TD 2 : Travail avec un jeu de données conséquent

Romuald CARETTE & Olivier GERARD

Objectif

Maintenant que les bases concernant l'AG sont comprises, il est temps d'appliquer ces bases avec un jeu de données, important et brut.
Votre travail ici consiste à trouver l'équation d'une courbe en vous basant sur un ensemble de points. Cette équation est de la forme suivante :

$$y = \sum_{i=1}^{5}(ci * xi)$$

Les données fournies sont de la forme suivante :

- **y** : la valeur y de l'équation
- **x1**...**x5** : les valeurs x_i correspondantes au y.

Attention, les valeurs x_i et c_i sont entières et comprises entre -1000 et 1000.
Tous les points sont présents sur la courbe.

Travail

Etape 1

La première étape vise à maîtriser l'utilisation d'un jeu de données. Ce jeu de données peut vous être fourni sous deux formats : CSV ou MongoExport. Le premier est équivalent à un format texte, avec des séparateurs prédéfinis. Le second est un fichier spécifique à MongoDB, permettant d'importer directement une collection extraite depuis une autre base MongoDB.
Il n'y a aucune différence de notation ou de contenu entre les deux sources. Ceci dit, il s'agit potentiellement de votre seule opportunité dans cette UE d'utiliser une base de type noSQL.

Etape 2

A partir de ce point, vous êtes indépendants, aucune indication ne sera fournie hormis des informations visant à clarifier des points présents dans cet énoncé.
La seconde étape consiste à trier les données. En effet, un fichier unique est confié à tous. Ce fichier contient les données propres à chacun d'entre vous. Cependant, les données présentes peuvent avoir été altérées et vous devrez extraire les valeurs inutilisables.
Les informations de cet énoncé devraient vous être suffisantes pour définir des critères de tri.

Etape 3

Reprenez l'énoncé du TD 1, remaniez les blocs de code le cas échéant, ajoutez la partie de code utilisant MongoDB ou la lecture du CSV. Vous pouvez remplacer le texte de description du TD 1 (sauf les titres) par des explications personnelles de votre code.

Etape 4

Vous êtes sûrs de votre travail. Vous fournissez à l'encadrant de votre séance votre fichier Jupyter, incluant les coefficients trouvés. Ces coefficients doivent être écrits dans un bloc de texte, et non pas seulement en tant que valeur imprimée par votre code.
Rappel : votre code doit être commenté, si nécessaire, et toute description de votre méthodologie est la bienvenue. Attention cependant, vous devez décrire votre méthode de tri des données.

Etape 5

Vous êtes libérés ! C'est à notre tour de travailler :

- La liste de coefficients est vérifiée
    - Vaut 5 points
    - Perte de 0.1 point par erreur de 1 par coefficient
- La description est consultée
    - Jusqu'à 5 points attribués
    - Perte potentielle de 3 points si absente ou erronée
        - Orthographe incluse !
- Le code est exécuté en entier
    - 5 points si les coefficients sont obtenus sous le délai de 15 minutes
    - 5 points dépendants du sens et de la propreté du code
    - Le respect de l'aléatoire du problème est toujours de mise
        - Même pénalité : code noté 0

Dernier rappel : ce TD 2 vaut pour 75% de la note TD, les 25% restants provenant du TD 1.