Les bases du machine learning avec Python et Scikit-Learn

Date : **13 au 15 mai 2025**

Lieu : Espace Vinci

Nombre de stagiaires : 10

Objectifs

- Identifier les problèmes pouvant faire appel à l'apprentissage automatique (machine learning)
- Découvrir les modèles classiques d'apprentissage supervisé (hors réseaux de neurones et deep learning)
- Évaluer les performances des modèles d'apprentissage
- Utiliser la librairie scikit-learn pour comprendre et mettre en œuvre les différentes étapes d'un traitement de données par apprentissage supervisé : choix du modèle, prétraitement, entrainement, validation croisée, évaluation

Public visé

Chercheurs et ingénieurs en lien avec des problématiques liées à la science des données et désirant se lancer dans la mise en œuvre de traitement de données par apprentissage automatique.

Modalités pédagogiques

La formation alternera des parties théoriques et de la mise en pratique par les stagiaires afin qu'ils développent une intuition par eux-mêmes. Il est recommandé de se focaliser sur un jeu de données unique mais représentatif des cas pour illustrer les trois jours et pouvoir se focaliser sur le contenu plutôt que sur la compréhension des données

Programme

1/ Introduction à l'apprentissage automatique

- Historique et motivations
- Panorama des différents types d'apprentissage

- L'apprentissage supervisé : classification vs. régression
- Exemples concrets

2/ Les modèles classiques

- Modèles linéaires (régression linéaire / logistique)
- Arbres de décision et modèles sous-jacents (random forests, boosted trees)
- Support vecteur machines (SVMs) et leurs kernels
- Quelques modèles très utilisés dans l'industrie (XGBoost, LightGBM)

3/ La réduction de dimensions

- La malédiction de la dimension
- Variables corrélées / décorrélées du problème
- L'apprentissage non-supervisé : PCA Manifold projection (t-SNE, UMAP, …)

4/ Pré-traitement des données

- Normalisation
- Encodage des variables qualitatives (one-hot vs. ordinal encoding)
- Augmentation de données (polynomial features)
- Traiter les données manquantes

5/ Pipelines scikit-learn

- Découverte du Pipeline scikit-learn
- Enchainement de pré-traitement et d'entrainement de modèle.
- Visualisation du Pipeline
- Introduction du ColumnTransformer

6/ Sélection des modèles

- Théorème du "No-free lunch"
- Comparer des modèles
- Scores / métriques de performance

7/ Optimisation des modèles

- Les hyper-paramètres
- Recherche en grille (grid search) vs. aléatoire (random search)
- Recherche d'hyperparamètres et cross-validation grâce au Pipeline
- Visualisation des résultats avec parallel plots (ex : plotly ou hiplots)

Conditions d'inscription

Date limite d'inscription : 03/02/2025

Inscription : https://formation.ifsem.cnrs.fr/

Renseignements :
ifsem-formation.contact@cnrs.fr