TD5 – Régression logistique pour la classification

L'objectif de ce TD est d'implémenter en *Python* le modèle de régression logistique étudié en cours. Des expérimentations seront ensuite menées pour évaluer la performance du modèle sur des jeux d'essai.

1. Présentation des éléments fournis

Plusieurs fichiers *Python* sont fournis afin d'aider à l'organisation du programme. La plupart d'entre eux sont à compléter afin de rendre le programme fonctionnel.

1. td5_main.py : contient le programme principal
2. lecture_donnees.py : contient la fonction permettant de lire les données à partir d'un fichier texte
3. normalisation.py : contient la fonction permettant de normaliser les données
4. sigmoide.py : contient la fonction permettant de calculer la valeur de la fonction sigmoïde
5. calcul_cout.py : contient la fonction permettant de calculer la valeur de la fonction cout
6. descente_gradient.py : contient la fonction permettant de réaliser la descente du gradient pour l'apprentissage des paramètres du modèle de régression logistique
7. prediction.py : contient la fonction permettant de prédire la classe des données par application du modèle de régression logistique
8. taux_classification.py : contient la fonction permettant de calculer le taux de classification (proportion d'éléments bien classés)
9. affichage.py : contient la fonction permettant d'afficher les données en 2 dimensions (cas de deux variables prédictives) et leur classe d'appartenance (une couleur par classe)

Afin de tester les fonctions développées, un fichier de données est fourni : notes.txt. Ce fichier contient un jeu de données caractérisées par deux variables prédictives (les notes d'étudiants obtenues à deux examens) (les deux premières colonnes). La variable cible (troisième colonne) indique si l'étudiant est admis (1) ou non (0) à l'Université. La problématique associée à ce jeu de données est donc de pouvoir prédire si un étudiant pourra être admis à l'Université compte-tenu des notes obtenues aux deux examens.

2. Ecriture du programme

Compléter les fichiers fournis pour rendre le programme fonctionnel. Il est conseillé de suivre l'ordre des fonctions indiqué dans la section précédente, et de tester le programme à chaque étape.

3. Travail à rendre

Le programme développé apprend et évalue le modèle de régression logistique sur les données d'apprentissage. Or, pour évaluer les performances réelles d'un modèle de prédiction, il est nécessaire de l'appliquer sur des données de test, différentes des données d'apprentissage.
Par ailleurs, la régression logistique permet une classification binaire. Pour un problème multi-classes, une possibilité est d'utiliser la stratégie un contre tous, comme vu en cours.

Le travail à réaliser consiste donc d'une part à développer les fonctions nécessaires afin de permettre un découpage des données en deux sous-ensembles : apprentissage et test. Ce découpage devra être paramétré par un nombre réel (entre 0 et 1) indiquant le ratio de données d'apprentissage par rapport aux données de test.

D'autre part, le programme devra être adapté à une classification multi-classes. Vous l'appliquerez à un jeu de données de votre choix (contenant au moins 3 classes) afin de l'évaluer. De nombreux jeux de données peuvent être trouvés ici : https://archive-beta.ics.uci.edu/ml/datasets

Ce travail sera à rendre pour le jeudi 27 octobre 2022. Vous devrez alors rédiger un compte-rendu électronique au format pdf indiquant votre démarche, vos résultats (tableaux de données et graphes), vos commentaires et vos interprétations des données. Ce rapport ainsi que votre programme *Python* devront être regroupés dans un fichier archive (.zip) qui sera déposé sur le site moodle.