Python Data Science

Nicolas rousset

Formation Data Science, day 1

Python écosystème

- Python est largement utilisé avec un écosystème large qui fournit tout ce que vous pouvez faire en programmation avec une saveur “open source” (hors interaction avec le logiciel propriétaire)
- Il possède une forte intégration dans tous les principaux fournisseurs d'hébergement Web (AWS, Google Cloud, OVH)
- Les bibliothèques Python, en particulier pandas, sont très conviviales et offrent beaucoup de possibilités intégrées (comme l'exportation d'un ensemble de données vers XML)
- Vous avez des bibliothèques pour lire / écrire excel, word files, latex ...

Numpy: arrays efficaces

Cette bibliothèque est une clé de tout l'écosystème de l'analyse scientifique / des données Python, car elle fournit la structure de type C de bas niveau qui permet un calcul haute performance.

Numpy consiste à manipuler des tableaux, des matrices ou des tenseurs (matrice de toutes dimensions).

Il fournit également des outils d'analyse de base sur ces structures (corrélation, moyenne, écart-type, etc . . .)

Vous pourriez ne pas l'utiliser directement, car elle est assez bas-niveau, mais vous devez savoir qu'elle est utilisée par toutes les bibliothèques, même la manipulation d'images. Par exemple, si vous souhaitez faire en sorte qu'une bibliothèque d'apprentissage automatique interagisse avec Pandas, elle passera par une structure numpy qu'ils utilisent toutes les deux.

Scipy

Scipy est un catalogue de fonctions mathématiques et d'algorithmes:
- l'intégration - Optimisation - équations différentielles - Valeurs propres

Il est donc utile dans de nombreux domaines spécialisés, mais pas tellement dans l'analyse des données. Je ne pense pas qu'il vaut la peine d'investir dans la compréhension directement, vous l'utiliserez lors de la saisie de "faire quelque chose de mthématique" dans StackOverflow qui fournira un code Scipy.

Pandas: Excel pour les programmeurs

Il s'agit de la bibliothèque principale utilisée par l'analyse des données, en particulier en mode d'étude. L'objet principal est le DataFrame, qui est une feuille Excel pour les programmeurs Python, il permet de gérer l'ensemble de données avec des variables de types mixtes: chaîne, int, float, date . . .

Il offre également beaucoup de fonctionnalités intégrées pratiques, y compris, mais pas seulement: - histogrammes - Exporter vers HTML / XML / LATEX - Analyse des données simples

Scikit-learn: apprentissage automatique

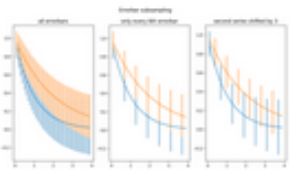
Une bibliothèque très populaire pour les algorithmes d'apprentissage automatique:

- régression linéaire et logistique
- Clustering (KNN)
- classification
- Réduction de la dimension (PCA)

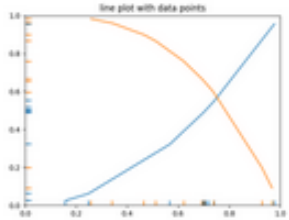
Matplotlib: dessin de graphiques

Une des bibliothèques principales de dessin de graphiques. La page principale est:

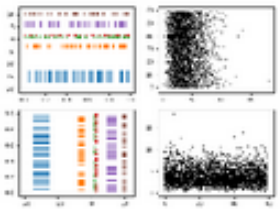
https://matplotlib.org/stable/gallery/index



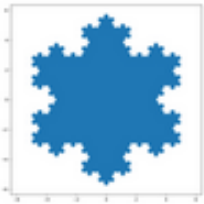
Errorbar subsampling



EventCollection Demo

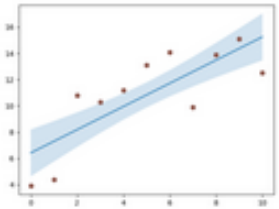


Eventplot Demo



Filled polygon

Fill Between and Alpha



Filling the area between lines

Tensorflow / Pytorch / Keras

L'une des raisons des pics de popularité de Python pour le moment, ce sont les principales bibliothèques du réseau neuronal / intelligence artificielle.

Tensorflow / keras sont développés par Google Pytorch est développé par Facebook

Opencv

Une des principales librairies opensource de traitement d'image et de traitement vidéo. La librairie est écrite en C++ (comme beaucoup de librairies scientifiques).

On peut également y trouver des algorithmes de machine learning, mais elle est moins populaire pour cet usage que scikit learn, tensorflow et d'autres librairies.

Seaborn

Une bibliothèque pour la visualisation du graphique et des données

Geopandas / cartopy

Pour la manipulation et la visualisation des données géographiques