Aldeano Paul
Département informatique

Documentation technique - Plateforme d'analyse et de machine learning

Introduction :

La plateforme d'analyse et de machine learning est un projet initié par Dominique Benmouffek, professeure aux Mines de Nancy, et Damien Couillard, lors de sa 3ème année au département informatique. Elle consiste à rendre accessible les algorithmes de machine learning sous la forme d'un outil simple d'utilisation. Les objectifs énoncés lors de la conception par M. Couillard sont :

- Extraire des données de différents formats.
- Visualiser ses données importées
- Afficher ses données sous forme de graphiques
- Appliquer des algorithmes d'analyse de données à ses données
- Appliquer des algorithmes de machine learning à ses données

Après une première plateforme implémentée par M. Couillard (cf. le rapport de projet qu'il a pu réaliser et le mien), j'ai fait le choix de changer de technologie pour la partie graphique. Pour cause, le besoin en technologies différentes (Docker, reactJS) de la première plateforme, auxquelles j'ai eu du mal à m'adapter, m'a poussé à la recréer entièrement en python à l'aide d'une librairie graphique. Les dépendances techniques se résument donc à python et ses librairies, et rendent à priori le projet plus abordable pour une personne souhaitant le faire évoluer.

Dépendances techniques :

Le projet est réalisé en quasi totalité sous python 3 (3.8.10 précisément). La partie visuelle est créée à l'aide d'un outil de design appelé QT designer. Cet outil génère des fichiers avec la terminaison .ui ensuite convertibles en python.

Avertissement : j'ai essayé de travailler sous Windows 10 et sous Linux (Ubuntu 20). Pour une raison inconnue, certaines librairies python ne voulaient pas s'installer correctement avec Windows. Il est probable que ce problème ne soit pas rencontré par d'autres élèves car la plupart des tutoriels sont sous Windows pour ces librairies.

Windows :

Télécharger Python 3 : https://www.python.org/downloads/

Suivre le tutoriel suivant (au moins les 3 premières vidéos, car l'accès à l'outil de design est expliqué dans la 3ème vidéo) :

https://www.youtube.com/playlist?list=PLzMcBGfZo4-lB8MZfHPLTEHO9zJDDLpYj

Pour obtenir les librairies utilisées, exécuter dans un terminal :

- sudo pip install requirements.txt

ou exécuter à la main :

- sudo pip install pandas
- sudo pip install numpy
- sudo pip install matplotlib
- sudo pip install sklearn

Linux :

Python 3 est généralement préinstallé sur Ubuntu et Debian. Si ce n'est pas le cas, exécuter :

- sudo apt install python3
- sudo apt install -y python3-pip

Ensuite, exécuter :

- sudo pip install pyqt5
- sudo pip install pyqt5-tools
- sudo apt-get install python3-pyqt5
- sudo apt-get install qtcreator pyqt5-dev-tools
- sudo apt-get install qttools5-dev-tools

Les deux premières commandes installent les librairies python nécessaires. Les 3 autres permettent d'accéder à l'outil de design en utilisant la commande

- designer <nom-fichier>

Pour obtenir les autres librairies utilisées dans le projet, exécuter :

- sudo pip install requirements.txt

ou exécuter à la main :

- sudo pip install pandas
- sudo pip install numpy
- sudo pip install matplotlib
- sudo pip install sklearn

La plateforme :

Pour lancer la plateforme, exécutez le fichier main.py.

Pour importer un fichier, cliquez sur le bouton Import File. Le chemin d'accès du fichier choisi s'affiche dans la barre à côté du bouton.

Les différents algorithmes utilisables sont listés à gauche de l'écran. Cliquez sur un bouton pour vous rendre sur la page correspondante. Exception : la page Learning est dédiée à l'explication de la plateforme.

Linear regression :

Sélectionnez les colonnes de votre jeu de données pour l'axe X et l'axe Y. La première colonne est initialisée à 0. Cet algorithme est capable de prendre en compte des jeux de données en .txt et .csv. Cliquez sur le bouton "Plot Linear Regression" pour voir s'afficher le résultat.

Polynomial regression :

Sélectionnez les colonnes de votre jeu de données pour l'axe X et l'axe Y. La première colonne est initialisée à 0. Choisissez un degré de polynôme pour la courbe (entre 1 et 6, mais le degré maximum est augmentable dans le code du projet). Cet algorithme est capable de prendre en compte des jeux de données en .txt et .csv. Cliquez sur le bouton "Plot Polynomial Regression" pour voir s'afficher le résultat.

Clustering :

Les prérequis pour cet algorithmes sont :

- d'avoir un fichier .csv
- que la première valeur des colonnes soient leur titre
- que l'unique colonne permettant de classifier les données soit la dernière colonne
- que les autres colonnes comportent des valeurs numériques

Choisissez un nombre de clusters maximal à évaluer pour choisir le nombre de clusters optimal (entre 1 et 20). Cliquez sur le bouton "Elbow Method" pour afficher l'évaluation. Le nombre de clusters correspondant à la coudure de la courbe est le nombre optimal à sélectionner.

Entrez ce nombre de clusters pour le clustering et cliquez sur le bouton "Clustering with Kmeans"

Partie Design

Le tutoriel donné précédemment permet de se familiariser avec l'outil de design et de comprendre son fonctionnement. Il est relativement court et facile à suivre, autant sur Linux que Windows :

https://www.youtube.com/playlist?list=PLzMcBGfZo4-lB8MZfHPLTEHO9zJDDLpYj

Afin de faire un outil relativement agréable à l'utilisation, j'ai suivi un deuxième tutoriel qui a posé les bases du design graphique de mon application :

- https://www.youtube.com/watch?v=5u2805f0xFw&t=21s (partie 1)
- https://www.youtube.com/watch?v=RYdAf2NH0TY&t=125s (partie 2)

L'outil de design n'est pas capable de lier les boutons et autres widgets qui sont créés à des fonctions python. Pour ce faire, il est nécessaire de convertir en .py le fichier .ui avec la commande :

- python3 -m PyQt5.uic.pyuic -x <fichier.ui> -o <nouveau_fichier.py>

Ensuite, grâce au deuxième tutoriel ci-dessus, il est expliqué comment lier des fonctions python (contenues dans ui_fonctions.py) avec le fichier main.py qui régit le fonctionnement des widgets.