Université de Bretagne Occidentale

MASTER MARINE SCIENCES

PARCOURS PHYSIQUE OCÉAN ET CLIMAT

semestre 9 PM POC

OPTION (UNE SEULE AU CHOIX)

Bases de données - Big data

Présentation

Le volume et le diversité grandissante des données océanographiques et climatiques fournies par les satellites, les réseaux de mesures *in situ* et les modèles climatiques, nous permettent de mieux comprendre le système climatique. Cependant, les méthodes d'analyse traditionnelles ne sont plus forcément adaptées pour traiter efficacement un si grand volume et une si large diversité de données. Il est donc nécessaire de mettre en œuvre de nouvelles approches pour extraire les informations utiles des masses croissantes et complexes de données. Ces dernières années, les outils et méthodes d'apprentissages ont permis d'énormes progrès dans les domaines allant de la recherche sur le web à la bio-informatique. On peut ainsi anticiper l'impact décisif de ces outils pour les sciences de l'océan et du climat.

Les compétences en traitement de données acquises sont complémentaires à celles proposées dans le cursus de M2, *i.e.* mesures *in situ* et analyses de données, et permettent d'aller plus loin avec le traitement de grosses bases de données océanographiques et climatiques qui prennent une place grandissante dans les sciences du climats.

2 crédits ECTS

Volume horaire

Projet tutoré : 12h
Cours Magistral : 15h
Travaux Pratiques : 10h
Autres : 12h

Objectifs

Le module 'Introduction to Big Data' vise à introduire au près des étudiants en physique de l'océan et du climat (M2) l'approche et les outils associés au 'Big Data', ainsi que les éléments de base de méthodes statistiques d'apprentissage, et de fouilles de données massives.

Ces objectifs se déclinent en deux volets :

i) Familiarisation avec l'architecture et maîtrise des outils basiques (cloud, cluster, software...) de manipulation de grosses bases de données ;

ii) Introduction à l'analyse de données basée sur les méthodes de classification et d'apprentissage statistique simple.

Pré-requis nécessaires

Étudiants de M2 Physique de l'Océan et du Climat (Master Science de la Mer et du Littoral, IUEM/UBO) ; et en formation continue.

Les étudiants doivent posséder un solide bagage en mathématiques, probabilités/statistiques et des bases en informatique.

Compétences visées

Ce cour renforce également la composante 'Sciences de l'ingénieur' du M2. Les compétences acquises sont également valorisables auprès de l'industrie offrant des débouchés futurs et massifs pour ce type de compétences.

A l'issu du module les étudiants devront être capable :

-de mettre en place un environnement 'big data' en ligne

-d'utiliser les outils techniques de bases associée à un environnement 'big data'

-de réaliser des opérations simples et d'extraire des informations utiles d'une source de données massives grâce à cette environnement

-maîtriser les d'apprentissage statistique simple permettant la classification de séries de données spatio-temporelles (classification avec méthode Kmean ; modèles de mélanges Gaussians - GMM)

Descriptif

1) Module « Big Data »

Familiarisation avec les outils de manipulation de base données massives en utilisant notamment des solutions disponibles en ligne. Les outils sont principalement des espaces de stockage en ligne ('Cloud'), des nœuds de calculateurs ('Cluster'), et des suites de logiciels permettant d'exploiter de façon optimale l'environnement de machine sollicité et traiter de façon efficace de gros volumes de données. Les étudiants devront réaliser des opérations simples et extraire des informations utiles à partir d'un environnement 'big data' en ligne qu'ils auront mis en place avec leur jeu de données géophysiques (ex : données satellites de SST).

2) Module « Spatio/Temporal Data Mining»

Familiarisation avec les méthodes statistiques de fouille de données visant à identifier des schémas récurrents, ou pattern, dans des bases de données spatio/temporelle multi-paramètres et multi-dimensionnelles. Les méthodes explorées seront celles de classification et d'apprentissage supervisées et non-supervisées. Les étudiants devront apprendre les principes mathématiques de ces méthodes pour bien en comprendre les domaines d'application et à les utiliser dans des cas simples à l'aide de librairies logiciels standards.

Dans les deux cas les données utilisées seront des données géophysiques océanographiques/atmosphériques provenant de sources massives : données satellites à haute résolution, sortie de modèles de climat multiples, ...

1) « Big Data » : 24 h = 8 CM + 16 TP de manipulation de données sur le cloud (répartition des heures à définir). Simple à mettre en œuvre avec les outils disponibles en ligne, mais avec un coût (prévoir ~500$ pour une session).

2) « Spatio/temporal Data Mining » : 24h = 8h CM + 4h TP + 12h Projet. Un certains nombres de projets seront pré-préparés mais les propositions à l'initiative des étudiants sera favorisée.

Modalités de contrôle des connaissances

Session 1 ou session unique - Contrôle de connaissances

Nature de l'enseignement	Modalité	Nature	Durée (min.)	Coefficient	Remarques
	CC	Autre nature		100%	

Session 2 : Contrôle de connaissances

Nature de l'enseignement	Modalité	Nature	Durée (min.)	Coefficient	Remarques
	Report de notes	Autre nature		100%	report de note session 1