Ref : DNDAFP15 | **Durée :** 6 jours | **Tarif :** 3 000 € HT

Data Mining & Machine Learning Python

Contenu

Objectifs :

Connaître et savoir utiliser les librairies incontournables de Python pour la data science : La Scipy Stack
Connaître et utiliser les principales librairies de visualisation de données et notamment orientées cartographie
Savoir manipuler des données volumineuses ne tenant pas en mémoire
Avoir une bonne compréhension de l'écosystème scientifique de Python, savoir trouver ses librairies et juger de leur qualité

L'écosystème scientifique Python

Il n'est pas facile d'y voir clair dans l'écosystème scientifique de Python tant les librairies sont variées et nombreuses.
Cette présentation vous apportera une vue d'ensemble et les éléments clefs qui vous aideront à choisir vos librairies et outils de travail pour vos projets de data science avec Python.

Les incontournables : Numpy, Scipy, Pandas, Matplotlib et iPython qui sont le ciment de toutes les autres librairies scientifiques
Panorama des librairies et logiciels scientifiques par domaine
Les critères permettant de juger de la qualité d'une librairie
Calculer avec des nombres réels: comprendre les erreurs de calculs

Les nombres réels, dans la plupart des langages, dont Python, utilisent la norme en virgule flottante.

Celle-ci n'est pas précise et peut générer des erreurs de calcul parfois bien gênantes.

La représentation des nombres réels
Comprendre les erreurs de calculs et les contourner
La scipy stack

La librairie Numpy qui signifie Numeric Python est la première que vous devez apprendre. Elle constitue avec Scipy, Matplotlib et Pandas le socle sur lequel s'appuient toutes les autres librairies scientifiques.

Manipuler des tableaux de nombres : Numpy
Différences avec les listes Python
Création, sélection, filtres et principales fonctions -Visualiser ses données : Matplotlib
Les concepts de la librairie

Principaux graphiques : nuages de points, courbes, histogrammes, boxplot, ...
Fonctionnalités avancées : 3D, légendes, colorbar, manipuler les axes, annotations, ...
Analyse de données : Pandas
Les fondements de la librairie : Manipuler des données de type CSV et Excel
Séries et Dataframes
Index, sélection de données, filtres/recherche, agrégations, jointures et fonctions avancées
Manipuler des séries temporelles
Les fonctions mathématiques avancées: Scipy
Statistiques, optimisation, interpolations/régressions, traitement d'images

Visualisation de données

Bien que Matplotlib constitue la première librairie de visualisation que vous devrez
apprendre, elle possède 2 limites majeures: elle ne sait pas gérer les données volumineuses et n'est pas adaptée au Web. Mais Python a su développer un riche écosystème de visualisation de données qui devrait pouvoir répondre à toutes vos attentes.

Présentation de l'écosystème de visualisation de données de Python
Les librairies orientées Web: Bokeh, Altair et Plotly
Les "écosystèmes" PyViz et HoloViz
La visualisation de données volumineuses/big data avec DataShader
Les statistiques avec Seaborn

Visualiser des données géospatiales

Posséder des données disposant de coordonnées géospatiales apporte une toute autre dimension à leur représentation. Python est très bien outillé dans ce domaine.

Convertir ses données d'un système de coordonnées à l'autre
Cartographie interactive "à la Open Street Map/Google Maps" avec Folium/iPyleaflet
Cartographie statique avec Cartopy
Autres librairies géospatiales

Manipulation de données volumineuses

Numpy et Pandas sont 2 librairies incroyables, mais elles ont 2 limites majeures : elles ne savent pas traiter des données de très grande volumétrie qui ne tiennent pas en mémoire et ne savent pas toujours paralléliser leurs calculs.

Python a su développer des solutions.
Les librairies h5py, pytables, netcdf4, xarray, iris, parquet permettant de lire vos fichiers scientifiques
Paralléliser ses calculs avec Dask
Paralléliser ses calculs avec CuDF
Manipuler des dataframes gigantesques avec Dask

Personnalisation

Sous réserve de contraintes techniques ou de confidentialité, nous vous proposons de personnaliser la formation en réalisant des exercices directement sur vos données métiers.
Apprentissage et analyse statistique avec scikit learn & statsmodels *Revue des techniques *L’analyse discriminante *La régression

logistique *Les arbres de décision *Gestion des ensembles d’apprentissage et de test *Évaluation des modèles *Introduction à l’utilisation de Spark avec Python (pyspark).

Pré-requis

Pour suivre ce stage dans de bonnes conditions, il est recommandé d'avoir suivi en amont la formation Python - Bases et introduction aux librairies scientifiques.

Méthodes pédagogiques

Alternance d'apports théoriques, d'exercices pratiques et d'études de cas.