

TD3 - 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

Un fichier Python (`regression_logistique.py`) est fourni, contenant des fonctions à compléter afin de rendre le programme fonctionnel.

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 le fichier fourni afin rendre le programme fonctionnel. Il est conseillé de suivre l'ordre des fonctions présentes dans le fichier, et de les tester à chaque étape.

3. Extension du programme et nouvelles expérimentations

Dans un premier temps, le travail consiste à étudier l'influence de la stratégie de descente du gradient en implémentant dans trois fonctions la descente par lot, stochastique, et par mini-lots.

Dans un second temps, on s'intéressera à l'évaluation du modèle. En effet, 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. Il s'agit donc de 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.

Enfin, on souhaite pouvoir utiliser la régression logistique pour traiter un problème multi-classes. Or, la régression logistique dans son utilisation standard permet uniquement une classification binaire. Comme vu en cours, une possibilité est d'utiliser la stratégie « un contre tous ». Le programme devra être adapté à une classification multi-classes selon cette stratégie. 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.ics.uci.edu/>