

3MD3030 - Méthodes géométriques dans l'analyse des données

Responsables : **Fragkiskos MALLIAROS**

Langues d'enseignement : **FRANCAIS**

Campus où le cours est proposé : **CAMPUS DE PARIS - SACLAY**

Nombre d'heures d'études élèves (HEE) : **40**

Nombre d'heures présentielles d'enseignement (HPE) : **24**

Année académique : **2024-2025**

Niveau avancé : **non**

Présentation, objectifs généraux du cours :

L'analyse des données est le processus de nettoyage, de transformation, de modélisation ou de comparaison des données, afin de déduire des informations utiles et de mieux comprendre des phénomènes complexes. D'un point de vue géométrique, lorsqu'une instance (un phénomène physique, un individu, etc.) est donnée comme une collection de taille fixe d'observations à valeurs réelles, elle est naturellement identifiée à un point géométrique ayant ces observations comme coordonnées. Toute collection de telles instances est alors considérée comme un nuage de points échantillonné dans un espace métrique ou normé.

Ce cours passe en revue les constructions fondamentales liées à la manipulation de tels nuages de points, en mélangeant les idées de la géométrie et de la topologie computationnelles, des statistiques et de l'apprentissage automatique. L'accent est mis sur les méthodes qui non seulement présentent des garanties théoriques, mais qui fonctionnent également bien dans la pratique. En particulier, des références logicielles et des jeux de données d'exemple seront fournis pour illustrer les constructions.

Période(s) du cours (n° de séquence ou hors séquence) :

SM10

Prérequis :

Il n'y a pas de prérequis officiel pour ce cours. Cependant, il est attendu des étudiants qu'ils aient une bonne connaissance de :

- Des bases en algorithmique (notions de complexité).
- Bases en algèbre linéaire, géométrie, théorie des probabilités.

La maîtrise d'un langage de programmation (C/C++, python, R) est également attendue.

Plan détaillé du cours (contenu) :

Les plus proches voisins dans les espaces euclidiens et métriques : structures de données et algorithmes de recherche

Plus proches voisins dans les espaces euclidiens et métriques : analyse

Algorithmes de réduction de la dimensionnalité

Couvertures et nerfs : inférence géométrique et l'algorithme Mapper
Algorithmes de classification et introduction à l'homologie persistante
Tests d'hypothèses statistiques et tests à deux échantillons (TST)
Comparaison de distributions à haute dimension, comparaison de clustering
Signatures de forme : stabilité et aspects statistiques

Déroulement, organisation du cours :

Chaque cours comprend deux conférences (1h30' chacune), fournissant les bases théoriques, ainsi que des illustrations des méthodes sur des ensembles de données pratiques.

Organisation de l'évaluation :

L'évaluation du cours sera basée sur un projet, réalisé par des groupes de deux étudiants. Deux types de projets seront proposés, en fonction de la formation des étudiants. D'une part, les projets à tendance mathématique-algorithmique permettront d'approfondir la compréhension théorique d'un résultat de recherche récent. D'autre part, les projets de traitement de données permettront d'acquérir une compréhension fine de données complexes, en utilisant les méthodes (sélectionnées) étudiées pendant le cours.

Pour chaque projet, le code développé et un rapport seront rendus.

Acquis d'apprentissage visés dans le cours :

Les objectifs sont doubles :

maîtriser les principes fondamentaux de l'analyse des données géométriques,
acquérir l'expertise nécessaire pour décider quelles sont les méthodes les plus appropriées pour traiter des données d'un certain type.

Description des compétences acquises à l'issue du cours

:

- The students are expected to master the fundamentals of geometric data analysis, and to acquire expertise to decide which methods are practically best suited to deal with data of a certain type.

Bibliographie :

A reading list for each course is provided on the course web site:

<https://www-sop.inria.fr/abs/teaching/centrale-FGMDA/centrale-FGMDA--cazals-carriere.html#>