RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

DR Paris Ile de France Sud
Formation Continue

Inserm
La science pour la santé
From science to health

FOAD - Inter - Machine learning et méthodes statistiques appliquées aux processus de classification

Les 6, 7, 8 et 9 novembre 2023

Objectifs	- Maîtriser le vocabulaire spécifique aux méthodes d'apprentissage à finalité de classement - Identifier le contexte et les conditions d'application des méthodes d'apprentissage supervisé et non supervisé - Connaître les objectifs et les différences entre les méthodes de classement - Décrire la méthodologie inhérente à ces méthodes - Mettre en oeuvre et interpréter les résultats des méthodes d'apprentissage supervisé - Connaître les principaux indicateurs de cohérence liés aux méthodes d'apprentissage - Maitriser les paramètres permettant d'estimer la qualité de ces analyses
Public	Toute personne souhaitant comprendre le contexte d'utilisation, les concepts, et la mise en oeuvre des méthodes de classements et de prédiction - Ingénieurs et chargés d'études/d'analyses, Chercheurs, Doctorants, Statisticiens, Data scientists
Pré-requis	- Connaissances sur les outils statistiques de base : corrélation, écart-type, variance, intervalles de confiance, tests d'hypothèses. - Dans le cas où la formation serait effectuée avec le logiciel R, une connaissance de base de ce logiciel est préconisée Après votre inscription dans Sirène, compléter le questionnaire suivant : https://forms.office.com/Pages/ResponsePage.aspx?id=Sdmr0PDW80WLuFdUfKcVNhSl5D6av0tFtLyEn6XmsGhURE5aTzcyTlJLSVg4WUdERVo0QlAxSTdUQSQlQCNjPTEu *Si le lien ne fonctionne pas en cliquant dessus ->copier/coller dans votre barre de recherche Internet* **Sans ce questionnaire complété et votre inscription dans Sirène validé par votre supérieur hiérarchique avant la date limite d'inscription, votre demande de formation ne sera pas étudiée**
Programme	**VOLET 1 : LES ALGORITHMES** **Généralités sur les différentes méthodes d'apprentissage supervisé** - Différences entre méthodes supervisées et non supervisées - Objectifs de l'apprentissage supervisé : Objectifs de description - Objectifs de prédiction - Structure des jeux de données - Présentation générale de l'éventail des méthodes **La méthode knn** - Principe de la méthode des plus proches voisins - Algorithme de calcul - Distances entre individus - Choix des proximités - Définition du paramètre k **La régression logistique** - Variable explicative et variable expliquée (continue / binaire) - Différences entre la régression linéaire classique et la régression logistique - Variables explicatives qualitatives, variables explicatives quantitatives - Objectifs de la régression logistique - Définition du modèle Logit (courbe sigmoïde) - Conditions d'utilisation à respecter - Estimation et interprétation des coefficients du modèle - Test de significativité du modèle (validation du modèle) - Tests d'apport d'une variable (test de Wald, tests sur les rapports de vraisemblance) - Interprétation du Khi² de Wald - Odds-ratios - Parallèle odds ratios et risques relatifs - Mise en oeuvre et analyse des résultats d'un modèle de régression logistique multiple - Estimation et interprétation des coefficients du modèle multiple

L'analyse factorielle discriminante

- Structure du jeu de données et contexte d'application
- Objectifs détaillés de l'AFD
- Notions de classement et de discrimination
- Méthodologie de l'AFD
- Comparaison avec l'ACP
- Interprétation des sorties logiciel : cercle factoriels, corrélations variables x axes
- Qualité de l'AFD (de la discrimination obtenue) : Tests univariés et multivariés (lambda de Wilks) - Graphique des individus - Matrice de confusion (et éventuellement courbe ROC)
- Les confusions et erreurs à ne pas commettre

Les supports vecteurs machines (SVM)

- Démarche des svm
- Notions de marge
- Séparation linéaire
- Séparation non linéaire
- Fonction noyau

Les arbres de décision

- Structure du jeu de données
- Principes, vocabulaire et objectifs
- Notion d'échantillon d'apprentissage, de validation et de test
- Comparaison de méthodes de type régression linéaire / logistique aux arbres de décision
- Principe de la segmentation selon le type de variable : Arbre de régression ou arbre de classification
- Définir les conditions d'arrêt de construction d'un arbre : Notion de pré-élagage
- Définition des groupes après construction de l'arbre
- Indicateurs de qualité
- Comparaison d'arbre de décision selon un certain type d'algorithme : CHAID vs CART
- Avantages et inconvénients : limites des arbres de décision
- Mise en oeuvre et interprétation des résultats obtenus après application d'une analyse par arbre de décision

De l'arbre à la foret - Random Forest

- Pourquoi avoir recourt aux forêts aléatoires ?
- Principes et objectifs o Instabilité de l'arbre : Notion de Bagging - Les erreurs liées à l'échantillonnage (Out-Of-Bag) - Prédiction avec un algorithme de Forêt aléatoire : Les paramètres
- Evaluer l'importance des variables : Notion d'importance - Comportement de l'importance - Lien entre diversité des arbres et l'importance - Influence des paramètres
- Sélection de variables : Généralités et principes de la sélection - Procédure de sélection - Les paramètres de sélection - Validation

V**OLET 2 : VALIDATIONS DES METHODES, MESURE DES PERFORMANCES**

- Partitionnement des données disponibles : Jeu d'entrainement - Jeu de validation - Tests sur le jeu d'entrainement - Tests sur le jeu de validation
- La validation croisée : Leave one out : - K fold - Leave v out
- Compromise biais / variance
- Mesures des performances des classifications : Matrices de confusions - Courbe Roc - Aire sous la courbe (AUC) - Sensibilité & spécificité

Dates

Date : **Les 6, 7, 8 et 9 novembre 2023**
Inscriptions avant **le 20 septembre 2023** **sur** **https://www.sirene.inserm.fr/**

FOAD : formation organisée à distance

Formation à distance sur votre lieu de travail ou à domicile.
Au sein d'un groupe restreint, vous suivez une formation en direct avec un formateur dédié qui dispose d'une solution de visio-conférence

Conditions pour pouvoir suivre la formation : Ordinateur avec micro (caméra non obligatoire), accès à internet, deuxième écran facilitateur mais non obligatoire.

Contact

Catherine Marcilhac
INSERM – DR Paris IdF Sud
@ : formation.paris11@inserm.fr