UM6P Data Science Data Analytics Engineering

Programme 24 septembre au 5 octobre

Le cours est centré sur l'analyse prédictive supervisée et non-supervisée.

L'environnement de travail repose sur les notebooks Jupyter, (Python 3, scikit-learn).

Les éléments théoriques et mathématiques sont enseignés le matin, l'après-midi est consacrée aux travaux pratiques (LAB). Chaque journée débute par un récapitulatif des éléments vus la veille pour permettre aux étudiants de revenir sur les points qui restent à éclaircir. Les étudiants se verront proposés des lectures d’articles pour le lendemain.

## Projet final

Un projet final basé sur une compétition Kaggle fera l'objet de présentations le dernier jour. Les étudiants travailleront en binôme. Ce projet leur permettra d’utiliser la plupart des notions vus dans le cours et de les appliquer sur un dataset non trivial.

## Outils

Les étudiants travailleront sur des Notebook Jupyter, soit sur leur ordinateur personnels soit sur un environnement similaire mis à leur disposition sur Google Cloud. Le but étant de s’affranchir des contraintes induites par des ordinateurs personnels peu puissants mais aussi de pouvoir partager et présenter du code facilement.

# Jour 1 : Intro, data science et régression linéaire

* Matin :
  + Intro
  + Présentation du programme, de la compétition Kaggle, du travail attendu
  + Mise en place de l'environnement de travail (Jupyter notebook) et des outils (python, scikit-learn, numpy, pandas, ...),
  + Qu'est-ce que la data science ?
    - DS vs ML vs AI vs RL vs DL.
    - Finalités et étapes d'un projet.
    - Supervisée vs non-supervisée,
    - Régression vs classification.
  + Jeux de données classiques
* Après-midi :
  + LAB : Régression(s) sur des datasets simples
    - Régression linéaire,
    - Régression logistique,
    - Conditions sur les données, cas d'application, p-value, R^2, intervalles de confiance, ...

# Jour 2 : Modèles ou Algorithmes

* Matin :
  + Pratique du maximum de vraisemblance
  + Importance de la linéarité : définition et tests statistiques associés
  + Impact de la multi-colinéarité
  + Régression polynomiale
* Après-midi :
  + Présentation du projet final : compétition Kaggle.
  + LAB : On continue sur les différents types de régression avec des datasets non linéaires

# Jour 3 : Gradient stochastique

* Matin :
  + Recap : l'analyse prédictive supervisée : Modèle ou Algorithme, minimisation d'une fonction de cout, score, prédicteurs et outcome, ...
  + Découpage en test, train validation ; validation croisée
  + Algorithme du gradient stochastique, théorie, pratique et applications
* Après-midi :
  + LAB : Gradient stochastique, convergence, paramétrisation, visualisation, scikit-learn, batch mode, conditions
  + Cas : L’utilisation du gradient stochastique dans le service AWS ML

# Jour 4 : Arbres et Forets

* Matin :
  + Arbre de décisions, forets aléatoires, XGBoost
  + Ensembling et Boosting
  + Score de classification : matrice de confusion, AUC, F1, ...
* Après-midi :
  + Présentation d’un cas : classification d’échantillons de musique électronique pour Splice.com
  + LAB : forets aléatoires et XGBoost

# Jour 5 : Support Vector Machines et overfitting

* Matin :
  + Support Vector Machines avec diffèrent kernels
  + Overfitting, compromis biais-variance, régularisations L1 et L2, Ridge et Lasso
  + Courbes d’apprentissage : détecter et corriger l'overfitting
* Après-midi :
  + LAB : SVM
  + LAB : Régularisation

# Jour 6: Création et selection des variables

* Matin :
  + Modèle bayésien naif
  + Feature engineering approche manuelle, brute ou approche bayésienne
  + Améliorer les variables : Box-Cox, valeurs manquantes ou aberrantes
  + Bootstrapping
  + Selection des variables
  + Curse of dimensionality
  + Reduction de dimension
* Après-midi :
  + LAB : nettoyage des données, feature engineering et selection

# Jour 7: Classification déséquilibrée

* Matin :
  + Le paradoxe de la précision
  + Methodes de sous-echantillonnage et de sur-echantillonnage
  + SMOTE
* Après-midi :
  + Cas : prediction de resiliation dans l’assurance
  + LAB : sur dataset caravan
  + Point projet

# Jour 8 : Non-supervisé

* Matin :
  + Kmeans, Nearest Neighbor, T-SNE, ...
  + Approche semi supervisée
* Après-midi :
  + LAB
  + Point projet

# Jour 9 : Séries temporelles

* Matin :
  + Exponential smoothing
  + Modèles linéaires AR, MA, ARIMA
  + Décomposition en tendance, saisonnalité et résidus
* Après-midi :
  + Cas : Prediction de la demande pour un producteur d’electromenager
  + LAB sur séries temporelles

# Jour 10 : Neural Nets avec Keras

* Matin :
  + Intro aux reseaux de neurones
  + Backpropagation, fonction d'activation, architecture
  + Reseaux Convolution, Recurrents
  + Domaines d’applications
* Après-midi :
  + Présentations des projets étudiants

Les sujets suivants ne sont pas abordés mais pourrait remplacer la derniere demi-journée :

* Data visualisation avec matplotlib
* Traitement de données textuelles