concurrence à ses homologues.

Les dirigeants de la société ont engagé une société de conseil automobile pour comprendre les facteurs dont dépend le prix des voitures.

Sur la base de diverses études de marché, la société de conseil a rassemblé un vaste ensemble de données sur différents types de voitures sur le marché américain.

Les variables récoltées sont : le `prix`, `wheelbase` (empattement), `carlength` (longueur de voiture), `carwidth` (largeur de la voiture), `carheight` (hauteur de voiture), `curbweight` (poids à vide), `enginesize` (la taille du moteur), `boreratio` (rapport d'alésage) , `stroke` (course), `compressionratio` (ratio de compression), `horsepower` (puissance), `peakrpm` (régime de pointe), `citympg` (consommation en milieu urbain) et `highwaympg` (consommation en autoroute).

On fait donc une régression linéaire et on obtient le rapport suivant :

OLS Regression Results

Dep. Variable:	price	R-squared:	0.851
Model:	OLS	Adj. R-squared:	0.841
Method:	Least Squares	F-statistic:	83.78
Date:	Tue, 24 Oct 2023	Prob (F-statistic):	1.68e-71
Time:	12:02:21	Log-Likelihood:	-1937.5
No. Observations:	205	AIC:	3903.
Df Residuals:	191	BIC:	3949.
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-4.75e+04	1.53e+04	-3.111	0.002	-7.76e+04	-1.74e+04
carlength	-94.6752	55.557	-1.704	0.090	-204.259	14.909
carwidth	505.5716	246.013	2.055	0.041	20.319	990.824
carheight	163.1801	135.721	1.202	0.231	-104.524	430.884
curbweight	1.8846	1.737	1.085	0.279	-1.542	5.312
enginesize	117.3461	13.837	8.481	0.000	90.054	144.638
boreratio	-1002.5654	1195.798	-0.838	0.403	-3361.231	1356.100
stroke	-3034.6060	778.604	-3.897	0.000	-4570.373	-1498.839
compressionratio	298.1369	82.914	3.596	0.000	134.592	461.682
horsepower	30.8086	16.216	1.900	0.059	-1.177	62.795
peakrpm	2.3751	0.671	3.540	0.001	1.052	3.698
citympg	-320.3545	177.769	-1.802	0.073	-670.998	30.289
highwaympg	202.8221	159.760	1.270	0.206	-112.299	517.943
wheelbase	122.6169	100.465	1.220	0.224	-75.546	320.780

Omnibus:	24.541	Durbin-Watson:	0.930
Prob(Omnibus):	0.000	Jarque-Bera (JB):	81.326
Skew:	0.383	Prob(JB):	2.19e-18
Kurtosis:	5.989	Cond. No.	3.94e+05

1. Interpréter la première partie du rapport. Le modèle est-il pertinent ?
2. A la lecture de la deuxième partie du rapport, quelles sont les variables qui ont un impact sur le `prix` ? (Expliquez pourquoi sans détailler, à ce stade, votre démarche)
3. Le cabinet de conseil suggère que le poids à vide et le régime de pointe n'ont pas d'impacts sur le prix. Que pensez-vous de ces affirmations ? Justifiez votre réponse de manière rigoureuse.
4. Que convient-il de faire à présent pour chercher un meilleur modèle de régression ?

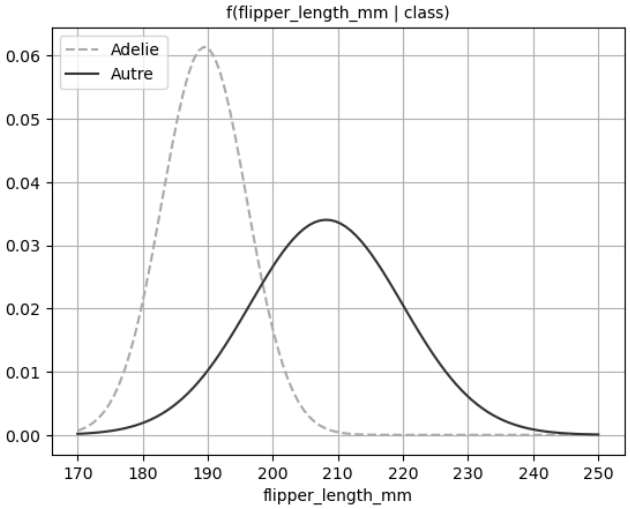
1. https ://www.kaggle.com/code/mustaphabouyardan/regression-lineaire-multiple

Exercice 4 : Classifieur de Bayes

Dans cet exercice, on considère un dataset [2] dont les données sont des mesures prises sur des manchots d'Antarctique. Ce dataset contient 4 descripteurs numériques par manchot ainsi que son espèce. Nous allons considérer un seul de ces descripteurs, la taille de la nageoire, appelé *flipper_length* dans la suite. On se pose le problème suivant : distinguer les manchots Adelie des autres espèces de manchots à partir de ce descripteur.

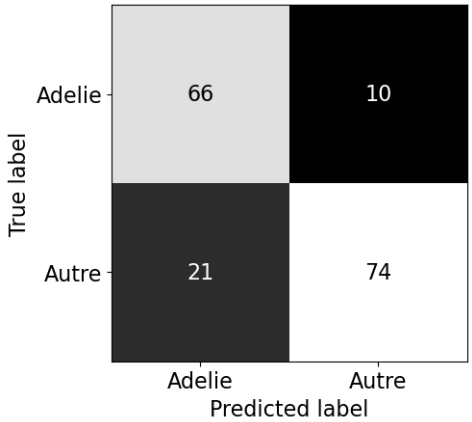
1. On veut tester le classifieur ML (maximum de vraisemblance) sur ce dataset en utilisant 50% des exemples pour l'entrainement et 50% pour le test.

   On estime donc sur l'ensemble d'entrainement les paramètres de la loi normale suivie par *flipper_length* pour chacune des deux classes (Adelie / Autre). On obtient la fonction de vraisemblance ci-dessous :



   Quelle est la moyenne de la classe Adelie ? Quelle est la classe de plus grande variance ? Quelle sera la classe prédite pour un manchot de taille de nageoire 200 mm ?

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



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

   Calculez la valeur de recall si on considère que la classe positive est la classe Adelie.

3. Quel est le nombre d'exemples du dataset ? Evaluez, avec les informations dont vous disposez, la proportion de manchots Adelie dans la population totale des manchots étudiés. On suppose que la proportion des deux classes dans l'ensemble de test est similaire à la proportion des deux classes dans la population totale.

4. On a $f(198|Adelie) = f(198|Autre)$. Quelle sera la classe prédite par un classifieur MAP pour un manchot de taille de nageoire 198mm ? Expliquez. Comment évolue la valeur de recall (classe positive Adelie) quand on passe du classifieur ML au classifieur MAP ? Expliquez.

2. https ://allisonhorst.github.io/palmerpenguins/