IA DE CONFIANCE

NOUVEAU PROGRAMME | AVANCÉ | ATELIER | RÉALISABLE EN ANGLAIS | FAISABLE À DISTANCE

FFCNCERCERXBD33	PRIX : 1 930 €	DURÉE : 2 JOURS

Pauses et déjeuners offerts

PRÉSENTATION

L'intelligence artificielle irrigue désormais des décisions à fort enjeu, et l'on attend des modèles qu'ils soient non seulement performants, mais aussi explicables, respectueux de la vie privée, robustes et équitables. Ces propriétés ne s'ajoutent pas après coup, elles se construisent dès la conception.

Cette formation outille les participants pour mesurer, diagnostiquer et améliorer la fiabilité d'un système d'apprentissage automatique sur quatre axes (explicabilité, confidentialité, robustesse, équité) à travers des méthodes concrètes et leur mise en œuvre.

OBJECTIFS

- Diagnostiquer les défauts de confiance d'un modèle d'apprentissage automatique à l'aide de métriques appropriées
- Mettre en œuvre les principales méthodes d'explicabilité et en apprécier la fidélité et les limites
- Implémenter des techniques de préservation de la vie privée et évaluer le compromis utilité/protection qu'elles induisent
- Concevoir et évaluer des modèles robustes face aux attaques adverses et à l'empoisonnement des données
- Mesurer l'(in)équité d'un système, appliquer des techniques d'atténuation et arbitrer entre des critères d'équité incompatibles

PROGRAMME

Introduction

Cadre et explicabilité par conception
- Propriétés d'une IA de confiance vues comme objectifs de conception ; tension performance / confiance
- Modèles interprétables par conception : modèles linéaires, arbres, modèles additifs généralisés, mécanismes d'attention
- Compromis précision / interprétabilité : quand un modèle simple suffit-il ?
- TP

Explicabilité ex-post
- Méthodes agnostiques au modèle : LIME, SHAP, importance par permutation, dépendances partielles
- Explications locales vs. globales et explications contrefactuelles
- Interprétation des réseaux profonds : cartes de saillance, Grad-CAM, sondes
- Pièges : instabilité, infidélité au modèle, illusions d'explication
- TP

Vie privée et robustesse
- Attaques par inférence d'appartenance et par inversion de modèle
- Confidentialité différentielle : mécanismes, budget, intégration à l'entraînement (DP-SGD)
- Apprentissage fédéré et génération de données synthétiques
- Attaques adverses et empoisonnement
- TP

Équité dans l'apprentissage
- Sources de biais dans le pipeline de modélisation
- Définitions formelles de l'équité et leur incompatibilité
- Atténuation en amont, pendant (sous contrainte), et en aval de l'entraînement
- Métriques d'équité, model cards comme support de conception
- TP

Synthèse et conclusion

PUBLIC/PRÉREQUIS

La formation s'adresse aux ingénieurs, chefs de projets souhaitant approfondir leurs connaissances en apprentissage statistique. Elle a pour but de détailler le développement des méthodes considérées ainsi que de fournir des éléments théoriques
justifiant leurs performances.

Des connaissances de base en statistiques ou en machine learning : notions de probabilités/statistiques élémentaires (variables aléatoires, loi d'une variable aléatoire, espérance, etc.) ainsi que d'une connaissance des enjeux des méthodes d'apprentissage sont nécessaires afin de tirer pleinement profit de cette formation.

RESPONSABLE(S)

Charlotte LACLAU
Enseignante-chercheuse au LTCI à Télécom Paris dans le département Image, Données et Signal. Ses intérêts de recherche portent sur l'apprentissage automatique et plus spécifiquement l'apprentissage de représentation pour des données complexes, avec un accent sur les graphes dynamiques et le texte. De plus, elle travaille sur l'analyse théorique des biais dans l'apprentissage automatique et le développement d'algorithmes équitables pour les données relationnelles.

MODALITÉS PÉDAGOGIQUES

Les concepts sont illustrés par de nombreux exemples utilisant des données simulées ainsi que des données réelles (données économétriques, météorologiques, applications en vision). Des séances de travaux pratiques en Python sont réalisées.