SCIENCE DES DONNÉES

DATA MINING

LE PROGRAMME ET LE CONTENU DU STAGE SONT ADAPTÉS AUX CURSUS ET AUX OBJECTIFS PROFESSIONNELS DES PARTICIPANTS. LA PÉDAGOGIE EN TOUT PETIT GROUPE PEUT ÊTRE BIEN DIFFÉRENCIÉE, NOTAMMENT PAR LA PRATIQUE DE CHACUN SUR SON POSTE PERSONNEL.

LES PRÉREQUIS

A minima, sont ceux d'un enseignement scientifique à Bac+2 avec un contenu en Statistique mais le stage peut aussi être calibré à Bac+4 ou Bac+5, au niveau de la formation dispensée en Science des Données dans la spécialité Mathématiques Appliquées de l'INSA Toulouse.

OBJECTIFS

L'objectif de cette formation est d'aborder et pratiquer une sélection des méthodes récentes de statistique et d'apprentissage machine appliquées à des données de grandes dimensions pour la fouille de données (data mining). L'accent est mis sur les techniques récentes d'exploration, classification non supervisé (clustering) et modélisation, prévision. Les applications sont réalisées avec le langage R ou en Python sous la forme de calepins (notebook jupyter) exécutables. Les supports pédagogiques sont disponibles sur le site : http://wikistat.fr.

Le programme peut être précisé à partir des mots-clefs ci-dessous en fonction des besoins et problématiques de l'entreprise ou des participants.

PROGRAMME DU STAGE

La formation alterne exposés méthodologiques des algorithmes concernés, et pratique sur des données réelles.

Introduction

- Changements de paradigmes en statistique : data mining, apprentissage statistique, big data analytics.

Principales méthodes

- Exploration multidimensionnelle (analyse en composantes principales)
- Classification non supervisée (clustering) par méthode hiérarchique ou partitionnement dynamique
- Estimation d'une erreur de prévision et risque.
- Modèle linéaire et régression logistique (sélection de modèle par sélection de variables et/ou pénalisation)
- Analyse discriminante décisionnelle et algorithme des k plus proches voisins
- Arbres binaires de décision (CART) pour la régression ou la classification
- Réseaux de neurones et introduction à l'apprentissage profond (deep learning)
- Agrégation de modèles (boosting, bagging, random forest)
- Introduction aux SVM (support vector machines ou séparateur à vaste marges)
- Détection d'anomalie ou d'observations atypiques

Étude de cas

En fonction des besoins et centres d'intérêt des participants.

- Pratique de ces méthodes avec le langage R et/ou Python sur différents types de jeux de données de complexité ou volume élémentaire à élevé : prédiction de pics d'ozones, reconnaissance d'activité humaine à partir de signaux enregistrés sur un smartphone, reconnaissance de caractère sur des images.

INFOS

DATE : à la demande

DURÉE DU STAGE : 3 jours | 21 heures

TARIF : 1 500€ | Documents pédagogiques et déjeuners inclus

RENSEIGNEMENTS & INSCRIPTION :
05 61 55 92 53 | fcq@insa-toulouse.fr

Une attestation de suivi de formation sera transmise à l'issue de celle-ci.

ANIMATION DE LA FORMATION :

- **Mélisande ALBERT**
- **Béatrice LAURENT-BONNEAU**
- **Olivier ROUSTANT**

Professeurs au département Génie Mathématiques et Modélisation | INSAT

Institut de Mathématiques de Toulouse